To my daughter Mia, my wife Anne,
my son Philippe, and my daughter Sylvie.
Preface

The motivations for writing these notes arose while I was coteaching a seminar on Special Topics in Machine Perception with Kostas Daniilidis in the Spring of 2004. In the Spring of 2005, I gave a version of my course Advanced Geometric Methods in Computer Science (CIS610), with the main goal of discussing statistics on diffusion tensors and shape statistics in medical imaging. This is when I realized that it was necessary to cover some material on Riemannian geometry but I ran out of time after presenting Lie groups and never got around to doing it! Then, in the Fall of 2006 I went on a wonderful and very productive sabbatical year in Nicholas Ayache’s group (ACSEPIOS) at INRIA Sophia Antipolis, where I learned about the beautiful and exciting work of Vincent Arsigny, Olivier Clatz, Hervé Delingette, Pierre Fillard, Grégoire Malandain, Xavier Pennec, Maxime Sermesant, and, of course, Nicholas Ayache, on statistics on manifolds and Lie groups applied to medical imaging. This inspired me to write chapters on differential geometry, and after a few additions made during Fall 2007 and Spring 2008, notably on left-invariant metrics on Lie groups, my little set of notes from 2004 had grown into the manuscript found here.

Let me go back to the seminar on Special Topics in Machine Perception given in 2004. The main theme of the seminar was group-theoretical methods in visual perception. In particular, Kostas decided to present some exciting results from Christopher Geyer’s Ph.D. thesis [75] on scene reconstruction using two parabolic catadioptric cameras (Chapters 4 and 5). Catadioptric cameras are devices which use both mirrors (catioptric elements) and lenses (dioptric elements) to form images. Catadioptric cameras have been used in computer vision and robotics to obtain a wide field of view, often greater than 180°, unobtainable from perspective cameras. Applications of such devices include navigation, surveillance, and visualization, among others. Technically, certain matrices called catadioptric fundamental matrices come up. Geyer was able to give several equivalent characterizations of these matrices (see Chapter 5, Theorem 5.2). To my surprise, the Lorentz group $\text{O}(3,1)$ (of the theory of special relativity) comes up naturally! The set of fundamental matrices turns out to form a manifold $\mathcal{F}$, and the question then arises: What is the dimension of this manifold? Knowing the answer to this question is not only theoretically important but it is also practically very significant, because it tells us what are the “degrees of freedom” of the problem.

Chris Geyer found an elegant and beautiful answer using some rather sophisticated concepts from the theory of group actions and Lie groups (Theorem 5.10): The space $\mathcal{F}$ is
isomorphic to the quotient
\[ \mathbf{O}(3, 1) \times \mathbf{O}(3, 1)/\mathbf{H}_F, \]
where \( \mathbf{H}_F \) is the stabilizer of any element \( F \) in \( \mathcal{F} \). Now, it is easy to determine the dimension of \( \mathbf{H}_F \) by determining the dimension of its Lie algebra, which is 3. As \( \dim \mathbf{O}(3, 1) = 6 \), we find that \( \dim \mathcal{F} = 2 \cdot 6 - 3 = 9 \).

Of course, a certain amount of machinery is needed in order to understand how the above results are obtained: group actions, manifolds, Lie groups, homogenous spaces, Lorentz groups, etc. As most computer science students, even those specialized in computer vision or robotics, are not familiar with these concepts, we thought that it would be useful to give a fairly detailed exposition of these theories.

During the seminar, I also used some material from my book, Gallier [72], especially from Chapters 11, 12 and 14. Readers might find it useful to read some of this material beforehand or in parallel with these notes, especially Chapter 14, which gives a more elementary introduction to Lie groups and manifolds. For the reader’s convenience, I have incorporated a slightly updated version of chapter 14 from [72] as Chapters 1 and 4 of this manuscript. In fact, during the seminar, I lectured on most of Chapter 5, but only on the “gentler” versions of Chapters 7, 8, 15, as in [72], and not at all on Chapter 28, which was written after the course had ended.

One feature worth pointing out is that we give a complete proof of the surjectivity of the exponential map \( \text{exp}: \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3) \), for the Lorentz group \( \mathbf{SO}_0(3, 1) \) (see Section 6.2, Theorem 6.17). Although we searched the literature quite thoroughly, we did not find a proof of this specific fact (the physics books we looked at, even the most reputable ones, seem to take this fact as obvious, and there are also wrong proofs; see the Remark following Theorem 6.4).

We are aware of two proofs of the surjectivity of \( \text{exp}: \mathfrak{so}(1, n) \to \mathbf{SO}_0(1, n) \) in the general case where \( n \) is arbitrary: One due to Nishikawa [137] (1983), and an earlier one due to Marcel Riesz [145] (1957). In both cases, the proof is quite involved (40 pages or so). In the case of \( \mathbf{SO}_0(1, 3) \), a much simpler argument can be made using the fact that \( \varphi: \mathbf{SL}(2, \mathbb{C}) \to \mathbf{SO}_0(1, 3) \) is surjective and that its kernel is \( \{ I, -I \} \) (see Proposition 6.16). Actually, a proof of this fact is not easy to find in the literature either (and, beware there are wrong proofs, again see the Remark following Theorem 6.4). We have made sure to provide all the steps of the proof of the surjectivity of \( \text{exp}: \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3) \). For more on this subject, see the discussion in Section 6.2, after Corollary 6.13.

One of the “revelations” I had while on sabbatical in Nicholas’ group was that many of the data that radiologists deal with (for instance, “diffusion tensors”) do not live in Euclidean spaces, which are flat, but instead in more complicated curved spaces (Riemannian manifolds). As a consequence, even a notion as simple as the average of a set of data does not make sense in such spaces. Similarly, it is not clear how to define the covariance matrix of a random vector.
Pennec [139], among others, introduced a framework based on Riemannian Geometry for defining some basic statistical notions on curved spaces and gave some algorithmic methods to compute these basic notions. Based on work in Vincent Arsigny’s Ph.D. thesis, Arsigny, Fillard, Pennec and Ayache [8] introduced a new Lie group structure on the space of symmetric positive definite matrices, which allowed them to transfer standard statistical concepts to this space (abusively called “tensors.”) One of my goals in writing these notes is to provide a rather thorough background in differential geometry so that one will then be well prepared to read the above papers by Arsigny, Fillard, Pennec, Ayache and others, on statistics on manifolds.

At first, when I was writing these notes, I felt that it was important to supply most proofs. However, when I reached manifolds and differential geometry concepts, such as connections, geodesics and curvature, I realized that how formidable a task it was! Since there are lots of very good book on differential geometry, not without regrets, I decided that it was best to try to “demistify” concepts rather than fill many pages with proofs. However, when omitting a proof, I give precise pointers to the literature. In some cases where the proofs are really beautiful, as in the Theorem of Hopf and Rinow, Myers’ Theorem or the Cartan-Hadamard Theorem, I could not resist to supply complete proofs!

Experienced differential geometers may be surprised and perhaps even irritated by my selection of topics. I beg their forgiveness! Primarily, I have included topics that I felt would be useful for my purposes, and thus, I have omitted some topics found in all respectable differential geometry book (such as spaces of constant curvature). On the other hand, I have occasionally included topics because I found them particularly beautiful (such as characteristic classes), even though they do not seem to be of any use in medical imaging or computer vision.

In the past two years, I have also come to realize that Lie groups and homogeneous manifolds, especially naturally reductive ones, are two of the most important topics for their role in applications. It is remarkable that most familiar spaces, spheres, projective spaces, Grassmannian and Stiefel manifolds, symmetric positive definite matrices, are naturally reductive manifolds. Remarkably, they all arise from some suitable action of the rotation group $\text{SO}(n)$, a Lie group, who emerges as the master player. The machinery of naturaly reductive manifolds, and of symmetric spaces (which are even nicer!), makes it possible to compute explicitly in terms of matrices all the notions from differential geometry (Riemannian metrics, geodesics, etc.) that are needed to generalize optimization methods to Riemannian manifolds. The interplay between Lie groups, manifolds, and analysis, yields a particularly effective tool. I tried to explain in some detail how these theories all come together to yield such a beautiful and useful tool.

I also hope that readers with a more modest background will not be put off by the level of abstraction in some of the chapters, and instead will be inspired to read more about these concepts, including fibre bundles!

I have also included chapters that present material having significant practical applications. These include
1. Chapter 20, on constructing manifolds from gluing data, has applications to surface reconstruction from 3D meshes.

2. Chapter 19, on homogeneous reductive spaces and symmetric spaces, has applications to robotics, machine learning, and computer vision. For example, Stiefel and Grassmannian manifolds come up naturally. Furthermore, in these manifolds, it is possible to compute explicitly geodesics, Riemannian distances, gradients and Hessians. This makes it possible to actually extend optimization methods such as gradient descent and Newton's method to these manifolds. A very good source on these topics is Absil, Mahony and Sepulchre [2].

3. Chapter 18, on the "Log-Euclidean framework," has applications in medical imaging.

4. Chapter 26, on spherical harmonics, has applications in computer graphics and computer vision.

5. Section 27.1 of Chapter 27 has applications to optimization techniques on matrix manifolds.

6. Chapter 30, on Clifford algebras and spinors, has applications in robotics and computer graphics.

Of course, as anyone who attempts to write about differential geometry and Lie groups, I faced the dilemma of including or not including a chapter on differential forms. Given that our intended audience probably knows very little about them, I decided to provide a fairly detailed treatment, including a brief treatment of vector-valued differential forms. Of course, this made it necessary to review tensor products, exterior powers, etc., and I have included a rather extensive chapter on this material.

I must acknowledge my debt to two of my main sources of inspiration: Berger’s Panoramic View of Riemannian Geometry [19] and Milnor’s Morse Theory [125]. In my opinion, Milnor’s book is still one of the best references on basic differential geometry. His exposition is remarkably clear and insightful, and his treatment of the variational approach to geodesics is unsurpassed. We borrowed heavily from Milnor [125]. Since Milnor’s book is typeset in “ancient” typewritten format (1973!), readers might enjoy reading parts of it typeset in \LaTeX. I hope that the readers of these notes will be well prepared to read standard differential geometry texts such as do Carmo [60], Gallot, Hulin, Lafontaine [73] and O’Neill [138], but also more advanced sources such as Sakai [150], Petersen [140], Jost [99], Knapp [106], and of course Milnor [125].

The chapters or sections marked with the symbol \(\clubsuit\) contain material that is typically more specialized or more advanced, and they can be omitted upon first (or second) reading. Chapter 23 and its successors deal with more sophisticated material that requires additional technical machinery.
Acknowledgement: I would like to thank Eugenio Calabi, Chris Croke, Ron Donagi, David Harbater, Herman Gluck, Alexander Kirillov, Steve Shatz and Wolfgang Ziller for their encouragement, advice, inspiration and for what they taught me. I also thank Kostas Daniilidis, Spyridon Leonardos, Marcelo Siqueira, and Roberto Tron for reporting typos and for helpful comments.
Contents

1 The Matrix Exponential; Some Matrix Lie Groups 15
   1.1 The Exponential Map ............................................. 15
   1.2 Some Classical Lie Groups ..................................... 25
   1.3 Symmetric and Other Special Matrices ......................... 30
   1.4 Exponential of Some Complex Matrices ....................... 33
   1.5 Hermitian and Other Special Matrices ....................... 36
   1.6 The Lie Group $\text{SE}(n)$ and the Lie Algebra $\mathfrak{se}(n)$ ........ 37

2 Basic Analysis: Review of Series and Derivatives 43
   2.1 Series and Power Series of Matrices ......................... 43
   2.2 The Derivative of a Function Between Normed Spaces ........ 52
   2.3 Linear Vector Fields and the Exponential ................... 65
   2.4 The Adjoint Representations .................................. 69

3 A Review of Point Set Topology 77
   3.1 Topological Spaces ........................................... 77
   3.2 Continuous Functions, Limits ................................ 82
   3.3 Connected Sets ............................................... 86
   3.4 Compact Sets ................................................ 91
   3.5 Quotient Spaces .............................................. 95

4 Introduction to Manifolds and Lie Groups 101
   4.1 Introduction to Embedded Manifolds ......................... 101
   4.2 Linear Lie Groups ............................................. 119

5 Groups and Group Actions 143
   5.1 Basic Concepts of Groups ..................................... 143
   5.2 Group Actions and Homogeneous Spaces, I .................... 149
   5.3 The Grassmann and Stiefel Manifolds ....................... 169
   5.4 Topological Groups $\otimes$ ................................. 173
### Contents

6 The Lorentz Groups ⊗
- 6.1 The Lorentz Groups $O(n, 1)$, $SO(n, 1)$ and $SO_0(n, 1)$ ................. 181
- 6.2 The Lie Algebra of the Lorentz Group $SO_0(n, 1)$ .............................. 195
- 6.3 Polar Forms for Matrices in $O(p, q)$ ............................................. 213
- 6.4 Pseudo-Algebraic Groups ................................................................. 220
- 6.5 More on the Topology of $O(p, q)$ and $SO(p, q)$ ............................ 222

7 Manifolds, Tangent Spaces, Cotangent Spaces 225
- 7.1 Charts and Manifolds ................................................................. 225
- 7.2 Tangent Vectors, Tangent Spaces ................................................ 242
- 7.3 Tangent Vectors as Derivations .................................................... 246
- 7.4 Tangent and Cotangent Spaces Revisited ⊗ ..................................... 255
- 7.5 Tangent Maps ................................................................................. 261
- 7.6 Submanifolds, Immersions, Embeddings ......................................... 266

8 Vector Fields, Integral Curves, Flows 271
- 8.1 Tangent and Cotangent Bundles .................................................... 271
- 8.2 Vector Fields, Lie Derivative .......................................................... 275
- 8.3 Integral Curves, Flows, One-Parameter Groups .................................. 283
- 8.4 Log-Euclidean Polyaffine Transformations ....................................... 292
- 8.5 Fast Polyaffine Transforms ............................................................. 294

9 Partitions of Unity, Covering Maps ⊗ 297
- 9.1 Partitions of Unity ............................................................................. 297
- 9.2 Covering Maps and Universal Covering Manifolds ............................ 304

10 Riemannian Metrics, Riemannian Manifolds 313
- 10.1 Frames ......................................................................................... 313
- 10.2 Riemannian Metrics ..................................................................... 315

11 Connections on Manifolds 321
- 11.1 Connections on Manifolds ............................................................ 322
- 11.2 Parallel Transport ......................................................................... 326
- 11.3 Connections Compatible with a Metric .......................................... 329

12 Geodesics on Riemannian Manifolds 337
- 12.1 Geodesics, Local Existence and Uniqueness .................................. 338
- 12.2 The Exponential Map ..................................................................... 344
- 12.3 Complete Riemannian Manifolds, Hopf-Rinow, Cut Locus ............... 353
- 12.4 Convexity, Convexity Radius ......................................................... 359
- 12.5 The Calculus of Variations Applied to Geodesics ......................... 360
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Curvature in Riemannian Manifolds</td>
<td>369</td>
</tr>
<tr>
<td>13.1</td>
<td>The Curvature Tensor</td>
<td>370</td>
</tr>
<tr>
<td>13.2</td>
<td>Sectional Curvature</td>
<td>378</td>
</tr>
<tr>
<td>13.3</td>
<td>Ricci Curvature</td>
<td>382</td>
</tr>
<tr>
<td>13.4</td>
<td>The Second Variation Formula and the Index Form</td>
<td>385</td>
</tr>
<tr>
<td>13.5</td>
<td>Jacobi Fields and Conjugate Points</td>
<td>390</td>
</tr>
<tr>
<td>13.6</td>
<td>Jacobi Field Applications in Topology and Curvature</td>
<td>404</td>
</tr>
<tr>
<td>13.7</td>
<td>Cut Locus and Injectivity Radius: Some Properties</td>
<td>409</td>
</tr>
<tr>
<td>14</td>
<td>Isometries, Submersions, Killing Vector Fields</td>
<td>413</td>
</tr>
<tr>
<td>14.1</td>
<td>Isometries and Local Isometries</td>
<td>414</td>
</tr>
<tr>
<td>14.2</td>
<td>Riemannian Covering Maps</td>
<td>418</td>
</tr>
<tr>
<td>14.3</td>
<td>Riemannian Submersions</td>
<td>421</td>
</tr>
<tr>
<td>14.4</td>
<td>Isometries and Killing Vector Fields</td>
<td>425</td>
</tr>
<tr>
<td>15</td>
<td>Lie Groups, Lie Algebra, Exponential Map</td>
<td>429</td>
</tr>
<tr>
<td>15.1</td>
<td>Lie Groups and Lie Algebras</td>
<td>431</td>
</tr>
<tr>
<td>15.2</td>
<td>Left and Right Invariant Vector Fields, Exponential Map</td>
<td>435</td>
</tr>
<tr>
<td>15.3</td>
<td>Homomorphisms, Lie Subgroups</td>
<td>442</td>
</tr>
<tr>
<td>15.4</td>
<td>The Correspondence Lie Groups–Lie Algebras</td>
<td>445</td>
</tr>
<tr>
<td>15.5</td>
<td>Semidirect Products of Lie Algebras and Lie Groups</td>
<td>447</td>
</tr>
<tr>
<td>15.6</td>
<td>Universal Covering Groups</td>
<td>456</td>
</tr>
<tr>
<td>15.7</td>
<td>The Lie Algebra of Killing Fields</td>
<td>457</td>
</tr>
<tr>
<td>16</td>
<td>The Derivative of $\exp$ and Dynkin’s Formula</td>
<td>459</td>
</tr>
<tr>
<td>16.1</td>
<td>The Derivative of the Exponential Map</td>
<td>459</td>
</tr>
<tr>
<td>16.2</td>
<td>The Product in Logarithmic Coordinates</td>
<td>461</td>
</tr>
<tr>
<td>16.3</td>
<td>Dynkin’s Formula</td>
<td>462</td>
</tr>
<tr>
<td>17</td>
<td>Metrics, Connections, and Curvature on Lie Groups</td>
<td>465</td>
</tr>
<tr>
<td>17.1</td>
<td>Left (resp. Right) Invariant Metrics</td>
<td>466</td>
</tr>
<tr>
<td>17.2</td>
<td>Bi-Invariant Metrics</td>
<td>467</td>
</tr>
<tr>
<td>17.3</td>
<td>Connections and Curvature of Left-Invariant Metrics</td>
<td>474</td>
</tr>
<tr>
<td>17.4</td>
<td>Simple and Semisimple Lie Algebras and Lie Groups</td>
<td>485</td>
</tr>
<tr>
<td>17.5</td>
<td>The Killing Form</td>
<td>487</td>
</tr>
<tr>
<td>17.6</td>
<td>Left-Invariant Connections and Cartan Connections</td>
<td>494</td>
</tr>
<tr>
<td>18</td>
<td>The Log-Euclidean Framework</td>
<td>499</td>
</tr>
<tr>
<td>18.1</td>
<td>Introduction</td>
<td>499</td>
</tr>
<tr>
<td>18.2</td>
<td>A Lie-Group Structure on $\text{SPD}(n)$</td>
<td>500</td>
</tr>
<tr>
<td>18.3</td>
<td>Log-Euclidean Metrics on $\text{SPD}(n)$</td>
<td>501</td>
</tr>
<tr>
<td>18.4</td>
<td>A Vector Space Structure on $\text{SPD}(n)$</td>
<td>504</td>
</tr>
<tr>
<td>18.5</td>
<td>Log-Euclidean Means</td>
<td>504</td>
</tr>
</tbody>
</table>
19 Manifolds Arising from Group Actions 507
19.1 Proper Maps ........................................ 508
19.2 Proper and Free Actions .............................. 510
19.3 Riemannian Submersions and Coverings ⊗ ............... 513
19.4 Reductive Homogeneous Spaces ......................... 517
19.5 Examples of Reductive Homogeneous Spaces ................ 526
19.6 Naturally Reductive Homogeneous Spaces .................. 530
19.7 Examples of Naturally Reductive Homogeneous Spaces ........ 536
19.8 A Glimpse at Symmetric Spaces ......................... 543
19.9 Examples of Symmetric Spaces ......................... 548
19.10 Types of Symmetric Spaces ............................ 562

20 Construction of Manifolds From Gluing Data ⊗ 567
20.1 Sets of Gluing Data for Manifolds ....................... 567
20.2 Parametric Pseudo-Manifolds .......................... 576

21 Tensor Algebras 579
21.1 Linear Algebra Preliminaries: Dual Spaces and Pairings .... 580
21.2 Tensors Products ..................................... 585
21.3 Bases of Tensor Products ................................ 596
21.4 Some Useful Isomorphisms for Tensor Products ............. 598
21.5 Duality for Tensor Products ................................ 602
21.6 Tensor Algebras ........................................ 606
21.7 Symmetric Tensor Powers ............................... 612
21.8 Bases of Symmetric Powers ............................. 617
21.9 Some Useful Isomorphisms for Symmetric Powers ............ 620
21.10 Duality for Symmetric Powers ........................... 620
21.11 Symmetric Algebras .................................... 623
21.12 Tensor Products of Modules over a Commutative Ring ....... 625

22 Exterior Tensor Powers and Exterior Algebras 629
22.1 Exterior Tensor Powers .................................. 629
22.2 Bases of Exterior Powers ................................ 634
22.3 Some Useful Isomorphisms for Exterior Powers ............... 637
22.4 Duality for Exterior Powers ............................. 637
22.5 Exterior Algebras ........................................ 640
22.6 The Hodge *-Operator .................................. 644
22.7 Testing Decomposability; Left and Right Hooks ⊗ .......... 647
22.8 The Grassmann-Plücker’s Equations and Grassmannians ⊕ .... 658
22.9 Vector-Valued Alternating Forms ......................... 662
22.10 The Pfaffian Polynomial ................................. 666
## CONTENTS

### 23 Differential Forms

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.1</td>
<td>Differential Forms on $\mathbb{R}^n$ and de Rham Cohomology</td>
</tr>
<tr>
<td>23.2</td>
<td>Differential Forms on Manifolds</td>
</tr>
<tr>
<td>23.3</td>
<td>Lie Derivatives</td>
</tr>
<tr>
<td>23.4</td>
<td>Vector-Valued Differential Forms</td>
</tr>
<tr>
<td>23.5</td>
<td>Differential Forms on Lie Groups</td>
</tr>
</tbody>
</table>

### 24 Integration on Manifolds

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.1</td>
<td>Orientation of Manifolds</td>
</tr>
<tr>
<td>24.2</td>
<td>Volume Forms on Riemannian Manifolds and Lie Groups</td>
</tr>
<tr>
<td>24.3</td>
<td>Integration in $\mathbb{R}^n$</td>
</tr>
<tr>
<td>24.4</td>
<td>Integration on Manifolds</td>
</tr>
<tr>
<td>24.5</td>
<td>Manifolds With Boundary</td>
</tr>
<tr>
<td>24.6</td>
<td>Integration on Regular Domains and Stokes’ Theorem</td>
</tr>
<tr>
<td>24.7</td>
<td>Integration on Riemannian Manifolds and Lie Groups</td>
</tr>
</tbody>
</table>

### 25 Distributions and the Frobenius Theorem

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.1</td>
<td>Tangential Distributions, Involutive Distributions</td>
</tr>
<tr>
<td>25.2</td>
<td>Frobenius Theorem</td>
</tr>
<tr>
<td>25.3</td>
<td>Differential Ideals and Frobenius Theorem</td>
</tr>
<tr>
<td>25.4</td>
<td>A Glimpse at Foliations</td>
</tr>
</tbody>
</table>

### 26 Spherical Harmonics and Linear Representations

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.1</td>
<td>Hilbert Spaces and Hilbert Sums</td>
</tr>
<tr>
<td>26.2</td>
<td>Spherical Harmonics on the Circle</td>
</tr>
<tr>
<td>26.3</td>
<td>Spherical Harmonics on the 2-Sphere</td>
</tr>
<tr>
<td>26.4</td>
<td>The Laplace-Beltrami Operator</td>
</tr>
<tr>
<td>26.5</td>
<td>Harmonic Polynomials, Spherical Harmonics and $L^2(S^n)$</td>
</tr>
<tr>
<td>26.6</td>
<td>Reproducing Kernel and Zonal Spherical Functions</td>
</tr>
<tr>
<td>26.7</td>
<td>More on the Gegenbauer Polynomials</td>
</tr>
<tr>
<td>26.8</td>
<td>The Funk-Hecke Formula</td>
</tr>
<tr>
<td>26.9</td>
<td>Linear Representations of Compact Lie Groups</td>
</tr>
<tr>
<td>26.10</td>
<td>Gelfand Pairs, Spherical Functions, Fourier Transform ( \ast )</td>
</tr>
</tbody>
</table>

### 27 The Laplace-Beltrami Operator and Harmonic Forms

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.1</td>
<td>The Gradient and Hessian Operators</td>
</tr>
<tr>
<td>27.2</td>
<td>The Hodge $\ast$ Operator on Riemannian Manifolds</td>
</tr>
<tr>
<td>27.3</td>
<td>The Laplace-Beltrami and Divergence Operators</td>
</tr>
<tr>
<td>27.4</td>
<td>Harmonic Forms, the Hodge Theorem, Poincaré Duality</td>
</tr>
<tr>
<td>27.5</td>
<td>The Connection Laplacian and the Bochner Technique</td>
</tr>
</tbody>
</table>

---

671  723  769  783  851
28 Bundles, Metrics on Bundles, Homogeneous Spaces 875
  28.1 Fibre Bundles ................................................. 875
  28.2 Vector Bundles ................................................. 892
  28.3 Operations on Vector Bundles ................................. 899
  28.4 Metrics on Bundles, Reduction, Orientation ................ 902
  28.5 Principal Fibre Bundles ....................................... 907
  28.6 Proper and Free Actions, Homogeneous Spaces Revisited .... 914

29 Connections and Curvature in Vector Bundles 917
  29.1 Connections in Vector Bundles and Riemannian Manifolds .... 917
  29.2 Curvature and Curvature Form ................................ 927
  29.3 Parallel Transport ............................................. 934
  29.4 Connections Compatible with a Metric ......................... 937
  29.5 Duality between Vector Fields and Differential Forms ........ 946
  29.6 Pontrjagin Classes and Chern Classes, a Glimpse ............. 947
  29.7 Euler Classes and The Generalized Gauss-Bonnet Theorem .... 955

30 Clifford Algebras, Clifford Groups, Pin and Spin 959
  30.1 Introduction: Rotations As Group Actions ..................... 959
  30.2 Clifford Algebras .............................................. 961
  30.3 Clifford Groups .............................................. 971
  30.4 The Groups Pin(n) and Spin(n) ................................ 977
  30.5 The Groups Pin(p,q) and Spin(p,q) ............................ 983
  30.6 Periodicity of the Clifford Algebras Cl_{p,q} .................. 985
  30.7 The Complex Clifford Algebras Cl(n, C) ....................... 989
  30.8 The Groups Pin(p,q) and Spin(p,q) as double covers .......... 990
Chapter 1

The Matrix Exponential; Some Matrix Lie Groups

1.1 The Exponential Map

The purpose of this chapter and the next three is to give a “gentle” and fairly concrete introduction to manifolds, Lie groups and Lie algebras, our main objects of study.

Most texts on Lie groups and Lie algebras begin with prerequisites in differential geometry that are often formidable to average computer scientists (or average scientists, whatever that means!). We also struggled for a long time, trying to figure out what Lie groups and Lie algebras are all about, but this can be done! A good way to sneak into the wonderful world of Lie groups and Lie algebras is to play with explicit matrix groups such as the group of rotations in $\mathbb{R}^2$ (or $\mathbb{R}^3$) and with the exponential map. After actually computing the exponential $A = e^B$ of a $2 \times 2$ skew symmetric matrix $B$ and observing that it is a rotation matrix, and similarly for a $3 \times 3$ skew symmetric matrix $B$, one begins to suspect that there is something deep going on. Similarly, after the discovery that every real invertible $n \times n$ matrix $A$ can be written as $A = RP$, where $R$ is an orthogonal matrix and $P$ is a positive definite symmetric matrix, and that $P$ can be written as $P = e^S$ for some symmetric matrix $S$, one begins to appreciate the exponential map.

Our goal in this chapter is to give an elementary and concrete introduction to Lie groups and Lie algebras by studying a number of the so-called classical groups, such as the general linear group $\text{GL}(n, \mathbb{R})$, the special linear group $\text{SL}(n, \mathbb{R})$, the orthogonal group $\text{O}(n)$, the
special orthogonal group $\text{SO}(n)$, and the group of affine rigid motions $\text{SE}(n)$, and their Lie algebras $\mathfrak{gl}(n, \mathbb{R})$ (all matrices), $\mathfrak{sl}(n, \mathbb{R})$ (matrices with null trace), $\mathfrak{o}(n)$, and $\mathfrak{so}(n)$ (skew symmetric matrices). Now, Lie groups are at the same time, groups, topological spaces, and manifolds, so we will also have to introduce the crucial notion of a manifold.

The inventors of Lie groups and Lie algebras (starting with Lie!) regarded Lie groups as groups of symmetries of various topological or geometric objects. Lie algebras were viewed as the “infinitesimal transformations” associated with the symmetries in the Lie group. For example, the group $\text{SO}(n)$ of rotations is the group of orientation-preserving isometries of the Euclidean space $\mathbb{E}^n$. The Lie algebra $\mathfrak{so}(n, \mathbb{R})$ consisting of real skew symmetric $n \times n$ matrices is the corresponding set of infinitesimal rotations. The geometric link between a Lie group and its Lie algebra is the fact that the Lie algebra can be viewed as the tangent space to the Lie group at the identity. There is a map from the tangent space to the Lie group, called the exponential map. The Lie algebra can be considered as a linearization of the Lie group (near the identity element), and the exponential map provides the “delinearization,” i.e., it takes us back to the Lie group. These concepts have a concrete realization in the case of groups of matrices and, for this reason, we begin by studying the behavior of the exponential maps on matrices.

We begin by defining the exponential map on matrices and proving some of its properties. The exponential map allows us to “linearize” certain algebraic properties of matrices. It also plays a crucial role in the theory of linear differential equations with constant coefficients. But most of all, as we mentioned earlier, it is a stepping stone to Lie groups and Lie algebras. On the way to Lie algebras, we derive the classical “Rodrigues-like” formulae for rotations and for rigid motions in $\mathbb{R}^2$ and $\mathbb{R}^3$. We give an elementary proof that the exponential map is surjective for both $\text{SO}(n)$ and $\text{SE}(n)$, not using any topology, just certain normal forms for matrices (see Gallier [72], Chapters 12 and 13).

The last section gives a quick introduction to manifolds, Lie groups and Lie algebras. Rather than defining abstract manifolds in terms of charts, atlases, etc., we consider the special case of embedded submanifolds of $\mathbb{R}^N$. This approach has the pedagogical advantage of being more concrete since it uses parametrizations of subsets of $\mathbb{R}^N$, which should be familiar to the reader in the case of curves and surfaces. The general definition of a manifold will be given in Chapter 7.

Also, rather than defining Lie groups in full generality, we define linear Lie groups using the famous result of Cartan (apparently actually due to Von Neumann) that a closed subgroup of $\text{GL}(n, \mathbb{R})$ is a manifold, and thus a Lie group. This way, Lie algebras can be “computed” using tangent vectors to curves of the form $t \mapsto A(t)$, where $A(t)$ is a matrix. This section is inspired from Artin [10], Chevalley [41], Marsden and Ratiu [121], Curtis [46], Howe [95], and Sattinger and Weaver [154].

Given an $n \times n$ (real or complex) matrix $A = (a_{ij})$, we would like to define the exponential
1.1. THE EXPONENTIAL MAP

$e^A$ of $A$ as the sum of the series

$$e^A = I_n + \sum_{p \geq 1} \frac{A^p}{p!} = \sum_{p \geq 0} \frac{A^p}{p!},$$

letting $A^0 = I_n$. The problem is, Why is it well-defined? The following proposition shows that the above series is indeed absolutely convergent. For the definition of absolute convergence see Chapter 2, Section 1.

**Proposition 1.1.** Let $A = (a_{ij})$ be a (real or complex) $n \times n$ matrix, and let

$$\mu = \max\{|a_{ij}| \mid 1 \leq i, j \leq n\}.$$

If $A^p = (a^{(p)}_{ij})$, then

$$|a^{(p)}_{ij}| \leq (n\mu)^p$$

for all $i, j$, $1 \leq i, j \leq n$. As a consequence, the $n^2$ series

$$\sum_{p \geq 0} \frac{a^{(p)}_{ij}}{p!}$$

converge absolutely, and the matrix

$$e^A = \sum_{p \geq 0} \frac{A^p}{p!}$$

is a well-defined matrix.

**Proof.** The proof is by induction on $p$. For $p = 0$, we have $A^0 = I_n$, $(n\mu)^0 = 1$, and the proposition is obvious. Assume that

$$|a^{(p)}_{ij}| \leq (n\mu)^p$$

for all $i, j$, $1 \leq i, j \leq n$. Then we have

$$|a^{(p+1)}_{ij}| = \left| \sum_{k=1}^{n} a^{(p)}_{ik} a_{kj} \right| \leq \sum_{k=1}^{n} |a^{(p)}_{ik}| |a_{kj}| \leq \mu \sum_{k=1}^{n} |a^{(p)}_{ik}| \leq n\mu (n\mu)^p = (n\mu)^{p+1},$$

for all $i, j$, $1 \leq i, j \leq n$. For every pair $(i, j)$ such that $1 \leq i, j \leq n$, since

$$|a^{(p)}_{ij}| \leq (n\mu)^p,$$

the series

$$\sum_{p \geq 0} \frac{|a^{(p)}_{ij}|}{p!}$$
is bounded by the convergent series
\[ e^{n\mu} = \sum_{p \geq 0} \frac{(n\mu)^p}{p!}, \]
and thus it is absolutely convergent. This shows that
\[ e^A = \sum_{k \geq 0} \frac{A^k}{k!} \]
is well defined. \( \square \)

It is instructive to compute explicitly the exponential of some simple matrices. As an example, let us compute the exponential of the real skew symmetric matrix
\[ A = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}. \]
We need to find an inductive formula expressing the powers \( A^n \). Let us observe that
\[ \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}^2 = -\theta^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \]
Then, letting
\[ J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \]
we have
\[ A^{4n} = \theta^{4n} I_2, \]
\[ A^{4n+1} = \theta^{4n+1} J, \]
\[ A^{4n+2} = -\theta^{4n+2} I_2, \]
\[ A^{4n+3} = -\theta^{4n+3} J, \]
and so
\[ e^A = I_2 + \frac{\theta}{1!} J - \frac{\theta^2}{2!} I_2 - \frac{\theta^3}{3!} J + \frac{\theta^4}{4!} I_2 + \frac{\theta^5}{5!} J - \frac{\theta^6}{6!} I_2 - \frac{\theta^7}{7!} J + \cdots. \]
Rearranging the order of the terms, we have
\[ e^A = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) I_2 + \left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right) J. \]
We recognize the power series for \( \cos \theta \) and \( \sin \theta \), and thus
\[ e^A = \cos \theta I_2 + \sin \theta J, \]
1.1. THE EXPONENTIAL MAP

that is
\[ e^A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. \]

Thus, \( e^A \) is a rotation matrix! This is a general fact. If \( A \) is a skew symmetric matrix, then \( e^A \) is an orthogonal matrix of determinant +1, i.e., a rotation matrix. Furthermore, every rotation matrix is of this form; i.e., the exponential map from the set of skew symmetric matrices to the set of rotation matrices is surjective. In order to prove these facts, we need to establish some properties of the exponential map.

But before that, let us work out another example showing that the exponential map is not always surjective. Let us compute the exponential of a real \( 2 \times 2 \) matrix with null trace of the form
\[ A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}. \]

We need to find an inductive formula expressing the powers \( A^n \). Observe that
\[ A^2 = (a^2 + bc)I_2 = -\det(A)I_2. \]

If \( a^2 + bc = 0 \), we have
\[ e^A = I_2 + A. \]

If \( a^2 + bc < 0 \), let \( \omega > 0 \) be such that \( \omega^2 = -(a^2 + bc) \). Then, \( A^2 = -\omega^2 I_2 \). We get
\[ e^A = I_2 + \frac{A}{1!} - \frac{\omega^2}{2!} I_2 - \frac{\omega^2}{3!} A + \frac{\omega^4}{4!} I_2 + \frac{\omega^4}{5!} A - \frac{\omega^6}{6!} I_2 - \frac{\omega^6}{7!} A + \cdots. \]

Rearranging the order of the terms, we have
\[ e^A = \left( 1 - \frac{\omega^2}{2!} + \frac{\omega^4}{4!} - \frac{\omega^6}{6!} + \cdots \right) I_2 + \frac{1}{\omega} \left( \omega - \frac{\omega^3}{3!} + \frac{\omega^5}{5!} - \frac{\omega^7}{7!} + \cdots \right) A. \]

We recognize the power series for \( \cos \omega \) and \( \sin \omega \), and thus
\[ e^A = \cos \omega I_2 + \frac{\sin \omega}{\omega} A = \begin{pmatrix} \cos \omega + \frac{\sin \omega}{\omega} a & \frac{\sin \omega}{\omega} b \\ \frac{\sin \omega}{\omega} c & \cos \omega - \frac{\sin \omega}{\omega} a \end{pmatrix}. \]

Note that
\[ \det(e^A) = \left( \cos \omega + \frac{\sin \omega}{\omega} a \right) \left( \cos \omega - \frac{\sin \omega}{\omega} a \right) - \frac{\sin^2 \omega}{\omega^2} bc \]
\[ = \cos^2 \omega - \frac{\sin^2 \omega}{\omega^2} (a^2 + bc) = \cos^2 \omega + \sin^2 \omega = 1. \]

If \( a^2 + bc > 0 \), let \( \omega > 0 \) be such that \( \omega^2 = a^2 + bc \). Then \( A^2 = \omega^2 I_2 \). We get
\[ e^A = I_2 + \frac{A}{1!} + \frac{\omega^2}{2!} I_2 + \frac{\omega^2}{3!} A + \frac{\omega^4}{4!} I_2 + \frac{\omega^4}{5!} A + \frac{\omega^6}{6!} I_2 + \frac{\omega^6}{7!} A + \cdots. \]
Rearranging the order of the terms, we have
\[ e^A = \left( 1 + \frac{\omega^2}{2!} + \frac{\omega^4}{4!} + \frac{\omega^6}{6!} + \cdots \right) I_2 + \frac{1}{\omega} \left( \frac{\omega^3}{3!} + \frac{\omega^5}{5!} + \frac{\omega^7}{7!} + \cdots \right) A. \]

If we recall that \( \cosh \omega = \left( e^\omega + e^{-\omega} \right)/2 \) and \( \sinh \omega = \left( e^\omega - e^{-\omega} \right)/2 \), we recognize the power series for \( \cosh \omega \) and \( \sinh \omega \), and thus
\[ e^A = \cosh \omega I_2 + \frac{\sinh \omega}{\omega} A = \left( \begin{array}{cc} \cosh \omega + \frac{\sinh \omega}{\omega} a & \sinh \omega b \\ \frac{\sinh \omega}{\omega} c & \cosh \omega - \frac{\sinh \omega}{\omega} a \end{array} \right), \]
and
\[ \det(e^A) = \left( \cosh \omega + \frac{\sinh \omega}{\omega} a \right) \left( \cosh \omega - \frac{\sinh \omega}{\omega} a \right) - \frac{\sinh^2 \omega}{\omega^2} bc = \cosh^2 \omega - \frac{\sinh^2 \omega}{\omega^2} (a^2 + bc) = \cosh^2 \omega - \sinh^2 \omega = 1. \]

In both cases
\[ \det(e^A) = 1. \]

This shows that the exponential map is a function from the set of \( 2 \times 2 \) matrices with null trace to the set of \( 2 \times 2 \) matrices with determinant 1. This function is not surjective. Indeed, \( \text{tr}(e^A) = 2 \cos \omega \) when \( a^2 + bc < 0 \), \( \text{tr}(e^A) = 2 \cosh \omega \) when \( a^2 + bc > 0 \), and \( \text{tr}(e^A) = 2 \) when \( a^2 + bc = 0 \). As a consequence, for any matrix \( A \) with null trace,
\[ \text{tr}(e^A) \geq -2, \]
and any matrix \( B \) with determinant 1 and whose trace is less than \(-2\) is not the exponential \( e^A \) of any matrix \( A \) with null trace. For example,
\[ B = \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}, \]
where \( a < 0 \) and \( a \neq -1 \), is not the exponential of any matrix \( A \) with null trace, since
\[ \frac{(a + 1)^2}{a} = \frac{a^2 + 2a + 1}{a} = \frac{a^2 + 1}{a} + 2 < 0, \]
which in turn implies \( \text{tr}(B) = a + \frac{1}{a} = a^2 + \frac{1}{a} < -2 \).

A fundamental property of the exponential map is that if \( \lambda_1, \ldots, \lambda_n \) are the eigenvalues of \( A \), then the eigenvalues of \( e^A \) are \( e^{\lambda_1}, \ldots, e^{\lambda_n} \). For this we need two propositions.

**Proposition 1.2.** Let \( A \) and \( U \) be (real or complex) matrices, and assume that \( U \) is invertible. Then
\[ e^{UAU^{-1}} = U e^A U^{-1}. \]
Proof. A trivial induction shows that
\[ UAP^{p}U^{-1} = (UA^{p}U^{-1})^{p}, \]
and thus
\[ e^{UA^{p}U^{-1}} = \sum_{p \geq 0} \frac{(UA^{p}U^{-1})^{p}}{p!} = \sum_{p \geq 0} \frac{UAP^{p}U^{-1}}{p!} = U \left( \sum_{p \geq 0} \frac{A^{p}}{p!} \right) U^{-1} = U e^{A} U^{-1}. \]

Say that a square matrix \( A \) is an upper triangular matrix if it has the following shape,
\[
\begin{pmatrix}
a_{11} & a_{12} & a_{13} & \ldots & a_{1,n-1} & a_{1,n} \\
0 & a_{22} & a_{23} & \ldots & a_{2,n-1} & a_{2,n} \\
0 & 0 & a_{33} & \ldots & a_{3,n-1} & a_{3,n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & a_{n-1,n-1} & a_{n-1,n} \\
0 & 0 & 0 & \ldots & 0 & a_{nn}
\end{pmatrix},
\]
i.e., \( a_{ij} = 0 \) whenever \( j < i, 1 \leq i, j \leq n \).

**Proposition 1.3.** Given any complex \( n \times n \) matrix \( A \), there is an invertible matrix \( P \) and an upper triangular matrix \( T \) such that
\[ A = PTP^{-1}. \]

Proof. We prove by induction on \( n \) that if \( f: \mathbb{C}^{n} \rightarrow \mathbb{C}^{n} \) is a linear map, then there is a basis \((u_{1}, \ldots, u_{n})\) with respect to which \( f \) is represented by an upper triangular matrix. For \( n = 1 \) the result is obvious. If \( n > 1 \), since \( \mathbb{C} \) is algebraically closed, \( f \) has some eigenvalue \( \lambda_{1} \in \mathbb{C} \), and let \( u_{1} \) be an eigenvector for \( \lambda_{1} \). We can find \( n - 1 \) vectors \((v_{2}, \ldots, v_{n})\) such that \((u_{1}, v_{2}, \ldots, v_{n})\) is a basis of \( \mathbb{C}^{n} \), and let \( W \) be the subspace of dimension \( n - 1 \) spanned by \((v_{2}, \ldots, v_{n})\). In the basis \((u_{1}, v_{2}, \ldots, v_{n})\), the matrix of \( f \) is of the form
\[
\begin{pmatrix}
a_{11} & a_{12} & \ldots & a_{1,n} \\
0 & a_{22} & \ldots & a_{2,n} \\
\vdots & \vdots & \ddots & \vdots \\
0 & a_{n,2} & \ldots & a_{nn}
\end{pmatrix},
\]
since its first column contains the coordinates of \( \lambda_{1}u_{1} \) over the basis \((u_{1}, v_{2}, \ldots, v_{n})\). Letting \( p: \mathbb{C}^{n} \rightarrow W \) be the projection defined such that \( p(u_{1}) = 0 \) and \( p(v_{i}) = v_{i} \) when \( 2 \leq i \leq n \),
the linear map \( g: W \to W \) defined as the restriction of \( p \circ f \) to \( W \) is represented by the \((n-1) \times (n-1)\) matrix \((a_{ij})_{2 \leq i,j \leq n}\) over the basis \((v_2, \ldots, v_n)\). By the induction hypothesis, there is a basis \((u_2, \ldots, u_n)\) of \( W \) such that \( g \) is represented by an upper triangular matrix \((b_{ij})_{1 \leq i,j \leq n-1}\).

However,

\[ \mathbb{C}^n = \mathbb{C}u_1 \oplus W, \]

and thus \((u_1, \ldots, u_n)\) is a basis for \( \mathbb{C}^n \). Since \( p \) is the projection from \( \mathbb{C}^n = \mathbb{C}u_1 \oplus W \) onto \( W \) and \( g: W \to W \) is the restriction of \( p \circ f \) to \( W \), we have

\[ f(u_1) = \lambda_1 u_1 \]

and

\[ f(u_{i+1}) = a_{1i} u_1 + \sum_{j=1}^{n-1} b_{ij} u_{j+1} \]

for some \( a_{1i} \in \mathbb{C} \), when \( 1 \leq i \leq n-1 \). But then the matrix of \( f \) with respect to \((u_1, \ldots, u_n)\) is upper triangular. Thus, there is a change of basis matrix \( P \) such that \( A = PT P^{-1} \) where \( T \) is upper triangular.

Remark: If \( E \) is a Hermitian space, the proof of Proposition 1.3 can be easily adapted to prove that there is an orthonormal basis \((u_1, \ldots, u_n)\) with respect to which the matrix of \( f \) is upper triangular. In terms of matrices, this means that there is a unitary matrix \( U \) and an upper triangular matrix \( T \) such that \( A = UT U^* \). This is usually known as Schur's lemma. Using this result, we can immediately rederive the fact that if \( A \) is a Hermitian matrix, i.e \( A = A^* \), then there is a unitary matrix \( U \) and a real diagonal matrix \( D \) such that \( A = UDU^* \).

If \( A = PT P^{-1} \) where \( T \) is upper triangular, then \( A \) and \( T \) have the same characteristic polynomial. This is because if \( A \) and \( B \) are any two matrices such that \( A = PBP^{-1} \), then

\[
\det(A - \lambda I) = \det(PBP^{-1} - \lambda IP^{-1}),
\]

\[
= \det(P(B - \lambda I)P^{-1}),
\]

\[
= \det(P) \det(B - \lambda I) \det(P^{-1}),
\]

\[
= \det(P) \det(B - \lambda I) \det(P)^{-1},
\]

\[
= \det(B - \lambda I).
\]

Furthermore, it is well known that the determinant of a matrix of the form

\[
\begin{pmatrix}
\lambda_1 - \lambda & a_{12} & a_{13} & \ldots & a_{1n-1} & a_{1n} \\
0 & \lambda_2 - \lambda & a_{23} & \ldots & a_{2n-1} & a_{2n} \\
0 & 0 & \lambda_3 - \lambda & \ldots & a_{3n-1} & a_{3n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & \lambda_{n-1} - \lambda & a_{n-1n} \\
0 & 0 & 0 & \ldots & 0 & \lambda_n - \lambda
\end{pmatrix}
\]
is \((\lambda_1 - \lambda) \cdots (\lambda_n - \lambda)\), and thus the eigenvalues of \(A = PTP^{-1}\) are the diagonal entries of \(T\). We use this property to prove the following proposition.

**Proposition 1.4.** Given any complex \(n \times n\) matrix \(A\), if \(\lambda_1, \ldots, \lambda_n\) are the eigenvalues of \(A\), then \(e^{\lambda_1}, \ldots, e^{\lambda_n}\) are the eigenvalues of \(e^A\). Furthermore, if \(u\) is an eigenvector of \(A\) for \(\lambda_i\), then \(u\) is an eigenvector of \(e^A\) for \(e^{\lambda_i}\).

**Proof.** By Proposition 1.3 there is an invertible matrix \(P\) and an upper triangular matrix \(T\) such that \(A = PTP^{-1}\).

By Proposition 1.2, \(e^{PTP^{-1}} = Pe^TP^{-1}\).

Note that \(e^T = \sum_{p \geq 0} \frac{T^p}{p!}\) is upper triangular since \(T^p\) is upper triangular for all \(p \geq 0\). If \(\lambda_1, \lambda_2, \ldots, \lambda_n\) are the diagonal entries of \(T\), the properties of matrix multiplication, when combined with an induction on \(p\), imply that the diagonal entries of \(T^p\) are \(\lambda_1^p, \lambda_2^p, \ldots, \lambda_n^p\). This in turn implies that the diagonal entries of \(e^T\) are \(\sum_{p \geq 0} \frac{\lambda_i^p}{p!} = e^{\lambda_i}\) for \(i \leq i \leq n\). In the preceding paragraph we showed that \(A\) and \(T\) have the same eigenvalues, which are the diagonal entries \(\lambda_1, \ldots, \lambda_n\) of \(T\). Since \(e^A = e^{PTP^{-1}} = Pe^TP^{-1}\), and \(e^T\) is upper triangular, we use the same argument to conclude that both \(e^A\) and \(e^T\) have the same eigenvalues, which are the diagonal entries of \(e^T\), where the diagonal entries of \(e^T\) are of the form \(e^{\lambda_1}, \ldots, e^{\lambda_n}\).

Now, if \(u\) is an eigenvector of \(A\) for the eigenvalue \(\lambda\), a simple induction shows that \(u\) is an eigenvector of \(A^n\) for the eigenvalue \(\lambda^n\), from which is follows that

\[
e^A = \left[ I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots \right] u = u + Au + \frac{A^2}{2!}u + \frac{A^3}{3!}u + \ldots
\]

\[
= u + \lambda u + \frac{\lambda^2}{2!}u + \frac{\lambda^3}{3!}u + \cdots = \left[ 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \right] u = e^{\lambda}u,
\]

which shows that \(u\) is an eigenvector of \(e^A\) for \(e^{\lambda}\). \(\square\)

As a consequence, we can show that

\[
\det(e^A) = e^{tr(A)},
\]

where \(tr(A)\) is the *trace of \(A\)*, i.e., the sum \(a_{11} + \cdots + a_{nn}\) of its diagonal entries, which is also equal to the sum of the eigenvalues of \(A\). This is because the determinant of a matrix is equal to the product of its eigenvalues, and if \(\lambda_1, \ldots, \lambda_n\) are the eigenvalues of \(A\), then by Proposition 1.4, \(e^{\lambda_1}, \ldots, e^{\lambda_n}\) are the eigenvalues of \(e^A\), and thus

\[
\det \left( e^A \right) = e^{\lambda_1} \cdots e^{\lambda_n} = e^{\lambda_1 + \cdots + \lambda_n} = e^{tr(A)}.
\]

This shows that \(e^A\) is always an invertible matrix, since \(e^z\) is never null for every \(z \in \mathbb{C}\). In
fact, the inverse of $e^A$ is $e^{-A}$, but we need to prove another proposition. This is because it is generally not true that 
\[ e^{A+B} = e^A e^B, \]
unless $A$ and $B$ commute, i.e., $AB = BA$. We need to prove this last fact.

**Proposition 1.5.** Given any two complex $n \times n$ matrices $A, B$, if $AB = BA$, then 
\[ e^{A+B} = e^A e^B. \]

**Proof.** Since $AB = BA$, we can expand $(A + B)^p$ using the binomial formula:
\[ (A + B)^p = \sum_{k=0}^{p} \binom{p}{k} A^k B^{p-k}, \]
and thus
\[ \frac{1}{p!} (A + B)^p = \sum_{k=0}^{p} \frac{A^k B^{p-k}}{k!(p-k)!}. \]

Note that for any integer $N \geq 0$, we can write
\[ \sum_{p=0}^{2N} \frac{1}{p!} (A + B)^p = \sum_{p=0}^{2N} \sum_{k=0}^{p} \frac{A^k B^{p-k}}{k!(p-k)!}. \]

where there are $N(N+1)$ pairs $(k, l)$ in the second term. Letting
\[ \|A\| = \max \{|a_{ij}| \mid 1 \leq i, j \leq n\}, \quad \|B\| = \max \{|b_{ij}| \mid 1 \leq i, j \leq n\}, \]
and $\mu = \max(\|A\|, \|B\|)$, note that for every entry $c_{ij}$ in $(A^k/k!) (B^l/l!)$, the first inequality of Proposition 1.1, along with the fact that $N < \max(k, l)$ and $k + l \leq 2N$, implies that
\[ |c_{ij}| \leq n \frac{(n\mu)^k}{k!} \frac{(n\mu)^l}{l!} \leq \frac{n(n\mu)^{k+l}}{k! l!} \leq \frac{(n^2\mu)^{k+l}}{k! l!} \leq \frac{(n^2\mu)^{2N}}{N!}. \]

As a consequence, the absolute value of every entry in
\[ \sum_{\max(k, l) > N}^{\max(k, l) \leq 2N} \frac{A^k B^l}{k! l!} \]
is bounded by
\[ N(N+1) \frac{(n^2\mu)^{2N}}{N!}, \]
which goes to 0 as \( N \to \infty \). To see why this is the case, note that

\[
\lim_{N \to \infty} N(N + 1) \frac{(n^2 \mu)^{2N}}{N!} = \lim_{N \to \infty} N(N + 1) \frac{(n^2 \mu)^{2N}}{N(N - 1)(N - 2)!} = \lim_{N \to \infty} \frac{(n^4 \mu^2)^{N-2}}{(N - 2)!} = 0,
\]

where the last equality follows from the well known identity \( \lim_{N \to \infty} \frac{x^N}{N!} = 0 \). From this it immediately follows that

\[
e^{A+B} = e^A e^B.
\]

Now, using Proposition 1.5, since \( A \) and \( -A \) commute, we have

\[
e^A e^{-A} = e^{A-A} = e^0 = I_n,
\]

which shows that the inverse of \( e^A \) is \( e^{-A} \).

We will now use the properties of the exponential that we have just established to show how various matrices can be represented as exponentials of other matrices.

### 1.2 The Lie Groups \( \text{GL}(n, \mathbb{R}), \text{SL}(n, \mathbb{R}), \text{O}(n), \text{SO}(n), \text{the Lie Algebras } \mathfrak{gl}(n, \mathbb{R}), \mathfrak{sl}(n, \mathbb{R}), \mathfrak{o}(n), \mathfrak{so}(n), \text{and the Exponential Map} \)

First, we recall some basic facts and definitions. The set of real invertible \( n \times n \) matrices forms a group under multiplication, denoted by \( \text{GL}(n, \mathbb{R}) \). The subset of \( \text{GL}(n, \mathbb{R}) \) consisting of those matrices having determinant +1 is a subgroup of \( \text{GL}(n, \mathbb{R}) \), denoted by \( \text{SL}(n, \mathbb{R}) \). It is also easy to check that the set of real \( n \times n \) orthogonal matrices forms a group under multiplication, denoted by \( \text{O}(n) \). The subset of \( \text{O}(n) \) consisting of those matrices having determinant +1 is a subgroup of \( \text{O}(n) \), denoted by \( \text{SO}(n) \). We will also call matrices in \( \text{SO}(n) \) rotation matrices. Staying with easy things, we can check that the set of real \( n \times n \) matrices with null trace forms a vector space under addition, and similarly for the set of skew symmetric matrices.

**Definition 1.1.** The group \( \text{GL}(n, \mathbb{R}) \) is called the *general linear group*, and its subgroup \( \text{SL}(n, \mathbb{R}) \) is called the *special linear group*. The group \( \text{O}(n) \) of orthogonal matrices is called the *orthogonal group*, and its subgroup \( \text{SO}(n) \) is called the *special orthogonal group* (or group of rotations). The vector space of real \( n \times n \) matrices with null trace is denoted by \( \mathfrak{sl}(n, \mathbb{R}) \), and the vector space of real \( n \times n \) skew symmetric matrices is denoted by \( \mathfrak{so}(n) \).
Remark: The notation $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{so}(n)$ is rather strange and deserves some explanation. The groups $\text{GL}(n, \mathbb{R})$, $\text{SL}(n, \mathbb{R})$, $\text{O}(n)$, and $\text{SO}(n)$ are more than just groups. They are also topological groups, which means that they are topological spaces (viewed as subspaces of $\mathbb{R}^{n^2}$) and that the multiplication and the inverse operations are continuous (in fact, smooth). Furthermore, they are smooth real manifolds. Such objects are called Lie groups. The real vector spaces $\mathfrak{sl}(n)$ and $\mathfrak{so}(n)$ are what is called Lie algebras. However, we have not defined the algebra structure on $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{so}(n)$ yet. The algebra structure is given by what is called the Lie bracket, which is defined as

$$[A, B] = AB - BA.$$ 

Lie algebras are associated with Lie groups. What is going on is that the Lie algebra of a Lie group is its tangent space at the identity, i.e., the space of all tangent vectors at the identity (in this case, $I_n$). In some sense, the Lie algebra achieves a “linearization” of the Lie group. The exponential map is a map from the Lie algebra to the Lie group, for example,

$$\exp: \mathfrak{so}(n) \to \text{SO}(n)$$

and

$$\exp: \mathfrak{sl}(n, \mathbb{R}) \to \text{SL}(n, \mathbb{R}).$$

The exponential map often allows a parametrization of the Lie group elements by simpler objects, the Lie algebra elements.

One might ask, What happened to the Lie algebras $\mathfrak{gl}(n, \mathbb{R})$ and $\mathfrak{o}(n)$ associated with the Lie groups $\text{GL}(n, \mathbb{R})$ and $\text{O}(n)$? We will see later that $\mathfrak{gl}(n, \mathbb{R})$ is the set of all real $n \times n$ matrices, and that $\mathfrak{o}(n) = \mathfrak{so}(n)$.

The properties of the exponential map play an important role in studying a Lie group. For example, it is clear that the map

$$\exp: \mathfrak{gl}(n, \mathbb{R}) \to \text{GL}(n, \mathbb{R})$$

is well-defined, but since $\det(e^A) = e^{\text{tr}(A)}$, every matrix of the form $e^A$ has a positive determinant and exp is not surjective. Similarly, the fact $\det(e^A) = e^{\text{tr}(A)}$ implies that the map

$$\exp: \mathfrak{sl}(n, \mathbb{R}) \to \text{SL}(n, \mathbb{R})$$

is well-defined. However, we showed in Section 1.1 that it is not surjective either. As we will see in the next theorem, the map

$$\exp: \mathfrak{so}(n) \to \text{SO}(n)$$

is well-defined and surjective. The map

$$\exp: \mathfrak{o}(n) \to \text{O}(n)$$

\[1\] We refrain from defining manifolds right now, not to interrupt the flow of intuitive ideas.
is well-defined, but it is not surjective, since there are matrices in $O(n)$ with determinant $-1$.

**Remark:** The situation for matrices over the field $\mathbb{C}$ of complex numbers is quite different, as we will see later.

We now show the fundamental relationship between $SO(n)$ and $\mathfrak{so}(n)$.

**Theorem 1.6.** The exponential map

$$\exp: \mathfrak{so}(n) \to SO(n)$$

is well-defined and surjective.

**Proof.** First, we need to prove that if $A$ is a skew symmetric matrix, then $e^A$ is a rotation matrix. For this, we quickly check that

$$(e^A)^\top = e^{A^\top}.$$  

This is consequence of the definition $e^A = \sum_{p\geq 0} \frac{A^p}{p!}$ as a absolutely convergent series, the observation that $(A^p)^\top = (A^\top)^p$, and the linearity of the transpose map, i.e. $(A + B)^\top = A^\top + B^\top$. Then, since $A^\top = -A$, we get

$$(e^A)^\top = e^{A^\top} = e^{-A},$$

and so

$$(e^A)^\top e^A = e^{-A} e^A = e^{-A + A} = e^0 = I_n,$$

and similarly,

$$e^A (e^A)^\top = I_n,$$

showing that $e^A$ is orthogonal. Also,

$$\det(e^A) = e^{\text{tr}(A)},$$

and since $A$ is real skew symmetric, its diagonal entries are 0, i.e., $\text{tr}(A) = 0$, and so $\det(e^A) = +1$.

For the surjectivity, we use Theorem 12.5, from Chapter 12 of Gallier [72]. Theorem 12.5 says that for every orthogonal matrix $R$ there is an orthogonal matrix $P$ such that $R = PE P^\top$, where $E$ is a block diagonal matrix of the form

$$E = \begin{pmatrix} E_1 & \cdots \\ E_2 & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & E_p \end{pmatrix},$$
such that each block $E_i$ is either 1, $-1$, or a two-dimensional matrix of the form

$$E_i = \begin{pmatrix} \cos \theta_i & -\sin \theta_i \\ \sin \theta_i & \cos \theta_i \end{pmatrix},$$

with $0 < \theta_i < \pi$. Furthermore, if $R$ is a rotation matrix, then we may assume that $0 < \theta_i \leq \pi$ and that the scalar entries are $+1$. Then, we can form the block diagonal matrix

$$D = \begin{pmatrix} D_1 & & \\ & \ddots & \\ & & D_p \end{pmatrix}$$

such that each block $D_i$ is either 0 when $E_i$ consists of $+1$, or the two-dimensional matrix

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix}$$

when

$$E_i = \begin{pmatrix} \cos \theta_i & -\sin \theta_i \\ \sin \theta_i & \cos \theta_i \end{pmatrix},$$

and we let $A = PD P^\top$. It is clear that $A$ is skew symmetric. Since by Proposition 1.2,

$$e^A = e^{PD P^{-1}} = Pe^D P^{-1},$$

and since $D$ is a block diagonal matrix, we can compute $e^D$ by computing the exponentials of its blocks. If $D_i = 0$, we get $E_i = e^0 = +1$, and if

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix},$$

we showed earlier that

$$e^{D_i} = \begin{pmatrix} \cos \theta_i & -\sin \theta_i \\ \sin \theta_i & \cos \theta_i \end{pmatrix},$$

exactly the block $E_i$. Thus, $E = e^D$, and as a consequence,

$$e^A = e^{PD P^{-1}} = Pe^D P^{-1} = PEP^{-1} = PE P^\top = R.$$  

This shows the surjectivity of the exponential. \qed

When $n = 3$ (and $A$ is skew symmetric), it is possible to work out an explicit formula for $e^A$. For any $3 \times 3$ real skew symmetric matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$
letting \( \theta = \sqrt{a^2 + b^2 + c^2} \) and

\[
B = \begin{pmatrix}
a^2 & ab & ac \\
ab & b^2 & bc \\
ac & bc & c^2
\end{pmatrix},
\]

we have the following result known as Rodrigues’s formula (1840).

**Proposition 1.7.** The exponential map \( \exp: \mathfrak{so}(3) \to SO(3) \) is given by

\[
e^A = \cos \theta I_3 + \frac{\sin \theta}{\theta} A + \left(1 - \cos \theta\right) \frac{1}{\theta^2} B,
\]

or, equivalently, by

\[
e^A = I_3 + \frac{\sin \theta}{\theta} A + \left(1 - \cos \theta\right) \frac{1}{\theta^2} A^2
\]

if \( \theta \neq 0 \), with \( e^{03} = I_3 \).

**Proof sketch.** First, observe that

\[
A^2 = -\theta^2 I_3 + B,
\]

since

\[
A^2 = \begin{pmatrix}
0 & -c & b \\
c & 0 & -a \\
-b & a & 0
\end{pmatrix}
\begin{pmatrix}
0 & -c & b \\
c & 0 & -a \\
-b & a & 0
\end{pmatrix} = \begin{pmatrix}
-c^2 - b^2 & ba & ca \\
ab & -c^2 - a^2 & cb \\
ac & cb & -b^2 - a^2
\end{pmatrix}
\]

\[
= \begin{pmatrix}
-a^2 - b^2 - c^2 & 0 & 0 \\
0 & -a^2 - b^2 - c^2 & 0 \\
0 & 0 & -a^2 - b^2 - c^2
\end{pmatrix} + \begin{pmatrix}
a^2 & ba & ca \\
ab & b^2 & cb \\
ac & cb & c^2
\end{pmatrix}
\]

\[
= -\theta^2 I_3 + B,
\]

and that

\[
AB = BA = 0.
\]

From the above, deduce that

\[
A^3 = -\theta^2 A,
\]

and for any \( k \geq 0 \),

\[
A^{4k+1} = \theta^{4k} A, \\
A^{4k+2} = \theta^{4k} A^2, \\
A^{4k+3} = -\theta^{4k+2} A, \\
A^{4k+4} = -\theta^{4k+2} A^2.
\]
Then prove the desired result by writing the power series for \( e^A \) and regrouping terms so that the power series for \( \cos \) and \( \sin \) show up. In particular

\[
e^A = I_3 + \sum_{p \geq 1} \frac{A^p}{p!} = I_3 + \sum_{p \geq 0} \frac{A^{2p+1}}{(2p+1)!} + \sum_{p \geq 1} \frac{A^{2p}}{(2p)!}
\]

\[
= I_3 + \sum_{p \geq 0} \frac{(-1)^p \theta^{2p}}{(2p+1)!} A + \sum_{p \geq 1} \frac{(-1)^{p-1} \theta^{2(p-1)}}{(2p)!} A^2
\]

\[
= I_3 + \frac{A}{\theta} \sum_{p \geq 0} \frac{(-1)^p \theta^{2p+1}}{(2p+1)!} - \frac{A^2}{\theta^2} \sum_{p \geq 1} \frac{(-1)^p \theta^{2p}}{(2p)!}
\]

\[
= I_3 + \frac{\sin \theta}{\theta} A - \frac{A^2}{\theta^2} \sum_{p \geq 0} \frac{(-1)^p \theta^{2p}}{(2p)!} + \frac{A^2}{\theta^2}
\]

\[
= I_3 + \frac{\sin \theta}{\theta} A + \frac{(1 - \cos \theta)}{\theta^2} A^2.
\]

The above formulae are the well-known formulae expressing a rotation of axis specified by the vector \((a, b, c)\) and angle \(\theta\). Since the exponential is surjective, it is possible to write down an explicit formula for its inverse (but it is a multivalued function!). This has applications in kinematics, robotics, and motion interpolation.

### 1.3 Symmetric Matrices, Symmetric Positive Definite Matrices, and the Exponential Map

Recall that a real symmetric matrix is called positive (or positive semidefinite) if its eigenvalues are all positive or null, and positive definite if its eigenvalues are all strictly positive. We denote the vector space of real symmetric \(n \times n\) matrices by \(S(n)\), the set of symmetric positive matrices by \(SP(n)\), and the set of symmetric positive definite matrices by \(SPD(n)\).

The next proposition shows that every symmetric positive definite matrix \(A\) is of the form \(e^B\) for some unique symmetric matrix \(B\). The set of symmetric matrices is a vector space, but it is not a Lie algebra because the Lie bracket \([A, B]\) is not symmetric unless \(A\) and \(B\) commute, and the set of symmetric (positive) definite matrices is not a multiplicative group, so this result is of a different flavor as Theorem 1.6.

**Proposition 1.8.** For every symmetric matrix \(B\), the matrix \(e^B\) is symmetric positive definite. For every symmetric positive definite matrix \(A\), there is a unique symmetric matrix \(B\) such that \(A = e^B\).

**Proof.** We showed earlier that

\[
(e^B)^T = e^{B^T}.
\]
If $B$ is a symmetric matrix, then since $B^T = B$, we get

$$(e^B)^\top = e^{B^\top} = e^B,$$

and $e^B$ is also symmetric. Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the symmetric matrix $B$ are real and the eigenvalues of $e^B$ are $e^{\lambda_1}, \ldots, e^{\lambda_n}$, and since $e^\lambda > 0$ if $\lambda \in \mathbb{R}$, $e^B$ is positive definite.

To show the subjectivity of the exponential map, note that if $A$ is symmetric positive definite, then by Theorem 12.3 from Chapter 12 of Gallier [72], there is an orthogonal matrix $P$ such that $A = PD P^\top$, where $D$ is a diagonal matrix

$$D = \begin{pmatrix}
\lambda_1 & \cdots \\
& \lambda_2 & \cdots \\
& & \ddots & \cdots \\
& & & \cdots & \lambda_n
\end{pmatrix},$$

where $\lambda_i > 0$, since $A$ is positive definite. Letting

$$L = \begin{pmatrix}
\log \lambda_1 & \cdots \\
& \log \lambda_2 & \cdots \\
& & \ddots & \cdots \\
& & & \cdots & \log \lambda_n
\end{pmatrix},$$

by using the power series representation of $e^L$, it is obvious that $e^L = D$, with $\log \lambda_i \in \mathbb{R}$, since $\lambda_i > 0$.

Let

$$B = PLP^\top.$$ 

By Proposition 1.2, we have

$$e^B = e^{PLP^\top} = e^{P^{-1}LP} = Pe^LP^{-1} = Pe^L P^\top = PD P^\top = A.$$ 

Finally, we prove that if $B_1$ and $B_2$ are symmetric and $A = e^{B_1} = e^{B_2}$, then $B_1 = B_2$. We use an argument due to Chevalley [41] (see Chapter I, Proposition 5, pages 13-14). Since $B_1$ is symmetric, there is an orthonormal basis $(u_1, \ldots, u_n)$ of eigenvectors of $B_1$. Let $\mu_1, \ldots, \mu_n$ be the corresponding eigenvalues. Similarly, there is an orthonormal basis $(v_1, \ldots, v_n)$ of eigenvectors of $B_2$. We are going to prove that $B_1$ and $B_2$ agree on the basis $(v_1, \ldots, v_n)$, thus proving that $B_1 = B_2$.

Let $\mu$ be some eigenvalue of $B_2$, and let $v = v_i$ be some eigenvector of $B_2$ associated with $\mu$. We can write

$$v = \alpha_1 u_1 + \cdots + \alpha_n u_n.$$ 

Since $v$ is an eigenvector of $B_2$ for $\mu$ and $A = e^{B_2}$, by Proposition 1.4

$$A(v) = e^\mu v = e^\mu \alpha_1 u_1 + \cdots + e^\mu \alpha_n u_n.$$
On the other hand, 

\[ A(v) = A(\alpha_1 u_1 + \cdots + \alpha_n u_n) = \alpha_1 A(u_1) + \cdots + \alpha_n A(u_n), \]

and since \( A = e^{B_1} \) and \( B_1(u_i) = \mu_i u_i \), by Proposition 1.4 we get

\[ A(v) = e^{\mu_1} \alpha_1 u_1 + \cdots + e^{\mu_n} \alpha_n u_n. \]

Therefore, \( \alpha_i = 0 \) if \( \mu_i \neq \mu \). Letting 

\[ I = \{ i \mid \mu_i = \mu, i \in \{1, \ldots, n\} \}, \]

we have

\[ v = \sum_{i \in I} \alpha_i u_i. \]

Now, 

\[
B_1(v) = B_1 \left( \sum_{i \in I} \alpha_i u_i \right) = \sum_{i \in I} \alpha_i B_1(u_i) = \sum_{i \in I} \alpha_i \mu_i u_i
\]

\[ = \sum_{i \in I} \alpha_i \mu_i u_i = \mu \left( \sum_{i \in I} \alpha_i u_i \right) = \mu v, \]

since \( \mu_i = \mu \) when \( i \in I \). Since \( v \) is an eigenvector of \( B_2 \) for \( \mu \),

\[ B_2(v) = \mu v, \]

which shows that

\[ B_1(v) = B_2(v). \]

Since the above holds for every eigenvector \( v_i \), we have \( B_1 = B_2 \).

Proposition 1.8 can be reformulated as stating that the map \( \exp: S(n) \rightarrow \text{SPD}(n) \) is a bijection. It can be shown that it is a homeomorphism. In the case of invertible matrices, the polar form theorem can be reformulated as stating that there is a bijection between the topological space \( \text{GL}(n, \mathbb{R}) \) of real \( n \times n \) invertible matrices (also a group) and \( \text{O}(n) \times \text{SPD}(n) \).

As a corollary of the polar form theorem (Theorem 13.1 in Chapter 13 of Gallier [72]) and Proposition 1.8, we have the following result: For every invertible matrix \( A \) there is a unique orthogonal matrix \( R \) and a unique symmetric matrix \( S \) such that

\[ A = Re^S. \]

Thus, we have a bijection between \( \text{GL}(n, \mathbb{R}) \) and \( \text{O}(n) \times S(n) \). But \( S(n) \) itself is isomorphic to \( \mathbb{R}^{n(n+1)/2} \). Thus, there is a bijection between \( \text{GL}(n, \mathbb{R}) \) and \( \text{O}(n) \times \mathbb{R}^{n(n+1)/2} \). It can also be shown that this bijection is a homeomorphism. This is an interesting fact. Indeed, this
homeomorphism essentially reduces the study of the topology of $\text{GL}(n, \mathbb{R})$ to the study of the topology of $\text{O}(n)$. This is nice, since it can be shown that $\text{O}(n)$ is compact.

In $A = Re^S$, if $\det(A) > 0$, then $R$ must be a rotation matrix (i.e., $\det(R) = +1$), since $\det(e^S) > 0$. In particular, if $A \in \text{SL}(n, \mathbb{R})$, since $\det(A) = \det(R) = +1$, the symmetric matrix $S$ must have a null trace, i.e., $S \in \mathcal{S}(n) \cap \mathfrak{sl}(n, \mathbb{R})$. Thus, we have a bijection between $\text{SL}(n, \mathbb{R})$ and $\text{SO}(n) \times (\mathcal{S}(n) \cap \mathfrak{sl}(n, \mathbb{R}))$.

We can also show that the exponential map is a surjective map from the skew Hermitian matrices to the unitary matrices (use Theorem 12.7 from Chapter 12 in Gallier [72]).

### 1.4 The Lie Groups $\text{GL}(n, \mathbb{C})$, $\text{SL}(n, \mathbb{C})$, $\text{U}(n)$, $\text{SU}(n)$, the Lie Algebras $\mathfrak{gl}(n, \mathbb{C})$, $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, $\mathfrak{su}(n)$, and the Exponential Map

The set of complex invertible $n \times n$ matrices forms a group under multiplication, denoted by $\text{GL}(n, \mathbb{C})$. The subset of $\text{GL}(n, \mathbb{C})$ consisting of those matrices having determinant $+1$ is a subgroup of $\text{GL}(n, \mathbb{C})$, denoted by $\text{SL}(n, \mathbb{C})$. It is also easy to check that the set of complex $n \times n$ unitary matrices forms a group under multiplication, denoted by $\text{U}(n)$. The subset of $\text{U}(n)$ consisting of those matrices having determinant $+1$ is a subgroup of $\text{U}(n)$, denoted by $\text{SU}(n)$. We can also check that the set of complex $n \times n$ matrices with null trace forms a real vector space under addition, and similarly for the set of skew Hermitian matrices and the set of skew Hermitian matrices with null trace.

**Definition 1.2.** The group $\text{GL}(n, \mathbb{C})$ is called the *general linear group*, and its subgroup $\text{SL}(n, \mathbb{C})$ is called the *special linear group*. The group $\text{U}(n)$ of unitary matrices is called the *unitary group*, and its subgroup $\text{SU}(n)$ is called the *special unitary group*. The real vector space of complex $n \times n$ matrices with null trace is denoted by $\mathfrak{sl}(n, \mathbb{C})$, the real vector space of skew Hermitian matrices is denoted by $\mathfrak{u}(n)$, and the real vector space $\mathfrak{u}(n) \cap \mathfrak{sl}(n, \mathbb{C})$ is denoted by $\mathfrak{su}(n)$.

**Remarks:**

1. As in the real case, the groups $\text{GL}(n, \mathbb{C})$, $\text{SL}(n, \mathbb{C})$, $\text{U}(n)$, and $\text{SU}(n)$ are also topological groups (viewed as subspaces of $\mathbb{R}^{2n^2}$), and in fact, smooth real manifolds. Such objects are called *(real) Lie groups*. The real vector spaces $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$ are *Lie algebras* associated with $\text{SL}(n, \mathbb{C})$, $\text{U}(n)$, and $\text{SU}(n)$. The algebra structure is given by the *Lie bracket*, which is defined as

$$[A, B] = AB - BA.$$
It is also possible to define complex Lie groups, which means that they are topological groups and smooth complex manifolds. It turns out that $GL(n, \mathbb{C})$ and $SL(n, \mathbb{C})$ are complex manifolds, but not $U(n)$ and $SU(n)$.

One should be very careful to observe that even though the Lie algebras $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$ consist of matrices with complex coefficients, we view them as real vector spaces. The Lie algebra $\mathfrak{sl}(n, \mathbb{C})$ is also a complex vector space, but $\mathfrak{u}(n)$ and $\mathfrak{su}(n)$ are not! Indeed, if $A$ is a skew Hermitian matrix, $iA$ is not skew Hermitian, but Hermitian!

Again the Lie algebra achieves a “linearization” of the Lie group. In the complex case, the Lie algebras $\mathfrak{gl}(n, \mathbb{C})$ is the set of all complex $n \times n$ matrices, but $\mathfrak{u}(n) \neq \mathfrak{su}(n)$, because a skew Hermitian matrix does not necessarily have a null trace.

The properties of the exponential map also play an important role in studying complex Lie groups. For example, it is clear that the map

$$\exp: \mathfrak{gl}(n, \mathbb{C}) \to GL(n, \mathbb{C})$$

is well-defined, but this time, it is surjective! One way to prove this is to use the Jordan normal form. Similarly, since

$$\det (e^A) = e^{\text{tr}(A)},$$

the map

$$\exp: \mathfrak{sl}(n, \mathbb{C}) \to SL(n, \mathbb{C})$$

is well-defined, but it is not surjective! As we will see in the next theorem, the maps

$$\exp: \mathfrak{u}(n) \to U(n)$$

and

$$\exp: \mathfrak{su}(n) \to SU(n)$$

are well-defined and surjective.

**Theorem 1.9.** The exponential maps

$$\exp: \mathfrak{u}(n) \to U(n) \quad \text{and} \quad \exp: \mathfrak{su}(n) \to SU(n)$$

are well-defined and surjective.

**Proof.** First, we need to prove that if $A$ is a skew Hermitian matrix, then $e^A$ is a unitary matrix. Recall that $A^* = -\overline{A^T}$. Then since $(e^A)^T = e^{A^T}$, we readily deduce that

$$(e^A)^* = e^{A^*}.$$ 

Then, since $A^* = -A$, we get

$$(e^A)^* = e^{A^*} = e^{-A},$$

and

$$e^{iA} = \cos \theta + i \sin \theta.$$
and so 
\[(e^A)^* e^A = e^{-A} e^A = e^{-A+A} = e^0 = I_n,\]
and similarly, 
\[e^A (e^A)^* = I_n,\]
showing that \(e^A\) is unitary. Since 
\[\det(e^A) = e^{\text{tr}(A)},\]
if \(A\) is skew Hermitian and has null trace, then \(\det(e^A) = +1.\)

For the surjectivity we will use Theorem 12.7 in Chapter 12 of Gallier [72]. First, assume that \(A\) is a unitary matrix. By Theorem 12.7, there is a unitary matrix \(U\) and a diagonal matrix \(D\) such that \(A = U D U^*\). Furthermore, since \(A\) is unitary, the entries \(\lambda_1, \ldots, \lambda_n\) in \(D\) (the eigenvalues of \(A\)) have absolute value +1. Thus, the entries in \(D\) are of the form \(\cos \theta + i \sin \theta = e^{i\theta}\). Thus, we can assume that \(D\) is a diagonal matrix of the form

\[
D = \begin{pmatrix}
e^{i\theta_1} & & & \\
 & e^{i\theta_2} & & \\
 & & \ddots & \\
 & & & e^{i\theta_p}
\end{pmatrix}.
\]

If we let \(E\) be the diagonal matrix

\[
E = \begin{pmatrix}
i\theta_1 & & & \\
 & i\theta_2 & & \\
 & & \ddots & \\
 & & & i\theta_p
\end{pmatrix},
\]

it is obvious that \(E\) is skew Hermitian and that 
\[e^E = D.\]

Then, letting \(B = U E U^*\), we have 
\[e^B = A,\]
and it is immediately verified that \(B\) is skew Hermitian, since \(E\) is.

If \(A\) is a unitary matrix with determinant +1, since the eigenvalues of \(A\) are \(e^{i\theta_1}, \ldots, e^{i\theta_p}\) and the determinant of \(A\) is the product
\[e^{i\theta_1} \cdots e^{i\theta_p} = e^{i(\theta_1 + \cdots + \theta_p)}\]
of these eigenvalues, we must have 
\[\theta_1 + \cdots + \theta_p = 0,\]
and so, $E$ is skew Hermitian and has zero trace. As above, letting

$$B = U E U^*,$$

we have

$$e^B = A,$$

where $B$ is skew Hermitian and has null trace.

We now extend the result of Section 1.3 to Hermitian matrices.

### 1.5 Hermitian Matrices, Hermitian Positive Definite Matrices, and the Exponential Map

Recall that a Hermitian matrix is called positive (or positive semidefinite) if its eigenvalues are all positive or null, and positive definite if its eigenvalues are all strictly positive. We denote the real vector space of Hermitian $n \times n$ matrices by $\mathbf{H}(n)$, the set of Hermitian positive matrices by $\mathbf{HP}(n)$, and the set of Hermitian positive definite matrices by $\mathbf{HPD}(n)$.

The next proposition shows that every Hermitian positive definite matrix $A$ is of the form $e^B$ for some unique Hermitian matrix $B$. As in the real case, the set of Hermitian matrices is a real vector space, but it is not a Lie algebra because the Lie bracket $[A, B]$ is not Hermitian unless $A$ and $B$ commute, and the set of Hermitian (positive) definite matrices is not a multiplicative group.

**Proposition 1.10.** For every Hermitian matrix $B$, the matrix $e^B$ is Hermitian positive definite. For every Hermitian positive definite matrix $A$, there is a unique Hermitian matrix $B$ such that $A = e^B$.

**Proof.** It is basically the same as the proof of Theorem 1.10, except that a Hermitian matrix can be written as $A = U D U^*$, where $D$ is a real diagonal matrix and $U$ is unitary instead of orthogonal.

Proposition 1.10 can be reformulated as stating that the map $\exp: \mathbf{H}(n) \to \mathbf{HPD}(n)$ is a bijection. In fact, it can be shown that it is a homeomorphism. In the case of complex invertible matrices, the polar form theorem can be reformulated as stating that there is a bijection between the topological space $\mathrm{GL}(n, \mathbb{C})$ of complex $n \times n$ invertible matrices (also a group) and $\mathbf{U}(n) \times \mathbf{HPD}(n)$. As a corollary of the polar form theorem and Proposition 1.10, we have the following result: For every complex invertible matrix $A$, there is a unique unitary matrix $U$ and a unique Hermitian matrix $S$ such that

$$A = U e^S.$$

Thus, we have a bijection between $\mathrm{GL}(n, \mathbb{C})$ and $\mathbf{U}(n) \times \mathbf{H}(n)$. But $\mathbf{H}(n)$ itself is isomorphic to $\mathbb{R}^{n^2}$, and so there is a bijection between $\mathrm{GL}(n, \mathbb{C})$ and $\mathbf{U}(n) \times \mathbb{R}^{n^2}$. It can also be
shown that this bijection is a homeomorphism. This is an interesting fact. Indeed, this homeomorphism essentially reduces the study of the topology of $\text{GL}(n, \mathbb{C})$ to the study of the topology of $\text{U}(n)$. This is nice, since it can be shown that $\text{U}(n)$ is compact (as a real manifold).

In the polar decomposition $A = U e^S$, we have $|\det(U)| = 1$, since $U$ is unitary, and $\text{tr}(S)$ is real, since $S$ is Hermitian (since it is the sum of the eigenvalues of $S$, which are real), so that $\det(e^S) > 0$. Thus, if $\det(A) = 1$, we must have $\det(e^S) = 1$, which implies that $S \in \text{H}(n) \cap \text{sl}(n, \mathbb{C})$. Thus, we have a bijection between $\text{SL}(n, \mathbb{C})$ and $\text{SU}(n) \times (\text{H}(n) \cap \text{sl}(n, \mathbb{C}))$.

In the next section we study the group $\text{SE}(n)$ of affine maps induced by orthogonal transformations, also called rigid motions, and its Lie algebra. We will show that the exponential map is surjective. The groups $\text{SE}(2)$ and $\text{SE}(3)$ play play a fundamental role in robotics, dynamics, and motion planning.

### 1.6 The Lie Group $\text{SE}(n)$ and the Lie Algebra $\mathfrak{se}(n)$

First, we review the usual way of representing affine maps of $\mathbb{R}^n$ in terms of $(n+1) \times (n+1)$ matrices.

**Definition 1.3.** The set of affine maps $\rho$ of $\mathbb{R}^n$, defined such that

$$\rho(X) = RX + U,$$

where $R$ is a rotation matrix ($R \in \text{SO}(n)$) and $U$ is some vector in $\mathbb{R}^n$, is a group under composition called the group of *direct affine isometries, or rigid motions*, denoted by $\text{SE}(n)$.

Every rigid motion can be represented by the $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$$

in the sense that

$$\begin{pmatrix} \rho(X) \\ 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}$$

iff

$$\rho(X) = RX + U.$$

**Definition 1.4.** The vector space of real $(n+1) \times (n+1)$ matrices of the form

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix},$$

where $\Omega$ is a skew symmetric matrix and $U$ is a vector in $\mathbb{R}^n$, is denoted by $\mathfrak{se}(n)$. 

Remark: The group $\text{SE}(n)$ is a Lie group, and its Lie algebra turns out to be $\mathfrak{se}(n)$.

We will show that the exponential map $\exp : \mathfrak{se}(n) \to \text{SE}(n)$ is surjective. First, we prove the following key proposition.

**Proposition 1.11.** Given any $(n+1) \times (n+1)$ matrix of the form

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix}$$

where $\Omega$ is any matrix and $U \in \mathbb{R}^n$,

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix},$$

where $\Omega^0 = I_n$. As a consequence,

$$e^A = \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix},$$

where

$$V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!} = \sum_{k \geq 1} \frac{\Omega^{k-1}}{k!}.$$

**Proof.** A trivial induction on $k$ shows that

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix}.$$

Then we have

$$e^A = \sum_{k \geq 0} \frac{A^k}{k!},$$

$$= I_{n+1} + \sum_{k \geq 1} \frac{1}{k!} \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix},$$

$$= \begin{pmatrix} I_n + \sum_{k \geq 1} \frac{\Omega^k}{k!} & \sum_{k \geq 1} \frac{\Omega^{k-1}}{k!}U \\ 0 & 1 \end{pmatrix},$$

$$= \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix}.$$
We can now prove our main theorem. We will need to prove that $V$ is invertible when $\Omega$ is a skew symmetric matrix. It would be tempting to write $V$ as

$$V = \Omega^{-1}(e^{\Omega} - I).$$

Unfortunately, for odd $n$, a skew symmetric matrix of order $n$ is not invertible! Thus, we have to find another way of proving that $V$ is invertible. However, observe that we have the following useful fact:

$$V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!} = \int_0^1 e^{\Omega t} dt,$$

since $e^{\Omega t}$ is absolutely convergent and term by term integration yields

$$\int_0^1 e^{\Omega t} dt = \int_0^1 \sum_{k \geq 0} \frac{(\Omega t)^k}{k!} dt = \sum_{k \geq 0} \frac{1}{k!} \int_0^1 (\Omega t)^k dt$$

$$= \sum_{k \geq 0} \frac{\Omega^k}{k!} \int_0^1 t^k dt = \sum_{k \geq 0} \frac{\Omega^k}{k!} \left[ \frac{t^{k+1}}{k+1} \right]_0^1$$

$$= \sum_{k \geq 1} \frac{\Omega^{k-1}}{k!} = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!}.$$ 

This is what we will use in Theorem 1.12 to prove surjectivity.

**Theorem 1.12.** The exponential map

$$\exp : \mathfrak{se}(n) \to \mathbf{SE}(n)$$

is well-defined and surjective.

**Proof.** Since $\Omega$ is skew symmetric, $e^{\Omega}$ is a rotation matrix, and by Theorem 1.6, the exponential map

$$\exp : \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is surjective. Thus, it remains to prove that for every rotation matrix $R$, there is some skew symmetric matrix $\Omega$ such that $R = e^{\Omega}$ and

$$V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!}$$

is invertible. This is because Proposition 1.11 will then imply

$$e^{\begin{pmatrix} \Omega & V^{-1}U \\ 0 & 0 \end{pmatrix}} = \begin{pmatrix} e^{\Omega} & VV^{-1}U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}. $$
Theorem 12.5 from Chapter 12 of Gallier [72] says that for every orthogonal matrix \( R \) there is an orthogonal matrix \( P \) such that \( R = PEP^\top \), where \( E \) is a block diagonal matrix of the form

\[
E = \begin{pmatrix}
E_1 & \cdots & \\
\vdots & \ddots & \\
E_p & \cdots & \\
\end{pmatrix},
\]

such that each block \( E_i \) is either 1, \(-1\), or a two-dimensional matrix of the form

\[
E_i = \begin{pmatrix}
\cos \theta_i & -\sin \theta_i \\
\sin \theta_i & \cos \theta_i
\end{pmatrix},
\]

Furthermore, if \( R \) is a rotation matrix, then we may assume that \( 0 < \theta_i \leq \pi \) and that the scalar entries are \(+1\). Then, we can form the block diagonal matrix

\[
D = \begin{pmatrix}
D_1 & \cdots & \\
\vdots & \ddots & \\
D_p & \cdots & \\
\end{pmatrix}
\]

such that each block \( D_i \) is either 0 when \( E_i \) consists of \(+1\), or the two-dimensional matrix

\[
D_i = \begin{pmatrix}
0 & -\theta_i \\
\theta_i & 0
\end{pmatrix}
\]

when

\[
E_i = \begin{pmatrix}
\cos \theta_i & -\sin \theta_i \\
\sin \theta_i & \cos \theta_i
\end{pmatrix},
\]

with \( 0 < \theta_i \leq \pi \). If we let \( \Omega = PD P^\top \), then

\[
e^\Omega = R,
\]

as in the proof of Theorem 1.6. To compute \( V \), since \( \Omega = PD P^\top = PDP^{-1} \), observe that

\[
V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k + 1)!}
\]

\[
= I_n + \sum_{k \geq 1} \frac{PD^k P^{-1}}{(k + 1)!}
\]

\[
= P \left( I_n + \sum_{k \geq 1} \frac{D^k}{(k + 1)!} \right) P^{-1}
\]

\[
= PW P^{-1},
\]
1.6. THE LIE GROUP $\text{SE}(N)$ AND THE LIE ALGEBRA $\mathfrak{se}(N)$

where

$$ W = I_n + \sum_{k \geq 1} \frac{D^k}{(k+1)!}. $$

We can compute

$$ W = I_n + \sum_{k \geq 1} \frac{D^k}{(k+1)!} = \int_0^1 e^{Dt} dt, $$

by computing

$$ W = \begin{pmatrix} W_1 & \cdots & \vdots & \vdots \\ \vdots & W_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & W_p \end{pmatrix} $$

by blocks. Since

$$ e^{Di t} = \begin{pmatrix} \cos(\theta_i t) & -\sin(\theta_i t) \\ \sin(\theta_i t) & \cos(\theta_i t) \end{pmatrix} $$

when $D_i$ is a $2 \times 2$ skew symmetric matrix

$$ D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix} $$

and $W_i = \int_0^1 e^{Di t} dt$, we get

$$ W_i = \begin{pmatrix} \int_0^1 \cos(\theta_i t) dt & \int_0^1 -\sin(\theta_i t) dt \\ \int_0^1 \sin(\theta_i t) dt & \int_0^1 \cos(\theta_i t) dt \end{pmatrix} = \frac{1}{\theta_i^2} \begin{pmatrix} \sin(\theta_i) |_0^1 & \cos(\theta_i) |_0^1 \\ -\cos(\theta_i) |_0^1 & \sin(\theta_i) |_0^1 \end{pmatrix}, $$

that is,

$$ W_i = \frac{1}{\theta_i^2} \begin{pmatrix} \sin \theta_i & -(1 - \cos \theta_i) \\ 1 - \cos \theta_i & \sin \theta_i \end{pmatrix}, $$

and $W_i = 1$ when $D_i = 0$. Now, in the first case, the determinant is

$$ \frac{1}{\theta_i^2} ((\sin \theta_i)^2 + (1 - \cos \theta_i)^2) = \frac{2}{\theta_i^2} (1 - \cos \theta_i), $$

which is nonzero, since $0 < \theta_i \leq \pi$. Thus, each $W_i$ is invertible, and so is $W$, and thus, $V = PWP^{-1}$ is invertible.

In the case $n = 3$, given a skew symmetric matrix

$$ \Omega = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}, $$

letting $\theta = \sqrt{a^2 + b^2 + c^2}$, it it easy to prove that if $\theta = 0$, then
$e^A = \begin{pmatrix} I_3 & U \\ 0 & 1 \end{pmatrix}$,

and that if $\theta \neq 0$ (using the fact that $\Omega^3 = -\theta^2 \Omega$), then by adjusting the calculation found at the end of Section 1.2

$$e^\Omega = I_3 + \frac{\sin \theta}{\theta} \Omega + \frac{(1 - \cos \theta)}{\theta^2} \Omega^2$$

and

$$V = I_3 + \frac{(1 - \cos \theta)}{\theta^2} \Omega + \frac{(\theta - \sin \theta)}{\theta^3} \Omega^2.$$
Chapter 2

Basic Analysis: Review of Series and Derivatives

The goal of Chapter 3 is to define embedded submanifolds and (linear) Lie groups. Before doing this, we believe that some readers might appreciate a review of the basic properties of power series involving matrix coefficients and a review of the notion of the derivative of a function between two normed vector spaces. Those readers familiar with these concepts may proceed directly to Chapter 3.

2.1 Series and Power Series of Matrices

Since a number of important functions on matrices are defined by power series, in particular the exponential, we review quickly some basic notions about series in a complete normed vector space.

Given a normed vector space \((E, \| \|)\), a series is an infinite sum \(\sum_{k=0}^{\infty} a_k\) of elements \(a_k \in E\). We denote by \(S_n\) the partial sum of the first \(n + 1\) elements,

\[
S_n = \sum_{k=0}^{n} a_k.
\]

**Definition 2.1.** We say that the series \(\sum_{k=0}^{\infty} a_k\) converges to the limit \(a \in E\) if the sequence \((S_n)\) converges to \(a\), i.e. given \(\epsilon > 0\), there exists a nonnegative integer \(N\) such that for all \(n \geq N\)

\[
\|S_n - a\| = \left\| \sum_{k=n+1}^{\infty} a_k \right\| < \epsilon.
\]

In this case, we say that the series is convergent. We say that the series \(\sum_{k=0}^{\infty} a_k\) converges absolutely if the series of norms \(\sum_{k=0}^{\infty} \|a_k\|\) is convergent.
CHAPTER 2. BASIC ANALYSIS: REVIEW OF SERIES AND DERIVATIVES

To intuitively understand Definition 2.1, think of \((a_n)\) as a long string or "snake" of vector entries. We subdivide this snake into head, body, and tail by choosing \(m > n \geq 0\) and writing
\[
\sum_{k=0}^{\infty} a_k = H + B + T,
\]
where
\[
H = \sum_{k=0}^{n} a_k = a_1 + a_2 + \cdots + a_n,
\]
\[
B = \sum_{k=n+1}^{m} a_k = a_{n+1} + a_{n+2} + \cdots + a_m,
\]
\[
T = \sum_{k=m+1}^{\infty} a_k = a_{m+1} + a_{m+2} + \ldots.
\]
Note \(H\) stands for head, \(B\) stands for body, and \(T\) stands for tail. The convergence of \(\sum_{k=0}^{\infty} a_k\) means \(T\) is arbitrarily small whenever \(m\) is "large enough". In particular, we have the following useful proposition.

**Proposition 2.1.** If \(\sum_{k=0}^{\infty} a_k\) converges, then \(\lim_{k \to \infty} a_k = \lim_{n \to \infty} \|a_k\| = 0\). Given \(N \geq 0\) and a fixed positive value \(s > 0\) infinitely many times whenever \(k \geq N\), then \(\sum_{k=0}^{\infty} a_k\) diverges.

The belly of the snake may be characterized in terms of a Cauchy sequence.

**Definition 2.2.** Given a normed vector space, \(E\), we say that a sequence, \((a_n)_n\), with \(a_n \in E\), is a **Cauchy sequence** iff for every \(\epsilon > 0\), there is some \(N > 0\) so that for all \(m, n \geq N\),
\[
\|a_n - a_m\| < \epsilon.
\]

**Definition 2.3.** A normed vector space, \(E\), is **complete** iff every Cauchy sequence converges. A complete normed vector space is also called a **Banach space**, after Stefan Banach (1892-1945).

There are series that are convergent but not absolutely convergent; for example, the series
\[
\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}.
\]
If \(E\) is complete, the converse is an enormously useful result.

**Proposition 2.2.** Assume \((E, \| \|)\) is a complete normed vector space. If a series \(\sum_{k=0}^{\infty} a_k\) is absolutely convergent, then it is convergent.
2.1. SERIES AND POWER SERIES OF MATRICES

Proof. If $\sum_{k=0}^{\infty} a_k$ is absolutely convergent, then we prove that the sequence $(S_m)$ is a Cauchy sequence; that is, for every $\epsilon > 0$, there is some $p > 0$ such that for all $n \geq m \geq p$, 

$$\|S_n - S_m\| \leq \epsilon.$$ 

Observe that 

$$\|S_n - S_m\| = \|a_{m+1} + \cdots + a_n\| \leq \|a_{m+1}\| + \cdots + \|a_n\|,$$

and since the sequence $\sum_{k=0}^{n} \|a_k\|$ converges, it satisfies Cauchy’s criterion. Thus, the sequence $(S_m)$ also satisfies Cauchy’s criterion, and since $E$ is a complete vector space, the sequence $(S_m)$ converges. \( \square \)

Remark: It can be shown that if $(E, \|\|)$ is a normed vector space such that every absolutely convergent series is also convergent, then $E$ must be complete (see Schwartz [155]).

An important corollary of absolute convergence is that if the terms in series $\sum_{k=0}^{\infty} a_k$ are rearranged, then the resulting series is still absolutely convergent, and has the same sum. More precisely, let $\sigma$ be any permutation (bijection) of the natural numbers. The series $\sum_{k=0}^{\infty} a_{\sigma(k)}$ is called a rearrangement of the original series. The following result can be shown (see Schwartz [155]).

**Proposition 2.3.** Assume $(E, \|\|)$ is a normed vector space. If a series $\sum_{k=0}^{\infty} a_k$ is convergent as well as absolutely convergent, then for every permutation $\sigma$ of $\mathbb{N}$, the series $\sum_{k=0}^{\infty} a_{\sigma(k)}$ is convergent and absolutely convergent, and its sum is equal to the sum of the original series:

$$\sum_{k=0}^{\infty} a_{\sigma(k)} = \sum_{k=0}^{\infty} a_k.$$ 

In particular, if $(E, \|\|)$ is a complete normed vector space, then Proposition 2.3 holds. A series $\sum_{k=0}^{\infty} a_k$ is said to be unconditionally convergent (or commutatively convergent) if the series $\sum_{k=0}^{\infty} a_{\sigma(k)}$ is convergent for every permutation $\sigma$ of $\mathbb{N}$, and if all these rearrangements have the same sum. It can be show that if $E$ has finite dimension, then a series is absolutely convergent iff it is unconditionally convergent. However, this is false if $E$ has infinite dimension (but hard to prove).

If $E = \mathbb{C}$, then there are several conditions that imply the absolute convergence of a series. In the rest of this section, we omit most proofs, details of which can be found in introductory analysis books such as [4] and [155].

The ratio test is the following test. Suppose there is some $N > 0$ such that $a_n \neq 0$ for all $n \geq N$, and either

$$r = \lim_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right|$$

exists, or the sequence of ratios diverges to infinity, in which case we write $r = \infty$. Then, if $0 \leq r < 1$, the series $\sum_{k=0}^{n} a_k$ converges absolutely, else if $1 < r \leq \infty$, the series diverges.
If \((r_n)\) is a sequence of real numbers, recall that
\[
\limsup_{n \to \infty} r_n = \lim_{n \to \infty} \sup_{k \geq n} \{r_k\}.
\]
If \(r_n \geq 0\) for all \(n\), then either the sequence \((r_n)\) is unbounded, in which case \(\sup_{k \geq n+1} \{r_k\} \leq \sup_{k \geq n} \{r_k\}\), the sequence \((\sup_{k \geq n} r_n)_{n \geq 0}\) is nonincreasing and bounded from below by 0, so \(\limsup_{n \to \infty} r_n = r\) exists and is finite. In this case, it is easy to see that \(r\) is characterized as follows:

For every \(\epsilon > 0\), there is some \(N \in \mathbb{N}\) such that \(r_n < r + \epsilon\) for all \(n \geq N\), and \(r_n > r - \epsilon\) for infinitely many \(n\).

The notion of \(\limsup_{n \to \infty} r_n\) may also be characterized in terms of limits of subsequences. Take the family of all subsequences \(\{(r_{n_j})\}\) of \((r_n)\). Consider the set, \(L\), of all possible limits of these subsequences. Then \(\limsup_{n \to \infty} r_n\) is the largest element (possibly infinity) of \(L\).

For example if \((r_n) = (1, -1, 1, -1, \ldots)\), then \(L = \{-1, 1\}\) and \(\limsup_{n \to \infty} r_n = 1\).

Then, the root test is this. Let
\[
r = \limsup_{n \to \infty} |a_n|^{1/n}
\]
if the limit exists (is finite), else write \(r = \infty\). Then, if \(0 \leq r < 1\), the series \(\sum_{k=0}^{n} a_k\) converges absolutely, else if \(1 < r \leq \infty\), the series diverges.

The root test also applies if \((E, \|\|)\) is a complete normed vector space by replacing \(|a_n|\) by \(\|a_n\|\). Let \(\sum_{k \geq 0} a_k\) be a series of elements \(a_k \in E\) and let
\[
r = \limsup_{n \to \infty} \|a_n\|^{1/n}
\]
if the limit exists (is finite), else write \(r = \infty\). Then, if \(0 \leq r < 1\), the series \(\sum_{k=0}^{n} a_k\) converges absolutely, else if \(1 < r \leq \infty\), the series diverges.

A power series with coefficients \(a_k \in \mathbb{C}\) in the indeterminate \(z\) is a formal expression \(f(z)\) of the form
\[
f(z) = \sum_{k=0}^{\infty} a_k z^k,
\]
For any fixed value \(z \in \mathbb{C}\), the series \(f(z)\) may or may not converge. It always converges for \(z = 0\), since \(f(0) = a_0\). A fundamental fact about power series is that they have a radius of convergence.

**Proposition 2.4.** Given any power series
\[
f(z) = \sum_{k=0}^{\infty} a_k z^k,
\]
there is a nonnegative real $R$, possibly infinite, called the **radius of convergence** of the power series, such that if $|z| < R$, then $f(z)$ converges absolutely, else if $|z| > R$, then $f(z)$ diverges. Moreover (Hadamard), we have

$$R = \frac{1}{\limsup_{n \to \infty} |a_n|^{1/n}}.$$

Note that Proposition 2.4 does not say anything about the behavior of the power series for boundary values, that is, values of $z$ such that $|z| = R$.

**Proof.** Given $\sum_{n=0}^{\infty} A_n$, where $(A_n)$ is an arbitrary sequence of complex numbers, note that

$$\sum_{n=0}^{\infty} |A_n| = \sum_{n=0}^{\infty} \left| |A_n|^{\frac{1}{n}} \right|^n. $$

If $\limsup_{n \to \infty} |A_n|^{\frac{1}{n}} < 1$, then $\sum_{n=0}^{\infty} A_n$ converges absolutely.

To see why this is the case, observe that the definition of $\limsup_{n \to \infty} |A_n|^{\frac{1}{n}}$ implies that

$$|A_n|^{\frac{1}{n}} \leq \limsup_{n \to \infty} |A_n|^{\frac{1}{n}} + \epsilon, \quad \text{whenever } n > N(\epsilon).$$

Choose $\epsilon$ small enough so that $|A_n|^{\frac{1}{n}} \leq \limsup_{n \to \infty} |A_n|^{\frac{1}{n}} + \epsilon < r_1 < 1$.

Then

$$\sum_{n=N(\epsilon)+1}^{\infty} |A_n| \leq \sum_{n=N(\epsilon)+1}^{\infty} \left| |A_n|^{\frac{1}{n}} \right|^n \leq \sum_{n=N(\epsilon)+1}^{\infty} r_n^n = \frac{r_1^{N(\epsilon)+1}}{1 - r_1},$$

and an application of the comparison test implies that $\sum_{n=0}^{\infty} A_n$ converges absolutely. It is then a matter of setting $A_n = a_n z^n$ and requiring that

$$\limsup_{n \to \infty} |A_n|^{\frac{1}{n}} = |z| \limsup_{n \to \infty} |a_n|^{\frac{1}{n}} < 1.$$

If $\limsup_{n \to \infty} |A_n|^{\frac{1}{n}} = |z| \limsup_{n \to \infty} |a_n|^{\frac{1}{n}} > 1$, the definition of $\limsup_{n \to \infty} |a_n|^{\frac{1}{n}}$ implies that

$$1 < |z| \limsup_{n \to \infty} |a_n|^{\frac{1}{n}} - \epsilon, \quad \text{for infinitely many } n.$$

Then Proposition 2.1 implies that $\sum_{n=0}^{\infty} A_n = \sum_{n=0}^{\infty} a_n z^n$ diverges.

Even though the ratio test does not apply to every power series, it provides a useful way of computing the radius of convergence of a power series.

**Proposition 2.5.** Let $f(z) = \sum_{k=0}^{\infty} a_k z^k$ be a power series with coefficients $a_k \in \mathbb{C}$. Suppose there is some $N > 0$ such that $a_n \neq 0$ for all $n \geq N$, and either

$$R = \lim_{n \to \infty} \left| \frac{a_n}{a_{n+1}} \right|$$

exists, or the sequence on the righthand side diverges to infinity, in which case we write $R = \infty$. Then, the power series $\sum_{k=0}^{\infty} a_k z^k$ has radius of convergence $R$. 

For example, for the power series
\[ \exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}, \]
we have
\[ \left| \frac{a_k}{a_{k+1}} \right| = \frac{(k+1)!}{k!} = k + 1, \]
whose limit is \( \infty \), so the exponential is defined for all \( z \in \mathbb{C} \); its radius of convergence is \( \infty \). For the power series
\[ f(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+1)!}, \]
we have
\[ \left| \frac{a_k}{a_{k+1}} \right| = \frac{(k+2)!}{(k+1)!} = k + 2, \]
so \( f(z) \) also has infinite radius of convergence.

For the power series
\[ \log(1 + x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}, \]
if \( k \geq 1 \) we have
\[ \left| \frac{a_k}{a_{k+1}} \right| = \frac{k+1}{k}, \]
whose limit is 1, so \( \log(1 + x) \) has radius of convergence 1. For \( x = 1 \), the series converges to \( \log(2) \), but for \( x = -1 \), the series diverges.

Power series behave very well with respect to term by term differentiation and term by term integration.

**Proposition 2.6.** Suppose the power series \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) (with complex coefficients) has radius of convergence \( R > 0 \). Then, \( f'(z) \) exists if \( |z| < R \), the power series \( \sum_{k=1}^{\infty} k a_k z^{k-1} \) has radius of convergence \( R \), and
\[ f'(z) = \sum_{k=1}^{\infty} k a_k z^{k-1}. \]

**Proposition 2.7.** Suppose the power series \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) (with complex coefficients) has radius of convergence \( R > 0 \). Then \( F(z) = \int_0^z f(t) \, dt \) exists if \( |z| < R \), the power series \( \sum_{k=0}^{\infty} \frac{a_k}{k+1} z^{k+1} \) has radius of convergence \( R \), and
\[ F(z) = \sum_{k=0}^{\infty} \frac{a_k}{k+1} z^{k+1}. \]
Let us now assume that \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) is a power series with coefficients \( a_k \in \mathbb{C} \), and that its radius of convergence is \( R \). Given any matrix \( A \in \mathbb{M}_n(\mathbb{C}) \) we can form the power series obtained by substituting \( A \) for \( z \),

\[
f(A) = \sum_{k=0}^{\infty} a_k A^k.
\]

Let \( \| \| \) be any matrix norm on \( \mathbb{M}_n(\mathbb{C}) \). Then the following proposition regarding the convergence of the power series \( f(A) \) holds.

**Proposition 2.8.** Let \( f(z) = \sum_{k=1}^{\infty} a_k z^k \) be a power series with complex coefficients, write \( R \) for its radius of convergence, and assume that \( R > 0 \). For every \( \rho \) such that \( 0 < \rho < R \), the series \( f(A) = \sum_{k=1}^{\infty} a_k A^k \) is absolutely convergent for all \( A \in \mathbb{M}_n(\mathbb{C}) \) such that \( \|A\| \leq \rho \). Furthermore, \( f \) is continuous on the open ball \( B(R) = \{A \in \mathbb{M}_n(\mathbb{C}) | \|A\| < R\} \).

Note that unlike the case where \( A \in \mathbb{C} \), if \( \|A\| > R \), we cannot claim that the series \( f(A) \) diverges. This has to do with the fact that even for the operator norm we may have \( \|A^n\| < \|A\|^n \), a fact which should be contrasted to situation in \( \mathbb{C} \) where \( |a|^n = |a^n| \). We leave it as an exercise to find an example of a series and a matrix \( A \) with \( \|A\| > R \), and yet \( f(A) \) converges. Hint: Consider \( A \) to be nilpotent, i.e. \( A \neq 0 \) but \( A^k = 0 \) for some positive integer \( k \).

As an application of Proposition 2.8, the exponential power series

\[
e^A = \exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}
\]

is absolutely convergent for all \( A \in \mathbb{M}_n(\mathbb{C}) \), and continuous everywhere. Proposition 2.8 also implies that the series

\[
\log(I + A) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{A^k}{k}
\]

is absolutely convergent if \( \|A\| < 1 \).

Next, let us consider the generalization of the notion of a power series \( f(t) = \sum_{k=1}^{\infty} a_k t^k \) of a complex variable \( t \), where the coefficients \( a_k \) belong to a complete normed vector space \( (F, \| \|) \). Then, it is easy to see that Proposition 2.4 generalizes to this situation.

**Proposition 2.9.** Let \( (F, \| \|) \) be a complete normed vector space. Given any power series

\[
f(t) = \sum_{k=0}^{\infty} a_k t^k,
\]

with \( t \in \mathbb{R} \) and \( a_k \in F \), there is a nonnegative real \( R \), possibly infinite, called the **radius of convergence** of the power series, such that if \( |t| < R \), then \( f(t) \) converges absolutely, else if \( |t| > R \), then \( f(t) \) diverges. Moreover, we have

\[
R = \limsup_{n \to \infty} \|a_n\|^{1/n}.
\]
Proposition 2.10. Let \((F, \| \|)\) be a complete normed vector space. Suppose the power series \(f(t) = \sum_{k=0}^{\infty} a_k t^k\) (with coefficients \(a_k \in F\)) has radius of convergence \(R > 0\). Then \(f'(t)\) exists if \(|t| < R\), the power series \(\sum_{k=1}^{\infty} ka_k t^{k-1}\) has radius of convergence \(R\), and
\[
f'(t) = \sum_{k=1}^{\infty} ka_k t^{k-1}.
\]

Proposition 2.11. Let \((F, \| \|)\) be a complete normed vector space. Suppose the power series \(f(t) = \sum_{k=0}^{\infty} a_k t^k\) (with coefficients \(a_k \in F\)) has radius of convergence \(R > 0\). Then \(F(t) = \int_0^t f(z) \, dz\) exists if \(|t| < R\), the power series \(\sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1}\) has radius of convergence \(R\), and
\[
F(t) = \sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1}.
\]

So far we have considered series as individual entities. We end this section with a discussion on ways to combine pairs of series through addition, multiplication, and composition. Given a complete normed vector space \((E, \| \|)\), if \(\sum_{k=0}^{\infty} a_k\) and \(\sum_{k=0}^{\infty} b_k\) are two series with \(a_k, b_k \in E\), we can form the series \(\sum_{k=0}^{\infty} (a_k + b_k)\) whose \(k\)th terms is \(a_k + b_k\), and for any scalar \(\lambda\), the series \(\sum_{k=0}^{\infty} \lambda a_k\), whose \(k\)th terms is \(\lambda a_k\). It is easy to see that if \(\sum_{k=0}^{\infty} a_k\) and \(\sum_{k=0}^{\infty} b_k\) are absolutely convergent with sums \(A\) and \(B\), respectively, then the series \(\sum_{k=0}^{\infty} (a_k + b_k)\) and \(\sum_{k=0}^{\infty} \lambda a_k\) are absolutely convergent, and their sums are given by
\[
\sum_{k=0}^{\infty} (a_k + b_k) = A + B = \sum_{k=0}^{\infty} a_k + \sum_{k=0}^{\infty} b_k
\]
\[
\sum_{k=0}^{\infty} \lambda a_k = \lambda A = \lambda \sum_{k=0}^{\infty} a_k.
\]

If \(f(z) = \sum_{k=0}^{\infty} a_k z^k\) and \(g(z) = \sum_{k=0}^{\infty} b_k z^k\) are two power series with \(a_k, b_k \in E\), we can form the power series \(h(z) = \sum_{k=0}^{\infty} (a_k + b_k) z^k\), and for any scalar \(\lambda\), the power series \(s(z) = \sum_{k=0}^{\infty} \lambda a_k z^k\). We can show easily that if \(f(z)\) has radius of convergence \(R(f)\) and \(g(z)\) has radius of convergence \(R(g)\), then \(h(z)\) has radius of convergence \(\geq \min(R(f), R(g))\), and for every \(z\) such that \(|z| < \min(R(f), R(g))\), we have
\[
h(z) = f(z) + g(z).
\]
Furthermore, \(s(z)\) has radius of convergence \(\geq R(f)\), and for every \(z\) such that \(|z| < R(f)\), we have
\[
s(z) = \lambda f(z).
\]
2.1. SERIES AND POWER SERIES OF MATRICES

The above also applies to power series \( f(A) = \sum_{k=0}^{\infty} a_k A^k \) and \( g(A) = \sum_{k=0}^{\infty} b_k A^k \) with matrix argument \( A \in \text{M}_n(\mathbb{C}) \), with \(|z|\) replaced by \( \|A\| \).

Let us now consider the product of two series \( \sum_{k=0}^{\infty} a_k \) and \( \sum_{k=0}^{\infty} b_k \) where \( a_k, b_k \in \mathbb{C} \). The Cauchy product of these two series is the series \( \sum_{k=0}^{\infty} c_k \), where

\[
c_k = \sum_{i=0}^{k} a_i b_{k-i} \quad k \in \mathbb{N}.
\]

The following result can be shown (for example, see Cartan [35]).

**Proposition 2.12.** Let \( \sum_{k=0}^{\infty} a_k \) and \( \sum_{k=0}^{\infty} b_k \) be two series with coefficients \( a_k, b_k \in \mathbb{C} \). If both series converge absolutely to limits \( A \) and \( B \), respectively, then their Cauchy product \( \sum_{k=0}^{\infty} c_k \), converges absolutely, and if \( C \) is the limit of the Cauchy product, then \( C = AB \).

Next, if \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) and \( g(z) = \sum_{k=0}^{\infty} b_k z^k \) are two power series with coefficients \( a_k, b_k \in \mathbb{C} \), the product of the power series \( f(z) \) and \( g(z) \) is the power series \( h(z) = \sum_{k=0}^{\infty} c_k z^k \) where \( c_k \) is the Cauchy product

\[
c_k = \sum_{i=0}^{k} a_i b_{k-i} \quad k \in \mathbb{N}.
\]

**Proposition 2.13.** Let \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) and \( g(z) = \sum_{k=0}^{\infty} b_k z^k \) be two series with coefficients \( a_k, b_k \in \mathbb{C} \). If both series have a radius of convergence \( \geq \rho \), then their Cauchy product \( h(z) = \sum_{k=0}^{\infty} c_k z^k \) has radius of convergence \( \geq \rho \). Furthermore, for all \( z \), if \( |z| < \rho \), then

\[
h(z) = f(z)g(z).
\]

Proposition 2.13 still holds for power series \( f(A) = \sum_{k=0}^{\infty} a_k A^k \) and \( g(A) = \sum_{k=0}^{\infty} b_k A^k \) with matrix argument \( A \in \text{M}_n(\mathbb{R}) \), with \(|z| < \rho \) replaced by \( \|A\| < \rho \).

Finally, let us consider the substitution of power series. Let \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) and \( g(z) = \sum_{k=0}^{\infty} b_k z^k \) be two series with coefficients \( a_k, b_k \in \mathbb{C} \), and assume that \( a_0 = 0 \). Then, if we substitute \( f(z) \) for \( z \) in \( g(z) \), we get an expression

\[
g(f(z)) = \sum_{k=0}^{\infty} b_k \left( \sum_{n=0}^{\infty} a_n z^n \right)^k,
\]

and because \( a_0 = 0 \), when we expand the powers, there are only finitely many terms involving any monomial \( z^m \), since for \( k > m \), the power \( \left( \sum_{n=0}^{\infty} a_n z^n \right)^k \) has no terms of degree less than \( m \). Thus, we can regroup the terms of \( g(f(z)) \) involving each monomial \( z^m \), and the resulting power series is denoted by \( (g \circ f)(z) \). We have the following result (for example, see Cartan [35]).
Proposition 2.14. Let \( f(z) = \sum_{k=0}^{\infty} a_k z^k \) and \( g(z) = \sum_{k=0}^{\infty} b_k z^k \) be two power series with coefficients \( a_k, b_k \in \mathbb{C} \), and write \( R(f) \) for the radius of convergence of \( f(z) \) and \( R(g) \) for the radius of convergence of \( g(z) \). If \( R(f) > 0 \), \( R(g) > 0 \), and \( a_0 = 0 \), then for any \( r > 0 \) chosen so that \( \sum_{k=1}^{\infty} |a_k|r^k < R(g) \), the following hold:

1. The radius of convergence \( R(h) \) of \( h(z) = (g \circ f)(z) \) is at least \( r \).
2. For every \( z \), if \( |z| \leq r \), then \( |f(z)| < R(g) \), and
   \[
   h(z) = g(f(z)).
   \]

Proposition 2.14 still holds for power series \( f(A) = \sum_{k=0}^{\infty} a_k A^k \) and \( g(A) = \sum_{k=0}^{\infty} b_k A^k \) with matrix argument \( A \in M_n(\mathbb{C}) \), with \( |z| \leq r \) replaced by \( \|A\| \leq r \) and \( |f(z)| < R(g) \) replaced by \( \|f(z)\| < R(g) \).

As an application of Proposition 2.14, (see Cartan [35]) note that the formal power series

\[
E(A) = \sum_{k=1}^{\infty} \frac{A^k}{k!}
\]

and

\[
L(A) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{A^k}{k}
\]

are mutual inverses; that is,

\[
E(L(A)) = A, \quad L(E(A)) = A, \quad \text{for all } A.
\]

Observe that \( E(A) = e^A - I = \exp(A) - I \) and \( L(A) = \log(I + A) \). It follows that

\[
\log(\exp(A)) = A \quad \text{for all } A \text{ with } \|A\| < \log(2)
\]

\[
\exp(\log(I + A)) = I + A \quad \text{for all } A \text{ with } \|A\| < 1.
\]

2.2 The Derivative of a Function Between Normed Vector Spaces, a Review

In this section we review some basic notions of differential calculus, in particular, the derivative of a function \( f: E \to F \), where \( E \) and \( F \) are normed vector spaces. In most cases, \( E = \mathbb{R}^n \) and \( F = \mathbb{R}^m \). However, if we need to deal with infinite dimensional manifolds, then it is necessary to allow \( E \) and \( F \) to be infinite dimensional. We omit most proofs and refer the reader to standard analysis textbooks such as Lang [113, 112], Munkres [135], Choquet-Bruhat [44] or Schwartz [155, 156], for a complete exposition.

Let \( E \) and \( F \) be two normed vector spaces, let \( A \subseteq E \) be some open subset of \( E \), and let \( a \in A \) be some element of \( A \). Even though \( a \) is a vector, we may also call it a point.
The idea behind the derivative of the function \( f \) at \( a \) is that it is a *linear approximation* of \( f \) in a small open set around \( a \). The difficulty is to make sense of the quotient

\[
\frac{f(a + h) - f(a)}{h}
\]

where \( h \) is a vector. We circumvent this difficulty in two stages.

A first possibility is to consider the *directional derivative* with respect to a vector \( u \neq 0 \) in \( E \).

We can consider the vector \( f(a + tu) - f(a) \), where \( t \in \mathbb{R} \) (or \( t \in \mathbb{C} \)). Now,

\[
\frac{f(a + tu) - f(a)}{t}
\]

makes sense.

The idea is that in \( E \), the points of the form \( a + tu \) for \( t \) in some small interval \([-\epsilon, +\epsilon] \) in \( \mathbb{R} \) form a line segment \([r, s]\) in \( A \) containing \( a \), and that the image of this line segment defines a small curve segment on \( f(A) \). This curve segment is defined by the map \( t \mapsto f(a + tu) \), from \([r, s]\) to \( F \), and the directional derivative \( D_u f(a) \) defines the direction of the tangent line at \( a \) to this curve.

**Definition 2.4.** Let \( E \) and \( F \) be two normed spaces, let \( A \) be a nonempty open subset of \( E \), and let \( f: A \to F \) be any function. For any \( a \in A \), for any \( u \neq 0 \) in \( E \), the *directional derivative of \( f \) at \( a \) w.r.t. the vector \( u \)*, denoted by \( D_u f(a) \), is the limit (if it exists)

\[
\lim_{t \to 0, t \in U} \frac{f(a + tu) - f(a)}{t},
\]

where \( U = \{t \in \mathbb{R} \mid a + tu \in A, t \neq 0\} \) (or \( U = \{t \in \mathbb{C} \mid a + tu \in A, t \neq 0\} \)).

Since the map \( t \mapsto a + tu \) is continuous, and since \( A - \{a\} \) is open, the inverse image \( U \) of \( A - \{a\} \) under the above map is open, and the definition of the limit in Definition 2.4 makes sense.

**Remark:** Since the notion of limit is purely topological, the existence and value of a directional derivative is independent of the choice of norms in \( E \) and \( F \), as long as they are equivalent norms.

The directional derivative is sometimes called the *Gâteaux derivative*.

In the special case where \( E = \mathbb{R} \), \( F = \mathbb{R} \) and we let \( u = 1 \) (i.e., the real number 1, viewed as a vector), it is immediately verified that \( D_1 f(a) = f'(a) \). When \( E = \mathbb{R} \) (or \( E = \mathbb{C} \)) and \( F \) is any normed vector space, the derivative \( D_1 f(a) \), also denoted by \( f'(a) \), provides a suitable generalization of the notion of derivative.
However, when \( E \) has dimension \( \geq 2 \), directional derivatives present a serious problem, which is that their definition is not sufficiently uniform. Indeed, there is no reason to believe that the directional derivatives w.r.t. all nonzero vectors \( u \) share something in common. As a consequence, a function can have all directional derivatives at \( a \), and yet not be continuous at \( a \). Two functions may have all directional derivatives in some open sets, and yet their composition may not. Thus, we introduce a more uniform notion.

Given two normed spaces \( E \) and \( F \), recall that a linear map \( f : E \to F \) is continuous iff there is some constant \( C \geq 0 \) such that
\[
\| f(u) \| \leq C \| u \| \quad \text{for all} \quad u \in E.
\]

**Definition 2.5.** Let \( E \) and \( F \) be two normed spaces, let \( A \) be a nonempty open subset of \( E \), and let \( f : A \to F \) be any function. For any \( a \in A \), we say that \( f \) is differentiable at \( a \) if there is a continuous linear map, \( L : E \to F \), and a function, \( \epsilon(h) \), such that
\[
f(a + h) = f(a) + L(h) + \epsilon(h)\|h\|
\]
for every \( a + h \in A \), where
\[
\lim_{h \to 0, \ h \in U} \epsilon(h) = 0,
\]
with \( U = \{ h \in E \mid a + h \in A, \ h \neq 0 \} \). The linear map \( L \) is denoted by \( Df(a) \), or \( Df_a \), or \( df(a) \), or \( df_a \), or \( f'(a) \), and it is called the Fréchet derivative, or total derivative, or derivative, or total differential, or differential, of \( f \) at \( a \).

Since the map \( h \mapsto a + h \) from \( E \) to \( E \) is continuous, and since \( A \) is open in \( E \), the inverse image \( U \) of \( A - \{ a \} \) under the above map is open in \( E \), and it makes sense to say that
\[
\lim_{h \to 0, \ h \in U} \epsilon(h) = 0.
\]

Note that for every \( h \in U \), since \( h \neq 0 \), \( \epsilon(h) \) is uniquely determined since
\[
\epsilon(h) = \frac{f(a + h) - f(a) - L(h)}{\|h\|},
\]
and the value \( \epsilon(0) \) plays absolutely no role in this definition. It does no harm to assume that \( \epsilon(0) = 0 \), and we will assume this from now on.

**Remark:** Since the notion of limit is purely topological, the existence and value of a derivative is independent of the choice of norms in \( E \) and \( F \), as long as they are equivalent norms.

The following proposition shows that our new definition is consistent with the definition of the directional derivative and that the continuous linear map \( L \) is unique, if it exists.
Proposition 2.15. Let $E$ and $F$ be two normed spaces, let $A$ be a nonempty open subset of $E$, and let $f : A \to F$ be any function. For any $a \in A$, if $Df(a)$ is defined, then $f$ is continuous at $a$ and $f$ has a directional derivative $D_u f(a)$ for every $u \neq 0$ in $E$. Furthermore,

$$D_u f(a) = Df(a)(u)$$

and thus, $Df(a)$ is uniquely defined.

Proof. If $L = Df(a)$ exists, then for any nonzero vector $u \in E$, because $A$ is open, for any $t \in \mathbb{R} - \{0\}$ (or $t \in \mathbb{C} - \{0\}$) small enough, $a + tu \in A$, so

$$f(a + tu) = f(a) + L(tu) + \epsilon(tu)||tu||$$

which implies that

$$L(u) = \frac{f(a + tu) - f(a)}{t} - \frac{|t|}{t} \epsilon(tu)||u||,$$

and since $\lim_{t \to 0} \epsilon(tu) = 0$, we deduce that

$$L(u) = Df(a)(u) = D_u f(a).$$

Because

$$f(a + h) = f(a) + L(h) + \epsilon(h)||h||$$

for all $h$ such that $||h||$ is small enough, $L$ is continuous, and $\lim_{h \to 0} \epsilon(h)||h|| = 0$, we have $\lim_{h \to 0} f(a + h) = f(a)$, that is, $f$ is continuous at $a$. \qed

Observe that the uniqueness of $Df(a)$ follows from Proposition 2.15. Also, when $E$ is of finite dimension, it is easily shown that every linear map is continuous and this assumption is then redundant.

As an example, consider the map $f : M_n(\mathbb{R}) \to M_n(\mathbb{R})$ given by

$$f(A) = A^\top A - I,$$

where $M_n(\mathbb{R})$ denotes the vector space of all $n \times n$ matrices with real entries equipped with any matrix norm, since they are all equivalent; for example, pick the Frobenius norm $||A||_F = \sqrt{\text{tr}(A^\top A)}$. We claim that

$$Df(A)(H) = A^\top H + H^\top A,$$

for all $A$ and $H$ in $M_n(\mathbb{R})$.

We have

$$f(A + H) - f(A) - (A^\top H + H^\top A) = (A + H)^\top (A + H) - I - (A^\top A - I) - A^\top H - H^\top A$$

$$= A^\top A + A^\top H + H^\top A + H^\top H - A^\top A - A^\top H - H^\top A$$

$$= H^\top H.$$
It follows that
\[
\epsilon(H) = \frac{f(A + H) - f(A) - (A^T H + H^T A)}{\|H\|} = H^T H \frac{1}{\|H\|},
\]
and since our norm is the Frobenius norm,
\[
\|\epsilon(H)\| = \left\| H^T H \frac{1}{\|H\|} \right\| \leq \frac{\|H^T\|}{\|H\|} \|H\| = \|H\|,
\]
so
\[
\lim_{H \to 0} \epsilon(H) = 0,
\]
and we conclude that
\[
Df(A)(H) = A^T H + H^T A.
\]

If \(Df(a)\) exists for every \(a \in A\), we get a map \(Df : A \to \mathcal{L}(E; F)\), called the derivative of \(f\) on \(A\), and also denoted by \(df\). Here, \(\mathcal{L}(E; F)\) denotes the vector space of continuous linear maps from \(E\) to \(F\).

We now consider a number of standard results about derivatives. A function \(f : E \to F\) is said to be affine if there is some linear map \(\overrightarrow{f} : E \to F\) and some fixed vector \(c \in F\), such that
\[
f(u) = \overrightarrow{f}(u) + c
\]
for all \(u \in E\). We call \(\overrightarrow{f}\) the linear map associated with \(f\).

**Proposition 2.16.** Given two normed spaces \(E\) and \(F\), if \(f : E \to F\) is a constant function, then \(Df(a) = 0\), for every \(a \in E\). If \(f : E \to F\) is a continuous affine map, then \(Df(a) = \overrightarrow{f}\), for every \(a \in E\), where \(\overrightarrow{f}\) denotes the linear map associated with \(f\).

**Proposition 2.17.** Given a normed space \(E\) and a normed vector space \(F\), for any two functions \(f, g : E \to F\), for every \(a \in E\), if \(Df(a)\) and \(Dg(a)\) exist, then \(D(f + g)(a)\) and \(D(\lambda f)(a)\) exist, and
\[
D(f + g)(a) = Df(a) + Dg(a), \\
D(\lambda f)(a) = \lambda Df(a).
\]

Given two normed vector spaces \((E_1, \| \|_1)\) and \((E_2, \| \|_2)\), there are three natural and equivalent norms that can be used to make \(E_1 \times E_2\) into a normed vector space:

1. \(\|(u_1, u_2)\|_1 = \|u_1\|_1 + \|u_2\|_2\).
2. \(\|(u_1, u_2)\|_2 = (\|u_1\|^2_1 + \|u_2\|^2_2)^{1/2}\).
3. \(\|(u_1, u_2)\|_\infty = \max(\|u_1\|_1, \|u_2\|_2)\).
2.2. THE DERIVATIVE OF A FUNCTION BETWEEN NORMED SPACES

We usually pick the first norm. If $E_1$, $E_2$, and $F$ are three normed vector spaces, recall that a bilinear map $f: E_1 \times E_2 \to F$ is \textit{continuous} iff there is some constant $C \geq 0$ such that

$$
\|f(u_1, u_2)\| \leq C \|u_1\|_1 \|u_2\|_2 \quad \text{for all } u_1 \in E_1 \text{ and all } u_2 \in E_2.
$$

\textbf{Proposition 2.18.} Given three normed vector spaces $E_1$, $E_2$, and $F$, for any continuous bilinear map $f: E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, $Df(a, b)$ exists, and for every $u \in E_1$ and $v \in E_2$,

$$
Df(a, b)(u, v) = f(u, b) + f(a, v).
$$

\textbf{Proof.} Since $f$ is bilinear, a simple computation implies that

$$
f((a, b) + (u, v)) - f(a, b) - (f(u, b) + f(a, v)) = f(a + u, b + v) - f(a, b) - f(u, b) - f(a, v)
$$

$$
= f(a + u, b) + f(a + u, v) - f(a, b) - f(u, b) - f(a, v)
$$

$$
= f(a, b) + f(u, b) + f(a, v) + f(u, v) - f(a, b) - f(u, b) - f(a, v)
$$

$$
= f(u, v).
$$

We define

$$
\epsilon(u, v) = \frac{f((a, b) + (u, v)) - f(a, b) - (f(u, b) + f(a, v))}{\|(u, v)\|_1},
$$

and observe that the continuity of $f$ implies

$$
\|f((a, b) + (u, v)) - f(a, b) - (f(u, b) + f(a, v))\| = \|f(u, v)\|
$$

$$
\leq C \|u\|_1 \|v\|_2 \leq C (\|u\|_1 + \|v\|_2)^2.
$$

Hence

$$
\|\epsilon(u, v)\| = \left\| \frac{f(u, v)}{\|(u, v)\|_1} \right\| = \frac{\|f(u, v)\|}{\|(u, v)\|_1} \leq C \left( \frac{\|u\|_1 + \|v\|_2)^2}{\|u\|_1 + \|v\|_2} \right) = C (\|u\|_1 + \|v\|_2) = C \|f(u, v)\|_1,
$$

which in turn implies

$$
\lim_{(u, v) \to (0, 0)} \epsilon(u, v) = 0.
$$

We now state the very useful \textit{chain rule}.

\textbf{Theorem 2.19.} Given three normed spaces $E$, $F$, and $G$, let $A$ be an open set in $E$, and let $B$ an open set in $F$. For any functions $f: A \to F$ and $g: B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, if $Df(a)$ exists and $Dg(f(a))$ exists, then $D(g \circ f)(a)$ exists, and

$$
D(g \circ f)(a) = Dg(f(a)) \circ Df(a).
$$

Theorem 2.19 has many interesting consequences. We mention one corollary.
Proposition 2.20. Given two normed spaces $E$ and $F$, let $A$ be some open subset in $E$, let $B$ be some open subset in $F$, let $f : A \to B$ be a bijection from $A$ to $B$, and assume that $Df$ exists on $A$ and that $Df^{-1}$ exists on $B$. Then, for every $a \in A$,
\[ Df^{-1}(f(a)) = (Df(a))^{-1}. \]

Proposition 2.20 has the remarkable consequence that the two vector spaces $E$ and $F$ have the same dimension. In other words, a local property, the existence of a bijection $f$ between an open set $A$ of $E$ and an open set $B$ of $F$, such that $f$ is differentiable on $A$ and $f^{-1}$ is differentiable on $B$, implies a global property, that the two vector spaces $E$ and $F$ have the same dimension.

Let us mention two more rules about derivatives that are used all the time.

Let $\iota : \text{GL}(n, \mathbb{C}) \to M_n(\mathbb{C})$ be the function (inversion) defined on invertible $n \times n$ matrices by
\[ \iota(A) = A^{-1}. \]
Then, we have
\[ d\iota_A(H) = -A^{-1}HA^{-1}, \]
for all $A \in \text{GL}(n, \mathbb{C})$ and for all $H \in M_n(\mathbb{C})$. In particular, if $A = I$, then $d\iota_I(H) = -H$.

Next, if $f : M_n(\mathbb{C}) \to M_n(\mathbb{C})$ and $g : M_n(\mathbb{C}) \to M_n(\mathbb{C})$ are differentiable matrix functions, then
\[ d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B), \]
for all $A, B \in M_n(\mathbb{C})$. This is known as the product rule.

When $E$ is of finite dimension $n$, for any basis, $(u_1, \ldots, u_n)$, of $E$, we can define the directional derivatives with respect to the vectors in the basis $(u_1, \ldots, u_n)$ (actually, we can also do it for an infinite basis). This way, we obtain the definition of partial derivatives, as follows:

Definition 2.6. For any two normed spaces $E$ and $F$, if $E$ is of finite dimension $n$, for every basis $(u_1, \ldots, u_n)$ for $E$, for every $a \in E$, for every function $f : E \to F$, the directional derivatives $D_{u_j}f(a)$ (if they exist) are called the partial derivatives of $f$ with respect to the basis $(u_1, \ldots, u_n)$. The partial derivative $D_{u_j}f(a)$ is also denoted by $\partial_j f(a)$, or $\frac{\partial f}{\partial x_j}(a)$.

The notation $\frac{\partial f}{\partial x_j}(a)$ for a partial derivative, although customary and going back to Leibniz, is a “logical obscenity.” Indeed, the variable $x_j$ really has nothing to do with the formal definition. This is just another of these situations where tradition is just too hard to overthrow!
If both $E$ and $F$ are of finite dimension, for any basis $(u_1, \ldots, u_n)$ of $E$ and any basis $(v_1, \ldots, v_m)$ of $F$, every function $f: E \to F$ is determined by $m$ functions $f_i: E \to \mathbb{R}$ (or $f_i: E \to \mathbb{C}$), where
\[
f(x) = f_1(x)v_1 + \cdots + f_m(x)v_m,
\]
for every $x \in E$. Then, we get
\[
Df(a)(u_j) = Df_1(a)(u_j)v_1 + \cdots + Df_i(a)(u_j)v_i + \cdots + Df_m(a)(u_j)v_m,
\]
that is,
\[
Df(a)(u_j) = \partial_j f_1(a)v_1 + \cdots + \partial_j f_i(a)v_i + \cdots + \partial_j f_m(a)v_m.
\]
Since the $j$-th column of the $m \times n$-matrix representing $Df(a)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$ is equal to the components of the vector $Df(a)(u_j)$ over the basis $(v_1, \ldots, v_m)$, the linear map $Df(a)$ is determined by the $m \times n$-matrix
\[
J(f)(a) = (\partial_j f_i(a)), \quad J(f)(a) = \left( \frac{\partial f_i}{\partial x_j} (a) \right):
\]
\[
J(f)(a) = \begin{pmatrix}
\partial_1 f_1(a) & \partial_2 f_1(a) & \cdots & \partial_n f_1(a) \\
\partial_1 f_2(a) & \partial_2 f_2(a) & \cdots & \partial_n f_2(a) \\
\vdots & \vdots & \ddots & \vdots \\
\partial_1 f_m(a) & \partial_2 f_m(a) & \cdots & \partial_n f_m(a)
\end{pmatrix},
\]
or
\[
J(f)(a) = \begin{pmatrix}
\frac{\partial f_1}{\partial x_1} (a) & \frac{\partial f_1}{\partial x_2} (a) & \cdots & \frac{\partial f_1}{\partial x_n} (a) \\
\frac{\partial f_2}{\partial x_1} (a) & \frac{\partial f_2}{\partial x_2} (a) & \cdots & \frac{\partial f_2}{\partial x_n} (a) \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f_m}{\partial x_1} (a) & \frac{\partial f_m}{\partial x_2} (a) & \cdots & \frac{\partial f_m}{\partial x_n} (a)
\end{pmatrix}.
\]

This matrix is called the Jacobian matrix of $Df$ at $a$. When $m = n$, the determinant, $\det(J(f)(a))$, of $J(f)(a)$ is called the Jacobian of $Df(a)$.

We know that this determinant only depends on $Df(a)$, and not on specific bases. However, partial derivatives give a means for computing it.

When $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$, for any function $f: \mathbb{R}^n \to \mathbb{R}^m$, it is easy to compute the partial derivatives $\frac{\partial f_i}{\partial x_j} (a)$. We simply treat the function $f_i: \mathbb{R}^n \to \mathbb{R}$ as a function of its $j$-th argument, leaving the others fixed, and compute the derivative as the usual derivative.

**Example 2.1.** For example, consider the function $f: \mathbb{R}^2 \to \mathbb{R}^2$, defined by
\[
f(r, \theta) = (r \cos \theta, r \sin \theta).
\]
Then, we have
\[
J(f)(r, \theta) = \begin{pmatrix}
\cos \theta & -r \sin \theta \\
\sin \theta & r \cos \theta
\end{pmatrix}
\]
and the Jacobian (determinant) has value \(\det(J(f)(r, \theta)) = r\).

In the case where \(E = \mathbb{R}\) (or \(E = \mathbb{C}\)), for any function \(f: \mathbb{R} \to F\) (or \(f: \mathbb{C} \to F\)), the Jacobian matrix of \(Df(a)\) is a column vector. In fact, this column vector is just \(D_1f(a)\). Then, for every \(\lambda \in \mathbb{R}\) (or \(\lambda \in \mathbb{C}\)), \(Df(a)(\lambda) = \lambda D_1f(a)\). This case is sufficiently important to warrant a definition.

**Definition 2.7.** Given a function \(f: \mathbb{R} \to F\) (or \(f: \mathbb{C} \to F\), where \(F\) is a normed space, the vector
\[
Df(a)(1) = D_1f(a)
\]
is called the *vector derivative* or *velocity vector* (in the real case) at \(a\). We usually identify \(Df(a)\) with its Jacobian matrix \(D_1f(a)\), which is the column vector corresponding to \(D_1f(a)\). By abuse of notation, we also let \(Df(a)\) denote the vector \(Df(a)(1) = D_1f(a)\).

When \(E = \mathbb{R}\), the physical interpretation is that \(f\) defines a (parametric) curve that is the trajectory of some particle moving in \(\mathbb{R}^m\) as a function of time, and the vector \(D_1f(a)\) is the *velocity* of the moving particle \(f(t)\) at \(t = a\).

**Example 2.2.**

1. When \(A = (0, 1)\), and \(F = \mathbb{R}^3\), a function \(f: (0, 1) \to \mathbb{R}^3\) defines a (parametric) curve in \(\mathbb{R}^3\). If \(f = (f_1, f_2, f_3)\), its Jacobian matrix at \(a \in \mathbb{R}\) is
\[
J(f)(a) = \begin{pmatrix}
\frac{\partial f_1}{\partial t}(a) \\
\frac{\partial f_2}{\partial t}(a) \\
\frac{\partial f_3}{\partial t}(a)
\end{pmatrix}
\]
2. When \(E = \mathbb{R}^2\), and \(F = \mathbb{R}^3\), a function \(\varphi: \mathbb{R}^2 \to \mathbb{R}^3\) defines a parametric surface. Letting \(\varphi = (f, g, h)\), its Jacobian matrix at \(a \in \mathbb{R}^2\) is
\[
J(\varphi)(a) = \begin{pmatrix}
\frac{\partial f}{\partial u}(a) & \frac{\partial f}{\partial v}(a) \\
\frac{\partial g}{\partial u}(a) & \frac{\partial g}{\partial v}(a) \\
\frac{\partial h}{\partial u}(a) & \frac{\partial h}{\partial v}(a)
\end{pmatrix}
\]
3. When $E = \mathbb{R}^3$, and $F = \mathbb{R}$, for a function $f : \mathbb{R}^3 \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^3$ is

$$J(f)(a) = \begin{bmatrix} \frac{\partial f}{\partial x}(a) & \frac{\partial f}{\partial y}(a) & \frac{\partial f}{\partial z}(a) \end{bmatrix}.$$ 

More generally, when $f : \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^n$ is the row vector

$$J(f)(a) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(a) \\ \vdots \\ \frac{\partial f}{\partial x_n}(a) \end{bmatrix}.$$ 

Its transpose is a column vector called the gradient of $f$ at $a$, denoted by $\text{grad} f(a)$ or $\nabla f(a)$. Then, given any $v \in \mathbb{R}^n$, note that

$$Df(a)(v) = \frac{\partial f}{\partial x_1}(a) v_1 + \cdots + \frac{\partial f}{\partial x_n}(a) v_n = \text{grad} f(a) \cdot v,$$

the scalar product of $\text{grad} f(a)$ and $v$.

When $E$, $F$, and $G$ have finite dimensions, where $(u_1, \ldots, u_p)$ is a basis for $E$, $(v_1, \ldots, v_n)$ is a basis for $F$, and $(w_1, \ldots, w_m)$ is a basis for $G$, if $A$ is an open subset of $E$, $B$ is an open subset of $F$, for any functions $f : A \to F$ and $g : B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, letting $b = f(a)$, and $h = g \circ f$, if $Df(a)$ exists and $Dg(b)$ exists, by Theorem 2.19, the Jacobian matrix $J(h)(a) = J(g \circ f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(w_1, \ldots, w_m)$ is the product of the Jacobian matrices $J(g)(b)$ w.r.t. the bases $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_n)$, and $J(f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$:

$$J(h)(a) = \begin{pmatrix} \frac{\partial g_1}{\partial y_1}(b) & \frac{\partial g_1}{\partial y_2}(b) & \cdots & \frac{\partial g_1}{\partial y_n}(b) \\ \frac{\partial g_2}{\partial y_1}(b) & \frac{\partial g_2}{\partial y_2}(b) & \cdots & \frac{\partial g_2}{\partial y_n}(b) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial y_1}(b) & \frac{\partial g_m}{\partial y_2}(b) & \cdots & \frac{\partial g_m}{\partial y_n}(b) \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) & \cdots & \frac{\partial f_1}{\partial x_p}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) & \cdots & \frac{\partial f_2}{\partial x_p}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(a) & \frac{\partial f_n}{\partial x_2}(a) & \cdots & \frac{\partial f_n}{\partial x_p}(a) \end{pmatrix}.$$ 

Thus, we have the familiar formula

$$\frac{\partial h_i}{\partial x_j}(a) = \sum_{k=1}^n \frac{\partial g_i}{\partial y_k}(b) \frac{\partial f_k}{\partial x_j}(a).$$

Given two normed spaces $E$ and $F$ of finite dimension, given an open subset $A$ of $E$, if a function $f : A \to F$ is differentiable at $a \in A$, then its Jacobian matrix is well defined.

One should be warned that the converse is false. There are functions such that all the partial derivatives exist at some $a \in A$, but yet, the function is not differentiable at $a$, and not even continuous at $a$. 

\[\text{Diagram or Figure} \]
However, there are sufficient conditions on the partial derivatives for $Df(a)$ to exist, namely, continuity of the partial derivatives. If $f$ is differentiable on $A$, then $f$ defines a function $Df: A \to \mathcal{L}(E;F)$. It turns out that the continuity of the partial derivatives on $A$ is a necessary and sufficient condition for $Df$ to exist and to be continuous on $A$. To prove this, we need an important result known as the mean value theorem.

If $E$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$), given any two points $a, b \in E$, the closed segment $[a, b]$ is the set of all points $a + \lambda(b - a)$, where $0 \leq \lambda \leq 1$, $\lambda \in \mathbb{R}$, and the open segment $(a, b)$ is the set of all points $a + \lambda(b - a)$, where $0 < \lambda < 1$, $\lambda \in \mathbb{R}$. The following result is known as the mean value theorem.

**Proposition 2.21.** Let $E$ and $F$ be two normed vector spaces, let $A$ be an open subset of $E$, and let $f: A \to F$ be a continuous function on $A$. Given any $a \in A$ and any $h \neq 0$ in $E$, if the closed segment $[a, a + h]$ is contained in $A$, if $f: A \to F$ is differentiable at every point of the open segment $(a, a + h)$, and if

$$\sup_{x \in (a,a+h)} \|Df(x)\| \leq M$$

for some $M \geq 0$, then

$$\|f(a + h) - f(a)\| \leq M\|h\|.$$ 

As a corollary, if $L: E \to F$ is a continuous linear map, then

$$\|f(a + h) - f(a) - L(h)\| \leq M\|h\|,$$

where $M = \sup_{x \in (a,a+h)} \|Df(x) - L\|$.

A very useful result which is proved using the mean value theorem is the proposition below.

**Proposition 2.22.** Let $f: A \to F$ be any function between two normed vector spaces $E$ and $F$, where $A$ is an open subset of $E$. If $A$ is connected and if $Df(a) = 0$ for all $a \in A$, then $f$ is a constant function on $A$.

The mean value theorem also implies the following important result.

**Theorem 2.23.** Given two normed spaces $E$ and $F$, where $E$ is of finite dimension $n$ and where $(u_1, \ldots, u_n)$ is a basis of $E$, given any open subset $A$ of $E$, given any function $f: A \to F$, the derivative $Df: A \to \mathcal{L}(E;F)$ is defined and continuous on $A$ iff every partial derivative $\partial_j f$ (or $\frac{\partial f}{\partial x_j}$) is defined and continuous on $A$, for all $j$, $1 \leq j \leq n$. As a corollary, if $F$ is of finite dimension $m$, and $(v_1, \ldots, v_m)$ is a basis of $F$, the derivative $Df: A \to \mathcal{L}(E;F)$ is defined and continuous on $A$ iff every partial derivative $\partial_j f_i \left( \text{or} \frac{\partial f_i}{\partial x_j} \right)$ is defined and continuous on $A$, for all $i, j$, $1 \leq i \leq m$, $1 \leq j \leq n$. 
2.2. THE DERIVATIVE OF A FUNCTION BETWEEN NORMED SPACES

Definition 2.8. Given two normed spaces $E$ and $F$, and an open subset $A$ of $E$, we say that a function $f: A \to F$ is a $C^0$-function on $A$ if $f$ is continuous on $A$. We say that $f: A \to F$ is a $C^1$-function on $A$ if $Df$ exists and is continuous on $A$.

Let $E$ and $F$ be two normed spaces, let $U \subseteq E$ be an open subset of $E$ and let $f: E \to F$ be a function such that $Df(a)$ exists for all $a \in U$. If $Df(a)$ is injective for all $a \in U$, we say that $f$ is an immersion (on $U$) and if $Df(a)$ is surjective for all $a \in U$, we say that $f$ is a submersion (on $U$).

When $E$ and $F$ are finite dimensional with $\dim(E) = n$ and $\dim(F) = m$, if $m \geq n$, then $f$ is an immersion if the Jacobian matrix, $J(f)(a)$, has full rank $n$ for all $a \in E$ and if $n \geq m$, then $f$ is a submersion if the Jacobian matrix, $J(f)(a)$, has full rank $m$ for all $a \in E$.

A very important theorem is the inverse function theorem. In order for this theorem to hold for infinite dimensional spaces, it is necessary to assume that our normed spaces are complete. Fortunately, $\mathbb{R}$, $\mathbb{C}$, and every finite dimensional (real or complex) normed vector space is complete. A real (resp. complex) vector space, $E$, is a real (resp. complex) Hilbert space if it is complete as a normed space with the norm $\|u\| = \sqrt{\langle u, u \rangle}$ induced by its Euclidean (resp. Hermitian) inner product (of course, positive, definite).

Definition 2.9. Given two topological spaces $E$ and $F$ and an open subset $A$ of $E$, we say that a function $f: A \to F$ is a local homeomorphism from $A$ to $F$ if for every $a \in A$, there is an open set $U \subseteq A$ containing $a$ and an open set $V$ containing $f(a)$ such that $f$ is a one-to-one, onto, continuous function from $U$ to $V = f(U)$ which has continuous inverse $f^{-1}: V \to U$. If $B$ is an open subset of $F$, we say that $f: A \to F$ is a (global) homeomorphism from $A$ to $B$ if $f$ is a homeomorphism from $A$ to $B = f(A)$.

If $E$ and $F$ are normed spaces, we say that $f: A \to F$ is a local diffeomorphism from $A$ to $F$ if for every $a \in A$, there is an open set $U \subseteq A$ containing $a$ and an open set $V$ containing $f(a)$ such that $f$ is a bijection from $U$ to $V$, $f$ is a $C^1$-function on $U$, and $f^{-1}$ is a $C^1$-function on $V = f(U)$. We say that $f: A \to F$ is a (global) diffeomorphism from $A$ to $B$ if $f$ is a homeomorphism from $A$ to $B = f(A)$, $f$ is a $C^1$-function on $A$, and $f^{-1}$ is a $C^1$-function on $B$.

Note that a local diffeomorphism is a local homeomorphism. Also, as a consequence of Proposition 2.20, if $f$ is a diffeomorphism on $A$, then $Df(a)$ is a linear isomorphism for every $a \in A$.

Theorem 2.24. (Inverse Function Theorem) Let $E$ and $F$ be complete normed spaces, let $A$ be an open subset of $E$, and let $f: A \to F$ be a $C^1$-function on $A$. The following properties hold:
(1) For every \( a \in A \), if \( Df(a) \) is a linear isomorphism (which means that both \( Df(a) \) and \( (Df(a))^{-1} \) are linear and continuous),\(^1\) then there exist some open subset \( U \subseteq A \) containing \( a \), and some open subset \( V \) of \( F \) containing \( f(a) \), such that \( f \) is a diffeomorphism from \( U \) to \( V = f(U) \). Furthermore,

\[
Df^{-1}(f(a)) = (Df(a))^{-1}.
\]

For every neighborhood \( N \) of \( a \), the image \( f(N) \) of \( N \) is a neighborhood of \( f(a) \), and for every open ball \( U \subseteq A \) of center \( a \), the image \( f(U) \) of \( U \) contains some open ball of center \( f(a) \).

(2) If \( Df(a) \) is invertible for every \( a \in A \), then \( B = f(A) \) is an open subset of \( F \), and \( f \) is a local diffeomorphism from \( A \) to \( B \). Furthermore, if \( f \) is injective, then \( f \) is a diffeomorphism from \( A \) to \( B \).

Proofs of the Inverse function theorem can be found in Lang [112], Abraham and Marsden [1], Schwartz [156], and Cartan [36]. Part (1) of Theorem 2.24 is often referred to as the “(local) inverse function theorem.” It plays an important role in the study of manifolds and (ordinary) differential equations.

If \( E \) and \( F \) are both of finite dimension, the case where \( Df(a) \) is just injective or just surjective is also important for defining manifolds, using implicit definitions.

Suppose as before that \( f: A \to F \) is a function from some open subset \( A \) of \( E \), with \( E \) and \( F \) two normed vector spaces. If \( Df: A \to \mathcal{L}(E;F) \) exists for all \( a \in A \), then we can consider taking the derivative \( DDf(a) \) of \( Df \) at \( a \). If it exists, \( DDf(a) \) is a continuous linear map in \( \mathcal{L}(E;\mathcal{L}(E;F)) \), and we denote \( DDf(a) \) as \( D^2f(a) \). It is known that the vector space \( \mathcal{L}(E;\mathcal{L}(E;F)) \) is isomorphic to the vector space of continuous bilinear maps \( \mathcal{L}_2(E^2;F) \), so we can view \( D^2f(a) \) as a bilinear map in \( \mathcal{L}_2(E^2;F) \). It is also known by Schwarz’s lemma that \( D^2f(a) \) is symmetric (partial derivatives commute; see Schwartz [156]). Therefore, for every \( a \in A \), where it exists, \( D^2f(a) \) belongs to the space \( \text{Sym}_2(E^2;F) \) of continuous symmetric bilinear maps from \( E^2 \) to \( F \). If \( E \) has finite dimension \( n \) and \( F = \mathbb{R} \), with respect to any basis \( (e_1, \ldots, e_n) \) of \( E \), \( D^2f(a)(u,v) \) is the value of the quadratic form \( u^\top \text{Hess} f(a)v \), where

\[
\text{Hess} f(a) = \left( \frac{\partial^2 f}{\partial x_i \partial x_j} (a) \right)
\]

is the Hessian matrix of \( f \) at \( a \).

By induction, if \( D^m f: A \to \text{Sym}_m(E^m;F) \) exists for \( m \geq 1 \), where \( \text{Sym}_m(E^m;F) \) denotes the vector space of continuous symmetric multilinear maps from \( E^m \) to \( F \), and if \( DD^m f(a) \) exists for all \( a \in A \), we obtain the \((m+1)\)th derivative \( D^{m+1} f \) of \( f \), and

\(^1\)Actually, since \( E \) and \( F \) are Banach spaces, by the Open Mapping Theorem, it is sufficient to assume that \( Df(a) \) is continuous and bijective; see Lang [112].
$D^{m+1}f \in \text{Sym}_{m+1}(E^{m+1}; F)$, where $\text{Sym}_{m+1}(E^{m+1}; F)$ is the vector space of continuous symmetric multilinear maps from $E^{m+1}$ to $F$.

For any $m \geq 1$, we say that the map $f: A \to F$ is a $C^m$ function (or simply that $f$ is $C^m$) if $Df, D^2f, \ldots, D^mf$ exist and are continuous on $A$.

We say that $f$ is $C^\infty$ or smooth if $D^mf$ exists and is continuous on $A$ for all $m \geq 1$. If $E$ has finite dimension $n$, it can be shown that $f$ is smooth iff all of its partial derivatives

$$\frac{\partial^m f}{\partial x_{i_1} \cdots \partial x_{i_m}}(a)$$

are defined and continuous for all $a \in A$, all $m \geq 1$, and all $i_1, \ldots, i_m \in \{1, \ldots, n\}$.

The function $f: A \to F$ is a $C^m$ diffeomorphism between $A$ and $B = f(A)$ if $f$ is a bijection from $A$ to $B$ and if $f$ and $f^{-1}$ are $C^m$. Similarly, $f$ is a smooth diffeomorphism between $A$ and $B = f(A)$ if $f$ is a bijection from $A$ to $B$ and if $f$ and $f^{-1}$ are smooth.

### 2.3 Linear Vector Fields and the Exponential

We can apply Propositions 2.9 and 2.10 to the map $f: t \mapsto e^{tA}$, where $A$ is any matrix $A \in \text{M}_n(\mathbb{C})$. This power series has a infinite radius of convergence, and we have

$$f'(t) = \sum_{k=1}^{\infty} k^{t^{k-1}A^k} k! = A \sum_{k=1}^{\infty} \frac{t^{k-1}A^{k-1}}{(k-1)!} = Ae^{tA}.$$ 

Note that

$$Ae^{tA} = e^{tA}A.$$ 

Given some open subset $A$ of $\mathbb{R}^n$, a vector field $X$ on $A$ is a function $X: A \to \mathbb{R}^n$, which assigns to every point $p \in A$ a vector $X(p) \in \mathbb{R}^n$. Usually, we assume that $X$ is at least $C^1$ on $A$. For example, if $f: A \to \mathbb{R}$ is a $C^1$ function, then its gradient defines a vector field $X$; namely, $p \mapsto \text{grad } f(p)$. For example if $f: \mathbb{R}^2 \to \mathbb{R}$ is $f(x, y) = \cos(xy^2)$, the gradient vector field $X$ is $(-y^2 \sin(xy^2), -2xy \sin(xy^2)) \equiv (X_1, X_2)$. Note that

$$\frac{\partial X_1}{\partial y} = -2y \sin(xy^2) - 2xy^3 \cos(xy^2) = \frac{\partial X_2}{\partial x}.$$ 

In general, if $f$ is $C^2$, then its second partials commute; that is,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(p) = \frac{\partial^2 f}{\partial x_j \partial x_i}(p), \quad 1 \leq i, j \leq n,$$

so this vector field $X = (X_1, \ldots, X_n)$ has a very special property:

$$\frac{\partial X_i}{\partial x_j} = \frac{\partial X_j}{\partial x_i}, \quad 1 \leq i, j \leq n.$$
This is a necessary condition for a vector field to be the gradient of some function, but not a sufficient condition in general. The existence of such a function depends on the topological shape of the domain $A$. Understanding what are sufficient conditions to answer the above question led to the development of differential forms and cohomology.

**Definition 2.10.** Given a vector field $X : A \to \mathbb{R}^n$, for any point $p_0 \in A$, a $C^1$ curve $\gamma : (-\epsilon, \epsilon) \to \mathbb{R}^n$ (with $\epsilon > 0$) is an integral curve for $X$ with initial condition $p_0$ if $\gamma(0) = p_0$, and

$$\gamma'(t) = X(\gamma(t)) \quad \text{for all } t \in (-\epsilon, \epsilon).$$

Thus, an integral curve has the property that for every time $t \in (-\epsilon, \epsilon)$, the tangent vector $\gamma'(t)$ to the curve $\gamma$ at the point $\gamma(t)$ coincides with the vector $X(\gamma(t))$ given by the vector field at the point $\gamma(t)$.

**Definition 2.11.** Given a $C^1$ vector field $X : A \to \mathbb{R}^n$, for any point $p_0 \in A$, a local flow for $X$ at $p_0$ is a function

$$\varphi : J \times U \to \mathbb{R}^n,$$

where $J \subseteq \mathbb{R}$ is an open interval containing 0 and $U$ is an open subset of $A$ containing $p_0$, so that for every $p \in U$, the curve $t \mapsto \varphi(t, p)$ is an integral curve of $X$ with initial condition $p$.

The theory of ODE tells us that if $X$ is $C^1$, then for every $p_0 \in A$, there is a pair $(J, U)$ as above such that there is a unique $C^1$ local flow $\varphi : J \times U \to \mathbb{R}^n$ for $X$ at $p_0$.

Let us now consider the special class of vector fields induced by matrices in $M_n(\mathbb{R})$. For any matrix $A \in M_n(\mathbb{R})$, let $X_A$ be the vector field given by

$$X_A(p) = Ap \quad \text{for all } p \in \mathbb{R}^n.$$

Such vector fields are obviously $C^1$ (in fact, $C^\infty$).
Figure 2.2: A portion of local flow $\varphi: J \times U \to \mathbb{R}^2$. If $p$ is fixed and $t$ varies, the flow moves along one of the colored curves. If $t$ is fixed and $p$ varies, $p$ acts as a parameter for the individually colored curves.

The vector field induced by the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

is shown in Figure 2.3. Integral curves are circles of center $(0,0)$.

Then, it turns out that the local flows of $X_A$ are global, in the sense that $J = \mathbb{R}$ and $U = \mathbb{R}^n$, and that they are given by the matrix exponential.

**Proposition 2.25.** For any matrix $A \in M_n(\mathbb{R})$, for any $p_0 \in \mathbb{R}^n$, there is a unique local flow $\varphi: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ for the vector field $X_A$ given by

$$\varphi(t, p) = e^{tA}p,$$

for all $t \in \mathbb{R}$ and all $p \in \mathbb{R}^n$.

**Proof.** For any $p \in \mathbb{R}^n$, write $\gamma_p(t) = \varphi(t, p)$. We claim that $\gamma_p(t) = e^{tA}p$ is the unique integral curve for $X_A$ with initial condition $p$.

We have

$$\gamma'_p(t) = (e^{tA}p)'(t) = Ae^{tA}p = A\gamma_p(t) = X_A(\gamma_p(t)),$$

which shows that $\gamma_p$ is an integral curve for $X_A$ with initial condition $p$. 
Say \( \theta \) is another integral curve for \( X_A \) with initial condition \( p \). Let us compute the derivative of the function \( t \mapsto e^{-tA}\theta(t) \). Using the product rule and the fact that \( \theta'(t) = X_A(\theta(t)) = A\theta(t) \), we have

\[
(e^{-tA}\theta)'(t) = (e^{-tA})'(t)\theta(t) + e^{-tA}\theta'(t)
\]
\[
= e^{-tA}(-A)\theta(t) + e^{-tA}A\theta(t)
\]
\[
= -e^{-tA}A\theta(t) + e^{-tA}A\theta(t) = 0.
\]

Therefore, by Proposition 2.22, the function \( t \mapsto e^{-tA}\theta(t) \) is constant on \( \mathbb{R} \). Furthermore, since \( \theta(0) = p \), its value is \( p \), so

\[
e^{-tA}\theta(t) = p \quad \text{for all } t \in \mathbb{R}.
\]

Therefore, \( \theta(t) = e^{tA}p = \gamma_p(t) \), establishing uniqueness.

For \( t \) fixed, the map \( \Phi_t: p \mapsto e^{tA}p \) is a smooth diffeomorphism of \( \mathbb{R}^n \) (with inverse given by \( e^{-tA} \)). We can think of \( \Phi_t \) as the map which, given any \( p \), moves \( p \) along the integral curve \( \gamma_p \) from \( p \) to \( \gamma_p(t) = e^{tA}p \). For the vector field of Figure 2.3, each \( \Phi_t \) is the rotation

\[
e^{tA} = \begin{pmatrix}
\cos t & -\sin t \\
\sin t & \cos t
\end{pmatrix}.
\]

The map \( \Phi: \mathbb{R} \to \text{Diff}(\mathbb{R}^n) \) is a group homomorphism, because

\[
\Phi_s \circ \Phi_t = e^{sA}e^{tA}p = e^{(s+t)A}p = \Phi_{s+t} \quad \text{for all } s, t \in \mathbb{R}.
\]

Observe that \( \Phi_t(p) = \varphi(t, p) \). If we hold \( p \) fixed, we obtain the integral curve with initial condition \( p \), which is also called a flow line of the local flow. If we hold \( t \) fixed, we obtain a smooth diffeomorphism of \( \mathbb{R}^n \) (moving \( p \) to \( \varphi(t, p) \)). The family \( \{\Phi_t\}_{t \in \mathbb{R}} \) is called the 1-parameter group generated by \( X_A \), and \( \Phi \) is called the (global) flow generated by \( X_A \).
2.4. THE ADJOINT REPRESENTATIONS

In the case of $2 \times 2$ matrices, it is possible to describe explicitly the shape of all integral curves; see Rossmann [146] (Section 1.1).

We conclude this chapter by introducing the adjoint representations of $\text{GL}(n, \mathbb{R})$ and $\mathfrak{gl}(n, \mathbb{R})$.

2.4 The Adjoint Representations $\text{Ad}$ and $\text{ad}$ and the derivative of $\exp$

Given any two vector spaces $E$ and $F$, recall that the vector space of all linear maps from $E$ to $F$ is denoted by $\text{Hom}(E, F)$. The set of all invertible linear maps from $E$ to itself is a group (under composition) denoted $\text{GL}(E)$. When $E = \mathbb{R}^n$, we often denote $\text{GL}(\mathbb{R}^n)$ by $\text{GL}(n, \mathbb{R})$ (and if $E = \mathbb{C}^n$, we often denote $\text{GL}(\mathbb{C}^n)$ by $\text{GL}(n, \mathbb{C})$). The vector space $M_n(\mathbb{R})$ of all $n \times n$ matrices is also denoted by $\mathfrak{gl}(n, \mathbb{R})$ (and $M_n(\mathbb{C})$ by $\mathfrak{gl}(n, \mathbb{C})$). Then, $\text{GL}(\mathfrak{gl}(n, \mathbb{R}))$ is the group of all invertible linear maps from $\mathfrak{gl}(n, \mathbb{R}) = M_n(\mathbb{R})$ to itself.

For any matrix $A \in M_n(\mathbb{R})$ (or $A \in M_n(\mathbb{C})$), define the maps $L_A: M_n(\mathbb{R}) \to M_n(\mathbb{R})$ and $R_A: M_n(\mathbb{R}) \to M_n(\mathbb{R})$ by

$$L_A(B) = AB, \quad R_A(B) = BA,$$

for all $B \in M_n(\mathbb{R})$.

Observe that $L_A \circ R_B = R_B \circ L_A$ for all $A, B \in M_n(\mathbb{R})$.

For any matrix $A \in \text{GL}(n, \mathbb{R})$, let

$$\text{Ad}_A: M_n(\mathbb{R}) \to M_n(\mathbb{R}) \quad \text{(conjugation by $A$)}$$

be given by

$$\text{Ad}_A(B) = ABA^{-1} \quad \text{for all } B \in M_n(\mathbb{R}).$$

Observe that $\text{Ad}_A = L_A \circ R_{A^{-1}}$ and that $\text{Ad}_A$ is an invertible linear map with inverse $\text{Ad}_{A^{-1}}$.

The restriction of $\text{Ad}_A$ to invertible matrices $B \in \text{GL}(n, \mathbb{R})$ yields the map

$$\text{Ad}_A: \text{GL}(n, \mathbb{R}) \to \text{GL}(n, \mathbb{R})$$

also given by

$$\text{Ad}_A(B) = ABA^{-1} \quad \text{for all } B \in \text{GL}(n, \mathbb{R}).$$

This time, observe that $\text{Ad}_A$ is a group homomorphism (with respect to multiplication), since

$$\text{Ad}_A(BC) = ABCA^{-1} = ABA^{-1}ACA^{-1} = \text{Ad}_A(B)\text{Ad}_A(C), \quad \text{for all } B, C \in \text{GL}(n, \mathbb{R}).$$

In fact, $\text{Ad}_A$ is a group isomorphism (since its inverse is $\text{Ad}_{A^{-1}}$).

Beware that $\text{Ad}_A$ is not a linear map on $\text{GL}(n, \mathbb{R})$ because $\text{GL}(n, \mathbb{R})$ is not a vector space! Indeed, $\text{GL}(n, \mathbb{R})$ is not closed under addition.
CHAPTER 2. BASIC ANALYSIS: REVIEW OF SERIES AND DERIVATIVES

Nevertheless, we can define the derivative of \( \text{Ad}_A: M_n(\mathbb{R}) \rightarrow M_n(\mathbb{R}) \) with \( A \in \text{GL}(n, \mathbb{R}) \) and \( B, X \in M_n(\mathbb{R}) \) by

\[
\text{Ad}_A(B + X) - \text{Ad}_A(B) = A(B + X)A^{-1} - ABA^{-1} = AXA^{-1},
\]

which shows that \( d(\text{Ad}_A) \) exists and is given by

\[
d(\text{Ad}_A)_B(X) = AXA^{-1}, \quad \text{for all } X \in M_n(\mathbb{R}).
\]

In particular, for \( B = I \), we see that the derivative \( d(\text{Ad}_A)_I \) of \( \text{Ad}_A \) at \( I \) is a linear map of \( \mathfrak{gl}(n, \mathbb{R}) = M_n(\mathbb{R}) \) denoted by \( \text{Ad}(A) \) or \( \text{Ad}_A \) (or \( \text{Ad} A \)), and given by

\[
\text{Ad}_A(X) = AXA^{-1} \quad \text{for all } X \in \mathfrak{gl}(n, \mathbb{R}).
\]

The inverse of \( \text{Ad}_A \) is \( \text{Ad}_A^{-1} \), so \( \text{Ad}_A \in \text{GL}(\mathfrak{gl}(n, \mathbb{R})) \). Note that

\[
\text{Ad}_{AB} = \text{Ad}_A \circ \text{Ad}_B,
\]

so the map \( A \mapsto \text{Ad}_A \) is a group homomorphism denoted

\[
\text{Ad}: \text{GL}(n, \mathbb{R}) \rightarrow \text{GL}(\mathfrak{gl}(n, \mathbb{R})).
\]

The homomorphism \( \text{Ad} \) is called the adjoint representation of \( \text{GL}(n, \mathbb{R}) \).

We also would like to compute the derivative \( d(\text{Ad})_I \) of \( \text{Ad} \) at \( I \). If it exists, it is a linear map

\[
d(\text{Ad})_I: \mathfrak{gl}(n, \mathbb{R}) \rightarrow \text{Hom}(\mathfrak{gl}(n, \mathbb{R}), \mathfrak{gl}(n, \mathbb{R})).
\]

For all \( X, Y \in M_n(\mathbb{R}) \), with \( ||X|| \) small enough we have \( I + X \in \text{GL}(n, \mathbb{R}) \), and

\[
\text{Ad}_{I+X}(Y) - \text{Ad}_I(Y) - (XY - YX) = (I + X)Y(I + X)^{-1} - Y - XY + YX
\]

\[
= [(I + X)Y - Y(I + X) - XY(I + X) + YX(I + X)](I + X)^{-1}
\]

\[
= [Y + XY - Y - YX - XY - XYX + YX + YX^2](I + X)^{-1}
\]

\[
= (YX^2 - XYX)(I + X)^{-1}.
\]

Then, if we let

\[
\epsilon(X, Y) = \frac{(YX^2 - XYX)(I + X)^{-1}}{||X||},
\]

since \( || \cdot || \) is a matrix norm, we get

\[
||\epsilon(X, Y)|| = \frac{||YX^2 - XYX|| ||(I + X)^{-1}||}{||X||} \leq \frac{(||YX^2|| + ||XYX||) ||(I + X)^{-1}||}{||X||}
\]

\[
= \frac{(||X^2|| ||Y|| + ||X|| ||Y|| ||X||) ||(I + X)^{-1}||}{||X||} \leq \frac{2 ||Y|| ||X||^2 ||(I + X)^{-1}||}{||X||}
\]

\[
= 2 ||X|| ||Y|| ||(I + X)^{-1}||.
\]
Therefore, we proved that for $\|X\|$ small enough

$$\text{Ad}_{I+X}(Y) - \text{Ad}_I(Y) = (XY - YX) + \varepsilon(X, Y) \|X\|,$$

with $\|\varepsilon(X, Y)\| \leq 2 \|X\| \|Y\| \|(I + X)^{-1}\|$, and with $\varepsilon(X, Y)$ linear in $Y$.

Let $\text{ad}_X : \mathfrak{gl}(n, \mathbb{R}) \to \mathfrak{gl}(n, \mathbb{R})$ be the linear map given by

$$\text{ad}_X(Y) = XY - YX = [X, Y],$$

and $\text{ad}$ be the linear map

$$\text{ad} : \mathfrak{gl}(n, \mathbb{R}) \to \text{Hom}(\mathfrak{gl}(n, \mathbb{R}), \mathfrak{gl}(n, \mathbb{R}))$$

given by

$$\text{ad}(X) = \text{ad}_X.$$

We also define $\epsilon_X : \mathfrak{gl}(n, \mathbb{R}) \to \mathfrak{gl}(n, \mathbb{R})$ as the linear map given by

$$\epsilon_X(Y) = \epsilon(X, Y).$$

If $\|\epsilon_X\|$ is the operator norm of $\epsilon_X$, we have

$$\|\epsilon_X\| = \max_{\|Y\|=1} \|\epsilon(X, Y)\| \leq 2 \|X\| \|(I + X)^{-1}\|.$$

Then, the equation

$$\text{Ad}_{I+X}(Y) - \text{Ad}_I(Y) = (XY - YX) + \epsilon(X, Y) \|X\|,$$

which holds for all $Y$, yields

$$\text{Ad}_{I+X} - \text{Ad}_I = \text{ad}_X + \epsilon_X \|X\|,$$

and because $\|\epsilon_X\| \leq 2 \|X\| \|(I + X)^{-1}\|$, we have $\lim_{X \to 0} \epsilon_X = 0$, which shows that $d(\text{Ad})_I(X) = \text{ad}_X$; that is,

$$d(\text{Ad})_I = \text{ad}.$$

The notation $\text{ad}(X)$ (or $\text{ad} X$) is also used instead $\text{ad}_X$. The map $\text{ad}$ is a linear map

$$\text{ad} : \mathfrak{gl}(n, \mathbb{R}) \to \text{Hom}(\mathfrak{gl}(n, \mathbb{R}), \mathfrak{gl}(n, \mathbb{R}))$$

called the *adjoint representation* of $\mathfrak{gl}(n, \mathbb{R})$. The Lie algebra $\text{Hom}(\mathfrak{gl}(n, \mathbb{R}), \mathfrak{gl}(n, \mathbb{R}))$ of the group $\text{GL}(\mathfrak{gl}(n, \mathbb{R}))$ is also denoted by $\mathfrak{gl}(\mathfrak{gl}(n, \mathbb{R}))$. 
Since
\[
\text{ad}([X, Y])(Z) = \text{ad}(XY - YX)(Z) = (XY - YX)Z - Z(XY - YX)
\]
\[
= XYZ - YXZ - ZXY + ZYX
\]
whenever \(X, Y, Z \in \mathfrak{gl}(n, \mathbb{R})\), we find that
\[
\text{ad}([X, Y]) = \text{ad}(X)\text{ad}(Y) - \text{ad}(Y)\text{ad}(X) = [\text{ad}(X), \text{ad}(Y)].
\]
This means that \(\text{ad}\) is a Lie algebra homomorphism. It can be checked that this property is equivalent to the following identity known as the \textit{Jacobi identity}:
\[
[X, [Y, Z]] + [Z, [X, Y]] + [Y, [Z, X]] = 0,
\]
for all \(X, Y, Z \in \mathfrak{gl}(n, \mathbb{R})\). Note that
\[
ad_X = L_X - R_X.
\]

Finally, we prove a formula relating \(\text{Ad}\) and \(\text{ad}\) through the exponential. For this, we view \(\text{ad}_X\) and \(\text{Ad}_A\) as an \(n^2 \times n^2\) matrices, for example, over the basis \((E_{ij})\) of \(n \times n\) matrices whose entries are all 0 except for the entry of index \((i, j)\) which is equal to 1.

**Proposition 2.26.** For any \(X \in M_n(\mathbb{R}) = \mathfrak{gl}(n, \mathbb{R})\), we have
\[
\text{Ad}_{e^X} = e^{\text{ad}_X} = \sum_{k=0}^{\infty} \frac{\text{ad}_X^k}{k!};
\]
that is,
\[
e^X Y e^{-X} = e^{\text{ad}_X} Y = Y + [X, Y] + \frac{1}{2!} [X, [X, Y]] + \frac{1}{3!} [X, [X, [X, Y]]] + \cdots
\]
for all \(X, Y \in M_n(\mathbb{R})\).

**Proof.** Let
\[
A(t) = \text{Ad}_{e^{tx}},
\]
pick any \(Y \in M_n(\mathbb{R})\), and compute the derivative of \(A(t)Y\). By the product rule we have
\[
(A(t)Y)'(t) = (e^{tx}Ye^{-tx})'(t)
\]
\[
= Xe^{tx}Ye^{-tx} + e^{tx}Ye^{-tx}(-X)
\]
\[
= Xe^{tx}Ye^{-tx} - e^{tx}Ye^{-tx}X
\]
\[
= \text{ad}_X(\text{Ad}_{e^{tx}}Y) = \text{ad}_X(A(t)Y).
\]
2.4. THE ADJOINT REPRESENTATIONS

We also have \( A(0)Y = \text{Ad}_Y Y = Y \). Therefore, the curve \( t \mapsto A(t)Y \) is an integral curve for the vector field \( X_{\text{ad}_X} \) with initial condition \( Y \), and by Proposition 2.25 (with \( n \) replaced by \( n^2 \)), this unique integral curve is given by

\[
\gamma(t) = e^{t\text{ad}_X}Y,
\]

which proves our assertion.

It is also possible to find a formula for the derivative \( d\exp_A \) of the exponential map at \( A \), but this is a bit tricky. It can be shown that

\[
d(\exp)_A = e^A \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} (\text{ad}_A)^k = d(\exp)_A = e^{L_A} \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!} (L_A - R_A)^j,
\]

so

\[
d(\exp)_A(B) = e^A \left( B - \frac{1}{2!}[A, B] + \frac{1}{3!}[A, [A, B]] - \frac{1}{4!}[A, [A, [A, B]]] + \cdots \right).
\]

It is customary to write

\[
\frac{id - e^{-\text{ad}_A}}{\text{ad}_A}
\]

for the power series

\[
\sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} (\text{ad}_A)^k,
\]

and the formula for the derivative of \( \exp \) is usually stated as

\[
d(\exp)_A = e^A \left( \frac{id - e^{-\text{ad}_A}}{\text{ad}_A} \right).
\]

Most proofs I am aware of use some tricks involving ODE’s, but there is a simple and direct way to prove the formula based on the fact that \( \text{ad}_A = L_A - R_A \) and that \( L_A \) and \( R_A \) commute. First, one can show that

\[
d(\exp)_A = \sum_{h,k \geq 0} \frac{L_A^h R_A^k}{(h+k+1)!}.
\]

Thus, we need to prove that

\[
e^{L_A} \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!} (L_A - R_A)^j = \sum_{h,k \geq 0} \frac{L_A^h R_A^k}{(h+k+1)!}.
\]

To simplify notation, write \( a \) for \( L_A \) and \( b \) for \( L_B \). We wish to prove that

\[
e^a \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!} (a - b)^j = \sum_{h,k \geq 0} \frac{d^h b^k}{(h+k+1)!},
\]

(*)
assuming that \( ab = ba \). This can be done by finding the coefficient of the monomial \( a^h b^k \) on the left hand side. We find that this coefficient is

\[
\frac{1}{(h + k + 1)!} \sum_{i=0}^{h} (-1)^{h-i} \binom{h + k + 1}{i} \binom{h + k - i}{k}.
\]

Therefore, to prove (*) we need to prove that

\[
\sum_{i=0}^{h} (-1)^{h-i} \binom{h + k + 1}{i} \binom{h + k - i}{k} = 1.
\]

The above identity can be shown in various ways. A brute force method is to use induction. One can also use “negation of the upper index” and a Vandermonde convolution to obtain a two line proof. The details are left as an exercise.

The formula for the exponential tells us when the derivative \( d(\exp)_A \) is invertible. Indeed, if the eigenvalues of the matrix \( X \) are \( \lambda_1, \ldots, \lambda_n \), then the eigenvalues of the matrix

\[
\text{id} - e^{-X} X = \sum_{k=0}^{\infty} \frac{(-X)^k}{(k+1)!} X^k
\]

are

\[
\frac{1 - e^{-\lambda_j}}{\lambda_j} \quad \text{if } \lambda_j \neq 0, \text{ and } 1 \quad \text{if } \lambda_j = 0.
\]

To see why this is the case, assume \( \lambda \neq 0 \) is an eigenvalue of \( X \) with eigenvector \( u \), i.e. \( X u = \lambda u \). Then \( (-X)^k u = -\lambda^k u \) for any nonnegative integer \( k \) and

\[
\text{id} - e^{-X} X = \sum_{k=0}^{\infty} \frac{(-X)^k}{(k+1)!} u = \left[ 1 - \frac{X}{2!} + \frac{X^2}{3!} - \frac{X^3}{4!} + \frac{X^4}{5!} + \ldots \right] u
\]

\[
= \left[ 1 - \frac{1}{2!} \lambda + \frac{1}{3!} \lambda^2 - \frac{1}{4!} \lambda^3 + \frac{1}{5!} \lambda^4 + \ldots \right] u
\]

\[
= \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{(k+1)!} u = \frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{(-\lambda)^{k+1}}{(k+1)!} u
\]

\[
= \frac{1 - e^{-\lambda}}{\lambda} u.
\]

It follows that the matrix \( \frac{\text{id} - e^{-X}}{X} \) is invertible iff no \( \lambda_j \) if of the form \( k2\pi i \) for some \( k \in \mathbb{Z} \), so \( d(\exp)_A \) is invertible iff no eigenvalue of \( \text{ad}_A \) is of the form \( k2\pi i \) for some \( k \in \mathbb{Z} \). However, it can also be shown that if the eigenvalues of \( A \) are \( \lambda_1, \ldots, \lambda_n \), then the eigenvalues of \( \text{ad}_A \) are the \( \lambda_i - \lambda_j \), with \( 1 \leq i, j \leq n \). In conclusion, \( d(\exp)_A \) is invertible iff for all \( i, j \) we have

\[
\lambda_i - \lambda_j \neq k2\pi i, \quad k \in \mathbb{Z}.
\]
This suggests defining the following subset $\mathcal{E}(n)$ of $M_n(\mathbb{R})$. The set $\mathcal{E}(n)$ consists of all matrices $A \in M_n(\mathbb{R})$ whose eigenvalue $\lambda + i\mu$ of $A$ ($\lambda, \mu \in \mathbb{R}$) lie in the horizontal strip determined by the condition $-\pi < \mu < \pi$. Then, it is clear that the matrices in $\mathcal{E}(n)$ satisfy the condition $(\ast)$, so $d(\exp)_A$ is invertible for all $A \in \mathcal{E}(n)$. By the inverse function theorem, the exponential map is a local diffeomorphism between $\mathcal{E}(n)$ and $\exp(\mathcal{E}(n))$. Remarkably, more is true: the exponential map is a diffeomorphism between $\mathcal{E}(n)$ and $\exp(\mathcal{E}(n))$ (in particular, it is a bijection). This takes quite a bit of work to be proved. For example, see Mnemné and Testard [130], Chapter 3, Theorem 3.8.4 (see also Bourbaki [28], Chapter III, Section 6.9, Proposition 17, and also Theorem 6). We have the following result.

**Theorem 2.27.** The restriction of the exponential map to $\mathcal{E}(n)$ is a diffeomorphism of $\mathcal{E}(n)$ onto its image $\exp(\mathcal{E}(n))$. Furthermore, $\exp(\mathcal{E}(n))$ consists of all invertible matrices that have no real negative eigenvalues; it is an open subset of $\text{GL}(n, \mathbb{R})$; it contains the open ball $B(I, 1) = \{ A \in \text{GL}(n, \mathbb{R}) \mid \| A - I \| < 1 \}$, for every matrix norm $\| \|$ on $n \times n$ matrices.

Theorem 2.27 has some practical applications because there are algorithms for finding a real log of a matrix with no real negative eigenvalues; for more on applications of Theorem 2.27 to medical imaging, see Chapter 18.
Chapter 3

A Review of Point Set Topology

This chapter contains a review of the topological concepts necessary for studying differential geometry and contains the following material:

1. The definition of a topological space in terms of open sets;
2. The definition of a basis for a topology;
3. The definition of the subspace topology;
4. The definition of the product topology;
5. The definition of continuity and notion of a homeomorphism;
6. The definition of a limit of a sequence;
7. The definition of connectivity and path-wise connectivity;
8. The definition of compactness;
9. The definition of the quotient topology.

Readers familiar with this material may proceed to Chapter 5.

3.1 Topological Spaces

We begin with the notion of a topological space.

**Definition 3.1.** Given a set $E$, a topology on $E$ (or a topological structure on $E$), is defined as a family $\mathcal{O}$ of subsets of $E$, called open sets, which satisfy the following three properties:

1. For every finite family $(U_i)_{1 \leq i \leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.
(2) For every arbitrary family \((U_i)_{i \in I}\) of sets \(U_i \in \mathcal{O}\), we have \(\bigcup_{i \in I} U_i \in \mathcal{O}\), i.e., \(\mathcal{O}\) is closed under arbitrary unions.

(3) \(\emptyset \in \mathcal{O}\), and \(E \in \mathcal{O}\), i.e., \(\emptyset\) and \(E\) belong to \(\mathcal{O}\).

A set \(E\) together with a topology \(\mathcal{O}\) on \(E\) is called a \textit{topological space}. Given a topological space \((E, \mathcal{O})\), a subset \(F\) of \(E\) is a \textit{closed set} if \(F = E - U\) for some open set \(U \in \mathcal{O}\), i.e., \(F\) is the complement of some open set.

It is possible that an open set is also a closed set. For example, \(\emptyset\) and \(E\) are both open and closed. When a topological space contains a proper nonempty subset \(U\) which is both open and closed, the space \(E\) is said to be \textit{disconnected}.

The reader is probably familiar with a certain class of topological spaces known as metric spaces. Recall that a \textit{metric space} is a set \(E\) together with a function \(d : E \times E \to \mathbb{R}_+\), called a \textit{metric}, or distance, assigning a nonnegative real number \(d(x, y)\) to any two points \(x, y \in E\), and satisfying the following conditions for all \(x, y, z \in E\):

\[
\begin{align*}
(D1) \quad & d(x, y) = d(y, x). \quad \text{(symmetry)} \\
(D2) \quad & d(x, y) \geq 0, \text{ and } d(x, y) = 0 \text{ iff } x = y. \quad \text{(positivity)} \\
(D3) \quad & d(x, z) \leq d(x, y) + d(y, z). \quad \text{(triangle inequality)}
\end{align*}
\]

For example, let \(E = \mathbb{R}^n\) (or \(E = \mathbb{C}^n\)). We have the \textit{Euclidean metric}

\[
d_2(x, y) = \left( |x_1 - y_1|^2 + \cdots + |x_n - y_n|^2 \right)^{\frac{1}{2}}.
\]

This particular metric is called the \textit{Euclidean norm}, \(\|x - y\|_2\), where a \textit{norm on} \(E\) is a function \(\|\| : E \to \mathbb{R}_+\), assigning a nonnegative real number \(\|u\|\) to any vector \(u \in E\), and satisfying the following conditions for all \(x, y, z \in E\):

\[
\begin{align*}
(N1) \quad & \|x\| \geq 0, \text{ and } \|x\| = 0 \text{ iff } x = 0. \quad \text{(positivity)} \\
(N2) \quad & \|Ax\| = |A| \|x\|. \quad \text{(scaling)} \\
(N3) \quad & \|x + y\| \leq \|x\| + \|y\|. \quad \text{(triangle inequality)}
\end{align*}
\]

Given a metric space \(E\) with metric \(d\), for every \(a \in E\), for every \(\rho \in \mathbb{R}\), with \(\rho > 0\), the set

\[
B(a, \rho) = \left\{ x \in E \mid d(a, x) \leq \rho \right\}
\]

is called the \textit{closed ball of center} \(a\) \textit{and radius} \(\rho\), the set

\[
B_0(a, \rho) = \left\{ x \in E \mid d(a, x) < \rho \right\}
\]

is called the \textit{open ball of center} \(a\) \textit{and radius} \(\rho\), and the set

\[
S(a, \rho) = \left\{ x \in E \mid d(a, x) = \rho \right\}
\]

is called the \textit{sphere of center} \(a\) \textit{and radius} \(\rho\). It should be noted that \(\rho\) is finite (i.e., not \(+\infty\)). Clearly, \(B(a, \rho) = B_0(a, \rho) \cup S(a, \rho)\). Furthermore, any metric space \(E\) is a topological space with \(\mathcal{O}\) being the family of arbitrary unions of open balls.
One should be careful that, in general, the family of open sets is not closed under infinite intersections. For example, in \( \mathbb{R} \) under the metric \( |x - y| \), letting \( U_n = (-1/n, +1/n) \), each \( U_n \) is open, but \( \bigcap_n U_n = \{0\} \), which is not open.

A topological space \((E, \mathcal{O})\) is said to satisfy the \textit{Hausdorff separation axiom} (or \(T_2\)-separation axiom) if for any two distinct points \( a \neq b \) in \( E \), there exist two open sets \( U_a \) and \( U_b \) such that, \( a \in U_a, b \in U_b \), and \( U_a \cap U_b = \emptyset \). When the \(T_2\)-separation axiom is satisfied, we also say that \((E, \mathcal{O})\) is a \textit{Hausdorff space}.

Remark: Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough “small” open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed).

By taking complements, we can state properties of the closed sets dual to those of Definition 3.1. Thus, \( \emptyset \) and \( E \) are closed sets, and the closed sets are closed under finite unions and arbitrary intersections.

It is also worth noting that the Hausdorff separation axiom implies that for every \( a \in E \), the set \( \{a\} \) is closed. Indeed, if \( x \in E - \{a\} \), then \( x \neq a \), and so there exist open sets \( U_a \) and \( U_x \) such that \( a \in U_a, x \in U_x \), and \( U_a \cap U_x = \emptyset \). Thus, for every \( x \in E - \{a\} \), there is an open set \( U_x \) containing \( x \) and contained in \( E - \{a\} \), showing by (O3) that \( E - \{a\} \) is open, and thus that the set \( \{a\} \) is closed.

Given a topological space, \((E, \mathcal{O})\), given any subset \( A \) of \( E \), since \( E \in \mathcal{O} \) and \( E \) is a closed set, the family \( \mathcal{C}_A = \{F \mid A \subseteq F, \text{ \( F \) \ is a closed set}\} \) of closed sets containing \( A \) is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection \( \bigcap \mathcal{C}_A \) of the sets in the family \( \mathcal{C}_A \) is the smallest closed set containing \( A \). By a similar reasoning, the union of all the open subsets contained in \( A \) is the largest open set contained in \( A \).

**Definition 3.2.** Given a topological space \((E, \mathcal{O})\), for any subset \( A \) of \( E \), the smallest closed set containing \( A \) is denoted by \( \overline{A} \), and is called the \textit{closure} or \textit{adherence} of \( A \). A subset \( A \) of \( E \) is \textit{dense in} \( E \) if \( \overline{A} = E \). The largest open set contained in \( A \) is denoted by \( \overset{\circ}{A} \), and is called the \textit{interior} of \( A \). The set \( \text{Fr} A = \overline{A} \cap \overline{E - A} \) is called the \textit{boundary} (or \textit{frontier}) of \( A \). We also denote the boundary of \( A \) by \( \partial A \).

Remark: The notation \( \overline{A} \) for the closure of a subset \( A \) of \( E \) is somewhat unfortunate, since \( \overline{A} \) is often used to denote the set complement of \( A \) in \( E \). Still, we prefer it to more cumbersome notations such as \( \text{clo}(A) \), and we denote the complement of \( A \) in \( E \) by \( E - A \) (or sometimes, \( A^c \)).
By definition, it is clear that a subset $A$ of $E$ is closed iff $A = \overline{A}$. The set $\mathbb{Q}$ of rationals is dense in $\mathbb{R}$. It is easily shown that $\overline{A} = \hat{A} \cup \partial A$ and $\hat{A} \cap \partial A = \emptyset$. Another useful characterization of $\overline{A}$ is given by the following proposition. Since this a review chapter, we will not provide proofs of the theorems and propositions and instead refer the reader to Massey [122, 123], Armstrong [5], and Munkres [134].

**Proposition 3.1.** Given a topological space $(E, \mathcal{O})$, given any subset $A$ of $E$, the closure $\overline{A}$ of $A$ is the set of all points $x \in E$ such that for every open set $U$ containing $x$, $U \cap A \neq \emptyset$.

Often it is necessary to consider a subset $A$ of a topological space $E$, and to view the subset $A$ as a topological space. The following definition shows how to define a topology on a subset.

**Definition 3.3.** Given a topological space $(E, \mathcal{O})$, given any subset $A$ of $E$, the **subspace topology on $A$ induced by $\mathcal{O}$** is the family $\mathcal{U}$ of open sets defined such that

$$\mathcal{U} = \{ U \cap A \mid U \in \mathcal{O} \}$$

is the family of all subsets of $A$ obtained as the intersection of any open set in $\mathcal{O}$ with $A$. We say that $(A, \mathcal{U})$ has the **subspace topology**. If $(E, d)$ is a metric space, the restriction $d_A : A \times A \to \mathbb{R}^+$ of the metric $d$ to $A$ is called the **subspace metric**.

For example, if $E = \mathbb{R}^n$ and $d$ is the Euclidean metric, we obtain the subspace topology on the closed $n$-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i, 1 \leq i \leq n\}.$$

One should realize that every open set $U \in \mathcal{O}$ which is entirely contained in $A$ is also in the family $\mathcal{U}$, but $\mathcal{U}$ may contain open sets that are not in $\mathcal{O}$. For example, if $E = \mathbb{R}$ with $|x - y|$, and $A = [a, b]$, then sets of the form $[a, c)$, with $a < c < b$ belong to $\mathcal{U}$, but they are not open sets for $\mathbb{R}$ under $|x - y|$. However, there is agreement in the following situation.

**Proposition 3.2.** Given a topological space $(E, \mathcal{O})$, given any subset $A$ of $E$, if $\mathcal{U}$ is the subspace topology, then the following properties hold.

1. If $A$ is an open set $A \in \mathcal{O}$, then every open set $U \in \mathcal{U}$ is an open set $U \in \mathcal{O}$.
2. If $A$ is a closed set in $E$, then every closed set w.r.t. the subspace topology is a closed set w.r.t. $\mathcal{O}$.

The concept of product topology is also useful.

**Definition 3.4.** Given $n$ topological spaces $(E_i, \mathcal{O}_i)$, the **product topology on $E_1 \times \cdots \times E_n$** is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if

$$\mathcal{B} = \{ U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i, 1 \leq i \leq n \},$$

then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. The set, $E_1 \times \cdots \times E_n$, when given the product topology, is called the **product space**.
It can be verified that when \( E_i = \mathbb{R} \), with the standard topology induced by \(|x - y|\), the product topology on \( \mathbb{R}^n \) is the standard topology induced by the Euclidean norm. This equality between the two topologies suggests the following definition.

**Definition 3.5.** Two metrics \( d_1 \) and \( d_2 \) on a space \( E \) are *equivalent* if they induce the same topology \( \mathcal{O} \) on \( E \) (i.e., they define the same family \( \mathcal{O} \) of open sets). Similarly, two norms \( \| \|_1 \) and \( \| \|_2 \) on a space \( E \) are *equivalent* if they induce the same topology \( \mathcal{O} \) on \( E \).

Given a topological space \((E, \mathcal{O})\), it is often useful, as in Definition 3.4, to define the topology \( \mathcal{O} \) in terms of a subfamily \( B \) of subsets of \( E \).

**Definition 3.6.** We say that a family \( B \) of subsets of \( E \) is a *basis for the topology \( \mathcal{O} \)*, if \( B \) is a subset of \( \mathcal{O} \), and if every open set \( U \) in \( \mathcal{O} \) can be obtained as some union (possibly infinite) of sets in \( B \) (agreeing that the empty union is the empty set). A *subbasis for \( \mathcal{O} \)* is a family \( S \) of subsets of \( E \), such that the family \( B \) of all finite intersections of sets in \( S \) (including \( E \) itself, in case of the empty intersection) is a basis of \( \mathcal{O} \).

It is immediately verified that if a family \( B = (U_i)_{i \in I} \) is a basis for the topology of \((E, \mathcal{O})\), then \( E = \bigcup_{i \in I} U_i \), and the intersection of any two sets \( U_i, U_j \in B \) is the union of some sets in the family \( B \) (again, agreeing that the empty union is the empty set). Conversely, a family \( B \) with these properties is the basis of the topology obtained by forming arbitrary unions of sets in \( B \).

The following proposition gives useful criteria for determining whether a family of open subsets is a basis of a topological space.

**Proposition 3.3.** Given a topological space \((E, \mathcal{O})\) and a family \( B \) of open subsets in \( \mathcal{O} \) the following properties hold:

1. The family \( B \) is a basis for the topology \( \mathcal{O} \) iff for every open set \( U \in \mathcal{O} \) and every \( x \in U \), there is some \( B \in B \) such that \( x \in B \) and \( B \subseteq U \).

2. The family \( B \) is a basis for the topology \( \mathcal{O} \) iff

   (a) For every \( x \in E \), there is some \( B \in B \) such that \( x \in B \).

   (b) For any two open subsets, \( B_1, B_2 \in B \), for every \( x \in E \), if \( x \in B_1 \cap B_2 \), then there is some \( B_3 \in B \) such that \( x \in B_3 \) and \( B_3 \subseteq B_1 \cap B_2 \).

We now consider the fundamental property of continuity.
3.2 Continuous Functions, Limits

**Definition 3.7.** Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, and let \(f: E \to F\) be a function. For every \(a \in E\), we say that \(f\) is continuous at \(a\), if for every open set \(V \in \mathcal{O}_F\) containing \(f(a)\), there is some open set \(U \in \mathcal{O}_E\) containing \(a\), such that, \(f(U) \subseteq V\). We say that \(f\) is continuous if it is continuous at every \(a \in E\).

![Figure 3.1: A schematic illustration of Definition 3.7](image)

Define a neighborhood of \(a \in E\) as any subset \(N\) of \(E\) containing some open set \(O \in \mathcal{O}\) such that \(a \in O\). It is easy to see that Definition 3.7 is equivalent to the following statements.

**Proposition 3.4.** Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, and let \(f: E \to F\) be a function. For every \(a \in E\), the function \(f\) is continuous at \(a \in E\) iff for every neighborhood \(N\) of \(f(a) \in F\), then \(f^{-1}(N)\) is a neighborhood of \(a\). The function \(f\) is continuous on \(E\) iff \(f^{-1}(V)\) is an open set in \(\mathcal{O}_E\) for every open set \(V \in \mathcal{O}_F\).

If \(E\) and \(F\) are metric spaces defined by metrics \(d_1\) and \(d_2\), we can show easily that \(f\) is continuous at \(a\) iff

for every \(\epsilon > 0\), there is some \(\eta > 0\), such that, for every \(x \in E\),

\[
\text{if } d_1(a, x) \leq \eta, \text{ then } d_2(f(a), f(x)) \leq \epsilon.
\]

Similarly, if \(E\) and \(F\) are normed vector spaces defined by norms \(\| \cdot \|_1\) and \(\| \cdot \|_2\), we can show easily that \(f\) is continuous at \(a\) iff

for every \(\epsilon > 0\), there is some \(\eta > 0\), such that, for every \(x \in E\),

\[
\text{if } \|x - a\|_1 \leq \eta, \text{ then } \|f(x) - f(a)\|_2 \leq \epsilon.
\]
It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

If \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) are topological spaces, and \(f: E \to F\) is a function, for every nonempty subset \(A \subseteq E\) of \(E\), we say that \(f\) is continuous on \(A\) if the restriction of \(f\) to \(A\) is continuous with respect to \((A, \mathcal{U})\) and \((F, \mathcal{O}_F)\), where \(\mathcal{U}\) is the subspace topology induced by \(\mathcal{O}_E\) on \(A\).

Given a product \(E_1 \times \cdots \times E_n\) of topological spaces, as usual, we let \(\pi_i: E_1 \times \cdots \times E_n \to E_i\) be the projection function such that, \(\pi_i(x_1, \ldots, x_n) = x_i\). It is immediately verified that each \(\pi_i\) is continuous. In fact, it can be shown that the product topology is the smallest topology on \(E_1 \times \cdots \times E_n\) for which each \(\pi_i\) is continuous.

Given a topological space \((E, \mathcal{O})\), we say that a point \(a \in E\) is isolated if \(\{a\}\) is an open set in \(\mathcal{O}\). Then, if \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) are topological spaces, any function \(f: E \to F\) is continuous at every isolated point \(a \in E\). In the discrete topology, every point is isolated.

The following proposition is easily shown.

**Proposition 3.5.** Given topological spaces \((E, \mathcal{O}_E)\), \((F, \mathcal{O}_F)\), and \((G, \mathcal{O}_G)\), and two functions \(f: E \to F\) and \(g: F \to G\), if \(f\) is continuous at \(a \in E\) and \(g\) is continuous at \(f(a) \in F\), then \(g \circ f: E \to G\) is continuous at \(a \in E\). Given \(n\) topological spaces \((F_i, \mathcal{O}_i)\), for every function \(f: E \to F_1 \times \cdots \times F_n\), then \(f\) is continuous at \(a \in E\) iff every \(f_i: E \to F_i\) is continuous at \(a\), where \(f_i = \pi_i \circ f\).

One can also show that in a metric space \((E, d)\), the norm \(d: E \times E \to \mathbb{R}\) is continuous, where \(E \times E\) has the product topology, and that for a normed vector space \((E, \|\|)\), the norm \(\|\|: E \to \mathbb{R}\) is continuous.

Given a function \(f: E_1 \times \cdots \times E_n \to F\), we can fix \(n - 1\) of the arguments, say \(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\), and view \(f\) as a function of the remaining argument,

\[
x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),
\]

where \(x_i \in E_i\). If \(f\) is continuous, it is clear that each \(f_i\) is continuous.

One should be careful that the converse is false! For example, consider the function \(f: \mathbb{R} \times \mathbb{R} \to \mathbb{R}\), defined such that,

\[
f(x, y) = \frac{xy}{x^2 + y^2} \quad \text{if} \quad (x, y) \neq (0, 0), \quad \text{and} \quad f(0, 0) = 0.
\]

The function \(f\) is continuous on \(\mathbb{R} \times \mathbb{R} - \{(0, 0)\}\), but on the line \(y = mx\), with \(m \neq 0\), we have \(f(x, y) = \frac{m}{1+m^2} \neq 0\), and thus, on this line, \(f(x, y)\) does not approach 0 when \((x, y)\) approaches \((0, 0)\).

The following proposition is useful for showing that real-valued functions are continuous.
**Proposition 3.6.** If $E$ is a topological space, and $(\mathbb{R}, |x - y|)$ the reals under the standard topology, for any two functions $f: E \to \mathbb{R}$ and $g: E \to \mathbb{R}$, for any $a \in E$, for any $\lambda \in \mathbb{R}$, if $f$ and $g$ are continuous at $a$, then $f + g$, $\lambda f$, $f \cdot g$, are continuous at $a$, and $f/g$ is continuous at $a$ if $g(a) \neq 0$.

**Remark:** Proposition 3.6 is true if $\mathbb{R}$ is replaced with $\mathbb{C}$, where the $\mathbb{C}$ has the topology induced by the Euclidean norm on $\mathbb{R}^2$.

Using Proposition 3.6, we can show easily that every real or complex polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

**Definition 3.8.** Let $(E, O_E)$ and $(F, O_F)$ be topological spaces, and let $f: E \to F$ be a function. We say that $f$ is a homeomorphism between $E$ and $F$ if $f$ is bijective, and both $f: E \to F$ and $f^{-1}: F \to E$ are continuous.

One should be careful that a bijective continuous function $f: E \to F$ is not necessarily an homeomorphism. For example, if $E = \mathbb{R}$ with the discrete topology, and $F = \mathbb{R}$ with the standard topology, the identity is not a homeomorphism.

We now introduce the concept of limit of a sequence. Given any set $E$, a sequence is any function $x: \mathbb{N} \to E$, usually denoted by $(x_n)_{n \in \mathbb{N}}$, or $(x_n)_{n \geq 0}$, or even by $(x_n)$.

**Definition 3.9.** Given a topological space, $(E, O)$, we say that a sequence $(x_n)_{n \in \mathbb{N}}$ converges to some $a \in E$ if for every open set $U$ containing $a$, there is some $n_0 \geq 0$, such that, $x_n \in U$, for all $n \geq n_0$. We also say that $a$ is a limit of $(x_n)_{n \in \mathbb{N}}$.

When $E$ is a metric space with metric $d$, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $d(x_n, a) \leq \epsilon$, for all $n \geq n_0$.

When $E$ is a normed vector space with norm $\| \cdot \|$, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $\|x_n - a\| \leq \epsilon$, for all $n \geq n_0$.

The following proposition shows the importance of the Hausdorff separation axiom.

**Proposition 3.7.** Given a topological space $(E, O)$, if the Hausdorff separation axiom holds, then every sequence has at most one limit.

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit $b$ iff it converges to the same limit $b$ in any equivalent metric (and similarly for equivalent norms).

We still need one more concept of limit for functions.
Definition 3.10. Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be topological spaces, let \(A\) be some nonempty subset of \(E\), and let \(f : A \to F\) be a function. For any \(a \in A\) and any \(b \in F\), we say that \(f(x)\) approaches \(b\) as \(x\) approaches \(a\) with values in \(A\) if for every open set \(V \in \mathcal{O}_F\) containing \(b\), there is some open set \(U \in \mathcal{O}_E\) containing \(a\), such that, \(f(U \cap A) \subseteq V\). This is denoted by

\[
\lim_{x \to a, x \in A} f(x) = b.
\]

First, note that by Proposition 3.1, since \(a \in A\), for every open set \(U\) containing \(a\), we have \(U \cap A \neq \emptyset\), and the definition is nontrivial. Also, even if \(a \in A\), the value \(f(a)\) of \(f\) at \(a\) plays no role in this definition. When \(E\) and \(F\) are metric space with metrics \(d_1\) and \(d_2\), it can be shown easily that the definition can be stated as follows:

For every \(\epsilon > 0\), there is some \(\eta > 0\), such that, for every \(x \in A\),

if \(d_1(x, a) \leq \eta\), then \(d_2(f(x), b) \leq \epsilon\).

When \(E\) and \(F\) are normed vector spaces with norms \(\|\cdot\|_1\) and \(\|\cdot\|_2\), it can be shown easily that the definition can be stated as follows:

For every \(\epsilon > 0\), there is some \(\eta > 0\), such that, for every \(x \in A\),

if \(\|x - a\|_1 \leq \eta\), then \(\|f(x) - b\|_2 \leq \epsilon\).

We have the following result relating continuity at a point and the previous notion.

Proposition 3.8. Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be two topological spaces, and let \(f : E \to F\) be a function. For any \(a \in E\), the function \(f\) is continuous at \(a\) iff \(f(x)\) approaches \(f(a)\) when \(x\) approaches \(a\) (with values in \(E\)).

Another important proposition relating the notion of convergence of a sequence to continuity is stated without proof.

Proposition 3.9. Let \((E, \mathcal{O}_E)\) and \((F, \mathcal{O}_F)\) be two topological spaces, and let \(f : E \to F\) be a function.

1. If \(f\) is continuous, then for every sequence \((x_n)_{n \in \mathbb{N}}\) in \(E\), if \((x_n)\) converges to \(a\), then \((f(x_n))\) converges to \(f(a)\).

2. If \(E\) is a metric space, and \((f(x_n))\) converges to \(f(a)\) whenever \((x_n)\) converges to \(a\), for every sequence \((x_n)_{n \in \mathbb{N}}\) in \(E\), then \(f\) is continuous.

We now turn to connectivity properties of topological spaces.
3.3 Connected Sets

Connectivity properties of topological spaces play a very important role in understanding the topology of surfaces.

**Definition 3.11.** A topological space, $(E, \mathcal{O})$, is *connected* if the only subsets of $E$ that are both open and closed are the empty set and $E$ itself. Equivalently, $(E, \mathcal{O})$ is connected if $E$ cannot be written as the union $E = U \cup V$ of two disjoint nonempty open sets, $U, V$, or if $E$ cannot be written as the union $E = U \cup V$ of two disjoint nonempty closed sets. A topological space, $(E, \mathcal{O})$, is *disconnected* if it is not connected.

**Definition 3.12.** A subset, $S \subseteq E$, is *connected* if it is connected in the subspace topology on $S$ induced by $(E, \mathcal{O})$. Otherwise the subset $S$ is *disconnected* which means there exits open subsets $G$ and $H$ of $X$ such that $S$ is the disjoint union of the two nonempty subsets $S \cap H$ and $S \cap G$. See Figure 3.2. A connected open set is called a *region* and a closed set is a *closed region* if its interior is a connected (open) set.

![Figure 3.2: The graph of $z^2 - x^2 - y^2 = 1$ is disconnected in $\mathbb{R}^3$. Let $G = \{(x, y, z)|z > 0\}$ and $H = \{(x, y, z)|z < 0\}$.

Most readers have an intuitive notion of the meaning of connectivity, namely that the space $E$ is in “one piece.” In particular, the following standard proposition characterizing the connected subsets of $\mathbb{R}$ can be found in most topology texts (for example, Munkres [134], Schwartz [155]).

**Proposition 3.10.** A subset of the real line, $\mathbb{R}$, is connected iff it is an interval, i.e., of the form $[a, b]$, $(a, b]$, where $a = -\infty$ is possible, $[a, b)$, where $b = +\infty$ is possible, or $(a, b)$, where $a = -\infty$ or $b = +\infty$ is possible.
A characterization of the connected subsets of $\mathbb{R}^n$ is harder and requires the notion of arcwise connectedness which we discuss at the end of this section.

One of the most important properties of connected sets is that they are preserved by continuous maps.

**Proposition 3.11.** Given any continuous map, $f : E \to F$, if $A \subseteq E$ is connected, then $f(A)$ is connected.

An important corollary of Proposition 3.11 is that for every continuous function, $f : E \to \mathbb{R}$, where $E$ is a connected space, $f(E)$ is an interval. Indeed, this follows from Proposition 3.10. Thus, if $f$ takes the values $a$ and $b$ where $a < b$, then $f$ takes all values $c \in [a, b]$. This is property is the Intermediate Value Theorem.

Here are two more properties of connected subsets.

**Lemma 3.12.** Given a topological space, $E$, for any family, $(A_i)_{i \in I}$, of (nonempty) connected subsets of $E$, if $A_i \cap A_j \neq \emptyset$ for all $i, j \in I$, then the union $A = \bigcup_{i \in I} A_i$ of the family $(A_i)_{i \in I}$ is also connected.

In particular, the above lemma applies when the connected sets in a family $(A_i)_{i \in I}$ have a point in common.

**Lemma 3.13.** If $A$ is a connected subset of a topological space, $E$, then for every subset, $B$, such that $A \subseteq B \subseteq \overline{A}$, where $\overline{A}$ is the closure of $A$ in $E$, the set $B$ is connected.

In particular, Lemma 3.13 shows that if $A$ is a connected subset, then its closure, $\overline{A}$, is also connected.

Connectivity provides a equivalence relation among the points of $E$.

**Definition 3.13.** Given a topological space, $(E, \mathcal{O})$, we say that two points $a, b \in E$ are connected if there is some connected subset $A$ of $E$ such that $a \in A$ and $b \in A$.

An application of Lemma 3.12 verifies that “$a$ and $b$ are connected in $E$” is an equivalence relation. Thus, the above equivalence relation defines a partition of $E$ into nonempty disjoint connected components. The following proposition, proven via Lemmas 3.12 and 3.13, provides a way of constructing the connected components of $E$.

**Proposition 3.14.** Given any topological space, $E$, for any $a \in E$, the connected component containing $a$ is the largest connected set containing $a$. The connected components of $E$ are closed.

The connected components are the “pieces” of $E$. Intuitively, if a space is not connected, it is possible to define a continuous function which is constant on disjoint connected components and which takes possibly distinct values on disjoint components. This can be stated in terms of the concept of a locally constant function.
CHAPTER 3. A REVIEW OF POINT SET TOPOLOGY

Definition 3.14. Given two topological spaces, $X,Y$, a function, $f: X \to Y$, is locally constant if for every $x \in X$, there is an open set, $U \subseteq X$, such that $x \in X$ and $f$ is constant on $U$.

We claim that a locally constant function is continuous. In fact, we will prove that $f^{-1}(V)$ is open for every subset, $V \subseteq Y$ (not just for an open set $V$). It is enough to show that $f^{-1}(y)$ is open for every $y \in Y$, since for every subset $V \subseteq Y$,

$$f^{-1}(V) = \bigcup_{y \in V} f^{-1}(y),$$

and open sets are closed under arbitrary unions. However, either $f^{-1}(y) = \emptyset$ if $y \in Y - f(X)$ or $f$ is constant on $U = f^{-1}(y)$ if $y \in f(X)$ (with value $y$), and since $f$ is locally constant, for every $x \in U$, there is some open set, $W \subseteq X$, such that $x \in W$ and $f$ is constant on $W$, which implies that $f(w) = y$ for all $w \in W$ and thus, that $W \subseteq U$, showing that $U$ is a union of open sets and thus, is open. The following proposition shows that a space is connected iff every locally constant function is constant.

Proposition 3.15. A topological space is connected iff every locally constant function is constant.

The notion of a locally connected space is also useful.

Definition 3.15. A topological space, $(E, \mathcal{O})$, is locally connected if for every $a \in E$, for every neighborhood $V$ of $a$, there is a connected neighborhood $U$ of $a$ such that $U \subseteq V$.

As we shall see in a moment, it would be equivalent to require that $E$ has a basis of connected open sets.

There are connected spaces that are not locally connected and there are locally connected spaces that are not connected. The two properties are independent. For example, let $X$ be a set with the discrete topology. Since $\{x\}$ is open for every $x \in X$, the topological space $X$ is locally connected. However, if $|X| > 1$, then $X$, with the discrete topology, is not connected. On the other hand, the space consisting of the graph of the function

$$f(x) = \sin(1/x),$$

where $x > 0$, together with the portion of the $y$-axis, for which $-1 \leq y \leq 1$, is connected, but not locally connected. The open disk centered at $(0,1)$ with radius $\frac{1}{4}$ does not contain a connected neighborhood of $(0,1)$. See Figure 3.3.
Figure 3.3: Let $S$ be the graph of $f(x) = \sin(1/x)$ union the $y$-axis between $-1$ and $1$. This space is connected, but not locally connected.

**Proposition 3.16.** A topological space, $E$, is locally connected iff for every open subset $A$ of $E$, the connected components of $A$ are open.

Proposition 3.16 shows that in a locally connected space, the connected open sets form a basis for the topology. It is easily seen that $\mathbb{R}^n$ is locally connected. Manifolds are also locally connected.

Another very important property of surfaces and more generally, manifolds, is to be arcwise connected. The intuition is that any two points can be joined by a continuous arc of curve. This is formalized as follows.

**Definition 3.16.** Given a topological space, $(E, \mathcal{O})$, an *arc* (or path) is a continuous map $\gamma: [a, b] \to E$, where $[a, b]$ is a closed interval of the real line, $\mathbb{R}$. The point $\gamma(a)$ is the *initial point* of the arc and the point $\gamma(b)$ is the *terminal point* of the arc. We say that $\gamma$ is an arc joining $\gamma(a)$ and $\gamma(b)$. An arc is a *closed curve* if $\gamma(a) = \gamma(b)$. The set $\gamma([a, b])$ is the *trace* of the arc $\gamma$.

Typically, $a = 0$ and $b = 1$.

One should not confuse an arc $\gamma: [a, b] \to E$ with its trace. For example, $\gamma$ could be constant, and thus, its trace reduced to a single point.
An arc is a Jordan arc if $\gamma$ is a homeomorphism onto its trace. An arc $\gamma: [a,b] \rightarrow E$ is a Jordan curve if $\gamma(a) = \gamma(b)$ and $\gamma$ is injective on $[a,b]$. Since $[a,b]$ is connected, by Proposition 3.11, the trace $\gamma([a,b])$ of an arc is a connected subset of $E$.

Given two arcs $\gamma: [0,1] \rightarrow E$ and $\delta: [0,1] \rightarrow E$ such that $\gamma(1) = \delta(0)$, we can form a new arc defined as follows:

**Definition 3.17.** Given two arcs, $\gamma: [0,1] \rightarrow E$ and $\delta: [0,1] \rightarrow E$, such that $\gamma(1) = \delta(0)$, we can form their composition (or product), $\gamma \delta$, defined such that

$$
\gamma \delta(t) = \begin{cases} 
\gamma(2t) & \text{if } 0 \leq t \leq 1/2; \\
\delta(2t-1) & \text{if } 1/2 \leq t \leq 1.
\end{cases}
$$

The inverse $\gamma^{-1}$ of the arc $\gamma$ is the arc defined such that $\gamma^{-1}(t) = \gamma(1-t)$, for all $t \in [0,1]$.

It is trivially verified that Definition 3.17 yields continuous arcs.

**Definition 3.18.** A topological space, $E$, is arcwise connected if for any two points $a, b \in E$, there is an arc $\gamma: [0,1] \rightarrow E$ joining $a$ and $b$, such that $\gamma(0) = a$ and $\gamma(1) = b$. A topological space, $E$, is locally arcwise connected if for every $a \in E$, for every neighborhood $V$ of $a$, there is an arcwise connected neighborhood $U$ of $a$ such that $U \subseteq V$.

The space $\mathbb{R}^n$ is locally arcwise connected, since for any open ball, any two points in this ball are joined by a line segment. Manifolds and surfaces are also locally arcwise connected. Proposition 3.11 also applies to arcwise connectedness. The following theorem is crucial to the theory of manifolds and surfaces.

**Theorem 3.17.** If a topological space, $E$, is arcwise connected, then it is connected. If a topological space, $E$, is connected and locally arcwise connected, then $E$ is arcwise connected.

If $E$ is locally arcwise connected, the above argument shows that the connected components of $E$ are arcwise connected.

It is not true that a connected space is arcwise connected. For example, the space consisting of the graph of the function

$$f(x) = \sin(1/x),$$

where $x > 0$, together with the portion of the $y$-axis, for which $-1 \leq y \leq 1$, is connected, but not arcwise connected. See Figure 3.3.

A trivial modification of the proof of Theorem 3.17 shows that in a normed vector space, $E$, a connected open set is arcwise connected by polygonal lines (arcs consisting of line segments). This is because in every open ball, any two points are connected by a line segment. Furthermore, if $E$ is finite dimensional, these polygonal lines can be forced to be parallel to basis vectors.

We conclude this section with the following theorem regarding the connectivity of product spaces.
3.4. COMPACT SETS

**Theorem 3.18.** Let \( X \) and \( Y \) be topological spaces. The product space \( X \times Y \) is connected if and only if \( X \) and \( Y \) are connected.

**Remark:** Theorem 3.18 can be extended to the set \( \{X_i\}_{i=1}^n \), where \( n \) is a positive integer, \( n \geq 2 \).

We now consider compactness.

### 3.4 Compact Sets

The property of compactness is very important in topology and analysis. We provide a quick review geared towards the study of manifolds and for details, we refer the reader to Munkres [134], Schwartz [155]. In this section, we will need to assume that the topological spaces are Hausdorff spaces. This is not a luxury, as many of the results are false otherwise.

There are various equivalent ways of defining compactness. For our purposes, the most convenient way involves the notion of open cover.

**Definition 3.19.** Given a topological space, \( E \), for any subset \( A \) of \( E \), an open cover, \((U_i)_{i \in I}\) of \( A \), is a family of open subsets of \( E \) such that \( A \subseteq \bigcup_{i \in I} U_i \). An open subcover of an open cover, \((U_i)_{i \in I}\) of \( A \), is any subfamily, \((U_j)_{j \in J}\), which is an open cover of \( A \), with \( J \subseteq I \). An open cover, \((U_i)_{i \in I}\) of \( A \), is finite if \( I \) is finite.

**Definition 3.20.** The topological space, \( E \), is compact if it is Hausdorff and for every open cover, \((U_i)_{i \in I}\) of \( E \), there is a finite open subcover \((U_j)_{j \in J}\) of \( E \). Given any subset \( A \) of \( E \), we say that \( A \) is compact if it is compact with respect to the subspace topology. We say that \( A \) is relatively compact if its closure \( \overline{A} \) is compact.

It is immediately verified that a subset, \( A \), of \( E \) is compact in the subspace topology relative to \( A \) iff for every open cover, \((U_i)_{i \in I}\) of \( A \) by open subsets of \( E \), there is a finite open subcover \((U_j)_{j \in J}\) of \( A \). The property that every open cover contains a finite open subcover is often called the Heine-Borel-Lebesgue property. By considering complements, a Hausdorff space is compact iff for every family, \((F_i)_{i \in I}\) of closed sets, if \( \bigcap_{i \in I} F_i = \emptyset \), then \( \bigcap_{j \in J} F_j = \emptyset \) for some finite subset \( J \) of \( I \).

**Remark:** Definition 3.20 requires that a compact space be Hausdorff. There are books in which a compact space is not necessarily required to be Hausdorff. Following Schwartz, we prefer calling such a space quasi-compact.

Another equivalent and useful characterization can be given in terms of families having the finite intersection property. A family \((F_i)_{i \in I}\) of sets has the finite intersection property if \( \bigcap_{j \in J} F_j \neq \emptyset \) for every finite subset \( J \) of \( I \). We have the following proposition.
Chapter 3. A Review of Point Set Topology

**Proposition 3.19.** A topological Hausdorff space, \( E \), is compact iff for every family \( (F_i)_{i \in I} \) of closed sets having the finite intersection property, then \( \bigcap_{i \in I} F_i \neq \emptyset \).

Another useful consequence of compactness is as follows. For any family \( (F_i)_{i \in I} \) of closed sets such that \( F_{i+1} \subseteq F_i \) for all \( i \in I \), if \( \bigcap_{i \in I} F_i = \emptyset \), then \( F_i = \emptyset \) for some \( i \in I \). Indeed, there must be some finite subset \( J \) of \( I \) such that \( \bigcap_{j \in J} F_j \neq \emptyset \), and since \( F_{i+1} \subseteq F_i \) for all \( i \in I \), we must have \( F_j = \emptyset \) for the smallest \( F_j \) in \( (F_j)_{j \in J} \). Using this fact, we note that \( \mathbb{R} \) is not compact. Indeed, the family of closed sets, \( ([n, +\infty))_{n \geq 0} \), is decreasing and has an empty intersection.

Given a metric space, if we define a **bounded subset** to be a subset that can be enclosed in some closed ball (of finite radius), then any nonbounded subset of a metric space is not compact. However, a closed interval \([a, b]\) of the real line is compact, and by extension every closed set, \([a_1, b_1] \times \cdots \times [a_m, b_m]\), when considered as a subspace of \( \mathbb{R}^m \), is compact.

The following two propositions give very important properties of the compact sets, and they only hold for Hausdorff spaces.

**Proposition 3.20.** Given a topological Hausdorff space, \( E \), for every compact subset, \( A \), and every point \( b \) not in \( A \), there exist disjoint open sets, \( U \) and \( V \), such that \( A \subseteq U \) and \( b \in V \). As a consequence, every compact subset is closed.

**Proposition 3.21.** Given a topological Hausdorff space, \( E \), for every pair of compact disjoint subsets, \( A \) and \( B \), there exist disjoint open sets, \( U \) and \( V \), such that \( A \subseteq U \) and \( B \subseteq V \).

The following proposition shows that in a compact topological space, every closed set is compact.

**Proposition 3.22.** Given a compact topological space, \( E \), every closed set is compact.

**Remark:** Proposition 3.22 also holds for quasi-compact spaces, i.e., the Hausdorff separation property is not needed.

Putting Proposition 3.21 and Proposition 3.22 together, we note that if \( X \) is compact, then for every pair of disjoint closed, sets \( A \) and \( B \), there exist disjoint open sets, \( U \) and \( V \), such that \( A \subseteq U \) and \( B \subseteq V \). We say that \( X \) is a **normal** space.

**Proposition 3.23.** Given a compact topological space, \( E \), for every \( a \in E \), and for every neighborhood \( V \) of \( a \), there exists a compact neighborhood \( U \) of \( a \) such that \( U \subseteq V \).

It can be shown that in a normed vector space of finite dimension, a subset is compact iff it is closed and bounded. This is what we use to show that \( \text{SO}(n) \) is compact in \( \mathbb{R}^{n^2} \). For \( \mathbb{R}^n \), the proof is simple.

In a normed vector space of infinite dimension, there are closed and bounded sets that are not compact!
Another crucial property of compactness is that it is preserved under continuity.

**Proposition 3.24.** Let $E$ be a topological space and let $F$ be a topological Hausdorff space. For every compact subset, $A$ of $E$, and for every continuous map, $f: E \to F$, the subspace $f(A)$ is compact.

As a corollary of Proposition 3.24, if $E$ is compact, $F$ is Hausdorff, and $f: E \to F$ is continuous and bijective, then $f$ is a homeomorphism. Indeed, it is enough to show that $f^{-1}$ is continuous, which is equivalent to showing that $f$ maps closed sets to closed sets. However, closed sets are compact and Proposition 3.24 shows that compact sets are mapped to compact sets, which, by Proposition 3.20, are closed.

It can also be shown that if $E$ is a compact nonempty space and $f: E \to \mathbb{R}$ is a continuous function, then there are points $a, b \in E$ such that $f(a)$ is the minimum of $f(E)$ and $f(b)$ is the maximum of $f(E)$. Indeed, $f(E)$ is a compact subset of $\mathbb{R}$ and thus, a closed and bounded set which contains its greatest lower bound and its least upper bound.

Another useful notion is that of local compactness. Indeed, manifolds and surfaces are locally compact.

**Definition 3.21.** A topological space, $E$, is **locally compact** if it is Hausdorff and for every $a \in E$, there is some compact neighborhood, $K$, of $a$.

From Proposition 3.23, every compact space is locally compact but the converse is false. For example, the real line $\mathbb{R}$, which is not compact, is locally compact since each $x \in \mathbb{R}$, given any neighborhood $N$ of $x$, there exist $\epsilon > 0$ such that $x \in [x - \epsilon, x + \epsilon] \subseteq N$. Furthermore, it can be shown that a normed vector space of finite dimension is locally compact.

**Proposition 3.25.** Given a locally compact topological space, $E$, for every $a \in E$, and for every neighborhood $N$ of $a$, there exists a compact neighborhood $U$ of $a$ such that $U \subseteq N$.

Finally, in studying surfaces and manifolds, an important property is the existence of a countable basis for the topology.

**Definition 3.22.** A topological space $E$ is called **second-countable** if there is a countable basis for its topology, i.e., if there is a countable family $(U_i)_{i \geq 0}$ of open sets such that every open set of $E$ is a union of open sets $U_i$.

It is easily seen that $\mathbb{R}^n$ is second-countable and more generally, that every normed vector space of finite dimension is second-countable. We have the following properties regarding second-countabilility.

**Proposition 3.26.** Given a second-countable topological space $E$, every open cover $(U_i)_{i \in I}$ of $E$ contains some countable subcover.
As an immediate corollary of Proposition 3.26, a locally connected second-countable space has countably many connected components.

In second-countable Hausdorff spaces, compactness can be characterized in terms of accumulation points (this is also true for metric spaces).

**Definition 3.23.** Given a topological Hausdorff space, $E$, and given any sequence $(x_n)$ of points in $E$, a point, $l \in E$, is an accumulation point (or cluster point) of the sequence $(x_n)$ if every open set, $U$, containing $l$ contains $x_n$ for infinitely many $n$.

Clearly, if $l$ is a limit of the sequence $(x_n)$, then it is an accumulation point, since every open set, $U$, containing $a$ contains all $x_n$ except for finitely many $n$. The following proposition provides another characterization of an accumulation point.

**Proposition 3.27.** Given a second-countable topological Hausdorff space, $E$, a point, $l$, is an accumulation point of the sequence $(x_n)$ iff $l$ is the limit of some subsequence $(x_{n_k})$, of $(x_n)$.

**Remark:** Proposition 3.27 also holds for metric spaces.

The next proposition relates the existence of accumulation points to the notion of compactness.

**Proposition 3.28.** A second-countable topological Hausdorff space, $E$, is compact iff every sequence $(x_n)$ has some accumulation point.

**Remark:** It should be noted that the proof showing that if $E$ is compact, then every sequence has some accumulation point, holds for any arbitrary compact space (the proof does not use a countable basis for the topology). The converse also holds for metric spaces.

Closely related to Proposition 3.28 is the Bolzano-Weierstrass property which states that an infinite subset of a compact space has a limit point.

We end this section with a result about the product of compact spaces. But first we state the following proposition.

**Proposition 3.29.** Let $X$ and $Y$ be topological spaces. The product space $X \times Y$ is a Hausdorff space iff $X$ and $Y$ are Hausdorff spaces.

**Remark:** Proposition 3.29 is true for finite set of topological spaces, $\{X_i\}_{i=1}^n$, with $n \geq 2$.

**Proposition 3.30.** Let $\{X_i\}_{i=1}^n$ be a family of topological spaces. The product space $X_1 \times \cdots \times X_n$ is compact iff $X_i$ is compact for all $1 \leq i \leq n$. 
3.5 Quotient Spaces

In the final section of this chapter we discuss a topological construction, the quotient space, which plays important role in the study of orbifolds and homogenous manifolds. For example, real projective spaces and Grassmannians are obtained this way. In this situation, the natural topology on the quotient object is the quotient topology, but unfortunately, even if the original space is Hausdorff, the quotient topology may not be. Therefore, it is useful to have criteria that insure that a quotient topology is Hausdorff (or second-countable). We will present two criteria. First, let us review the notion of quotient topology. For more details, consult Munkres [134], Massey [122, 123], Armstrong [5], or Tu [170].

Definition 3.24. Given any topological space $X$ and any set $Y$, for any surjective function $f : X \to Y$, we define the quotient topology on $Y$ determined by $f$ (also called the identification topology on $Y$ determined by $f$), by requiring a subset $V$ of $Y$ to be open if $f^{-1}(V)$ is an open set in $X$. Given an equivalence relation $R$ on a topological space $X$, if $\pi : X \to X/R$ is the projection sending every $x \in X$ to its equivalence class $[x]$ in $X/R$, the space $X/R$ equipped with the quotient topology determined by $\pi$ is called the quotient space of $X$ modulo $R$. Thus a set $V$ of equivalence classes in $X/R$ is open iff $\pi^{-1}(V)$ is open in $X$, which is equivalent to the fact that $\bigcup_{[x] \in V}[x]$ is open in $X$.

It is immediately verified that Definition 3.24 defines topologies and that $f : X \to Y$ and $\pi : X \to X/R$ are continuous when $Y$ and $X/R$ are given these quotient topologies.

To intuitively understand the quotient space construction, start with a topological space $X$ and any set $Y$, for any surjective function $f : X \to Y$, we define the quotient topology on $Y$ determined by $f$ (also called the identification topology on $Y$ determined by $f$), by requiring a subset $V$ of $Y$ to be open if $f^{-1}(V)$ is an open set in $X$. Given an equivalence relation $R$ on a topological space $X$, if $\pi : X \to X/R$ is the projection sending every $x \in X$ to its equivalence class $[x]$ in $X/R$, the space $X/R$ equipped with the quotient topology determined by $\pi$ is called the quotient space of $X$ modulo $R$. Thus a set $V$ of equivalence classes in $X/R$ is open iff $\pi^{-1}(V)$ is open in $X$, which is equivalent to the fact that $\bigcup_{[x] \in V}[x]$ is open in $X$.

We demonstrate this construction by building a cylinder as a quotient of the rectangle $Q = [0, 2] \times [0, 1]$. The partition $\mathcal{R} = \bigcup_{i \in I} R_i$ of $Q$ is defined as follows:

i. $R_i = \{(x, y)\}$ where $0 < x < 2$ and $0 \leq y \leq 1$.

ii. $R_i = \{(0, y), (2, y)\}$ where $0 \leq y \leq 1$

Each $R_i$ is a point in $Y$ and the function $\pi : Q \to Y$ maps $(x, y)$ to the $R_i$ which contains it. The map $\pi$ “glues” together the left and right vertical edges of $Q$ and forms a cylinder.
A similar construction creates a Möbius strip as a quotient of \( Q = [0, 2] \times [0, 1] \). This time the partition \( \mathcal{R} = \bigcup_{i \in I} \mathcal{R}_i \) of \( Q \) is

i. \( \mathcal{R}_i = \{(x, y)\} \) where \( 0 < x < 2 \) and \( 0 \leq y \leq 1 \),

ii. \( \mathcal{R}_i = \{(0, y), (2, 1 - y)\} \) where \( 0 \leq y \leq 1 \).

This time the map \( \pi: Q \to Y \) “glues” the left and right vertical edges with a twist and forms a Möbius strip.

We can also build a torus as quotient of the unit square \( S = [0, 1] \times [0, 1] \) by giving \( S \) the following partition \( \mathcal{R} = \bigcup_{i \in I} \mathcal{R}_i \):
3.5. QUOTIENT SPACES

i. \( \mathcal{R}_i = \{(0, 0), (0, 1), (1, 1), (1, 0)\} \).

ii. \( \mathcal{R}_i = \{(0, y), (1, y)\} \) for \( 0 < y < 1 \).

iii. \( \mathcal{R}_i = \{(x, 0), (x, 1)\} \) for \( 0 < x < 1 \).

iv. \( \mathcal{R}_i = \{(x, y)\} \) for \( 0 < x < 1 \) and \( 0 < y < 1 \).

Once again each \( \mathcal{R}_i \) is a point in \( Y \) and the function \( \pi: Q \to Y \) maps \( (x, y) \) to the equivalence class \( \mathcal{R}_i \) containing it. Geometrically \( \pi \) takes \( S \), glues together the left and right edges to form a cylinder, then glues together the top and bottom of the cylinder to form the torus.

Figure 3.6: Constructing a torus as a quotient of a square
Although we visualized the proceeding three quotients spaces in $\mathbb{R}^3$, the quotient construction, namely $\pi: Q \to Y$, is abstract and independent of any pictorial representation.

One should be careful that if $X$ and $Y$ are topological spaces and $f: X \to Y$ is a continuous surjective map, $Y$ does not necessarily have the quotient topology determined by $f$. Indeed, it may not be true that a subset $V$ of $Y$ is open when $f^{-1}(V)$ is open. However, this will be true in two important cases.

**Definition 3.25.** A continuous map $f: X \to Y$ is an open map (or simply open) if $f(U)$ is open in $Y$ whenever $U$ is open in $X$, and similarly, $f: X \to Y$ is a closed map (or simply closed) if $f(F)$ is closed in $Y$ whenever $F$ is closed in $X$.

Then, $Y$ has the quotient topology induced by the continuous surjective map $f$ if either $f$ is open or $f$ is closed. Indeed, if $f$ is open, then assuming that $f^{-1}(V)$ is open in $X$, we have $f(f^{-1}(V)) = V$ open in $Y$. Now, since $f^{-1}(Y - B) = X - f^{-1}(B)$, for any subset $B$ of $Y$, a subset $V$ of $Y$ is open in the quotient topology iff $f^{-1}(Y - V)$ is closed in $X$. From this, we can deduce that if $f$ is a closed map, then $V$ is open in $Y$ iff $f^{-1}(V)$ is open in $X$.

Unfortunately, the Hausdorff separation property is not necessarily preserved under quotient. Nevertheless, it is preserved in some special important cases.

**Proposition 3.31.** Let $X$ and $Y$ be topological spaces, let $f: X \to Y$ be a continuous surjective map, and assume that $X$ is compact and that $Y$ has the quotient topology determined by $f$. Then $Y$ is Hausdorff iff $f$ is a closed map.

**Proof.** If $Y$ is Hausdorff, because $X$ is compact and $f$ is continuous, since every closed set $F$ in $X$ is compact, $f(F)$ is compact, and since $Y$ is Hausdorff, $f(F)$ is closed, and $f$ is a closed map.

For the converse, we use the fact that in a Hausdorff space $E$, if $A$ and $B$ are compact disjoint subsets of $E$, then there exist two disjoint open sets $U$ and $V$ such that $A \subseteq U$ and $B \subseteq V$.

Since $X$ is Hausdorff, every set $\{a\}$ consisting of a single element $a \in X$ is closed, and since $f$ is a closed map, $\{f(a)\}$ is also closed in $Y$. Since $f$ is surjective, every set $\{b\}$ consisting of a single element $b \in Y$ is closed. If $b_1, b_2 \in Y$ and $b_1 \neq b_2$, since $\{b_1\}$ and $\{b_2\}$ are closed in $Y$ and $f$ is continuous, the sets $f^{-1}(b_1)$ and $f^{-1}(b_2)$ are closed in $X$, thus compact, and by the fact stated above, there exists some disjoint open sets $U_1$ and $U_2$ such that $f^{-1}(b_1) \subseteq U_1$ and $f^{-1}(b_2) \subseteq U_2$. Since $f$ is closed, the sets $f(X - U_1)$ and $f(X - U_2)$ are closed, and thus the sets

\[
V_1 = Y - f(X - U_1) \\
V_2 = Y - f(X - U_2)
\]

are open, and it is immediately verified that $V_1 \cap V_2 = \emptyset$, $b_1 \in V_1$, and $b_2 \in V_2$. This proves that $Y$ is Hausdorff. \qed
Under the hypotheses of Proposition 3.31, it is easy to show that $Y$ is Hausdorff iff the set
\[ \{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\} \]
is closed in $X \times X$.

Another simple criterion uses continuous open maps. The following proposition is proved in Massey [122] (Appendix A, Proposition 5.3).

**Proposition 3.32.** Let $f: X \to Y$ be a surjective continuous map between topological spaces. If $f$ is an open map, then $Y$ is Hausdorff iff the set
\[ \{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\} \]
is closed in $X \times X$.

Note that the hypothesis of Proposition 3.32 implies that $Y$ has the quotient topology determined by $f$.

The following special case of Proposition 3.32 is discussed in Tu [170] (Section 7.5, Theorem 7.8). Given a topological space $X$ and an equivalence relation $R$ on $X$, we say that $R$ is *open* if the projection map $\pi: X \to X/R$ is an open map, where $X/R$ is equipped with the quotient topology. Then, if $R$ is an open equivalence relation on $X$, the topological space $X/R$ is Hausdorff iff $R$ is closed in $X \times X$.

The following proposition, also from Tu [170] (Section 7.5, Theorem 7.9), yields a sufficient condition for second-countability.

**Proposition 3.33.** If $X$ is a topological space and $R$ is an open equivalence relation on $X$, then for any basis $\{B_\alpha\}$ for the topology of $X$, the family $\{\pi(B_\alpha)\}$ is a basis for the topology of $X/R$, where $\pi: X \to X/R$ is the projection map. Consequently, if $X$ is second-countable, then so is $X/R$.

Examples of quotient spaces, such as the Grassmannian and Stiefel manifolds, are discussed in Chapter 5, since their definitions require the notion of a group acting on a set.
Chapter 4

Introduction to Manifolds and Lie Groups

4.1 Introduction to Embedded Manifolds

In this section we define precisely manifolds, Lie groups and Lie algebras. One of the reasons that Lie groups are nice is that they have a differential structure, which means that the notion of tangent space makes sense at any point of the group. Furthermore, the tangent space at the identity happens to have some algebraic structure, that of a Lie algebra. Roughly, the tangent space at the identity provides a “linearization” of the Lie group, and it turns out that many properties of a Lie group are reflected in its Lie algebra, and that the loss of information is not too severe. The challenge that we are facing is that unless our readers are already familiar with manifolds, the amount of basic differential geometry required to define Lie groups and Lie algebras in full generality is overwhelming.

Fortunately, most of the Lie groups that we will consider are subspaces of \( \mathbb{R}^N \) for some sufficiently large \( N \). In fact, most of them are isomorphic to subgroups of \( \text{GL}(N, \mathbb{R}) \) for some suitable \( N \), even \( \text{SE}(n) \), which is isomorphic to a subgroup of \( \text{SL}(n+1) \). Such groups are called linear Lie groups (or matrix groups). Since these groups are subspaces of \( \mathbb{R}^N \), in a first stage, we do not need the definition of an abstract manifold. We just have to define embedded submanifolds (also called submanifolds) of \( \mathbb{R}^N \) (in the case of \( \text{GL}(n, \mathbb{R}) \), \( N = n^2 \)). This is the path that we will follow. The general definition of manifold will be given in Chapter 7.

Let us now provide the definition of an embedded submanifold. For simplicity, we restrict our attention to smooth manifolds. For detailed presentations, see DoCarmo [59, 60], Milnor [127], Marsden and Ratiu [121], Berger and Gostiaux [20], or Warner [175]. For the sake of brevity, we use the terminology manifold (but other authors would say embedded submanifolds, or something like that).

The intuition behind the notion of a smooth manifold in \( \mathbb{R}^N \) is that a subspace \( M \) is a manifold of dimension \( m \) if every point \( p \in M \) is contained in some open subset set \( U \) of
M (in the subspace topology) that can be parametrized by some function \( \varphi: \Omega \rightarrow U \) from some open subset \( \Omega \) of the origin in \( \mathbb{R}^m \), and that \( \varphi \) has some nice properties that allow the definition of smooth functions on \( M \) and of the tangent space at \( p \). For this, \( \varphi \) has to be at least a homeomorphism, but more is needed: \( \varphi \) must be smooth, and the derivative \( \varphi'(0_m) \) at the origin must be injective (letting \( 0_m = (0, \ldots, 0) \)).

**Definition 4.1.** Given any integers \( N, m \), with \( N \geq m \geq 1 \), an \( m \)-dimensional smooth manifold in \( \mathbb{R}^N \), for short a manifold, is a nonempty subset \( M \) of \( \mathbb{R}^N \) such that for every point \( p \in M \) there are two open subsets \( \Omega \subseteq \mathbb{R}^m \) and \( U \subseteq M \), with \( p \in U \), and a smooth function \( \varphi: \Omega \rightarrow \mathbb{R}^N \) such that \( \varphi \) is a homeomorphism between \( \Omega \) and \( U = \varphi(\Omega) \), and \( \varphi'(t_0) \) is injective, where \( t_0 = \varphi^{-1}(p) \); see Figure 4.1. The function \( \varphi: \Omega \rightarrow U \) is called a (local) parametrization of \( M \) at \( p \). If \( 0_m \in \Omega \) and \( \varphi(0_m) = p \), we say that \( \varphi: \Omega \rightarrow U \) is centered at \( p \).

![Figure 4.1: A manifold in \( \mathbb{R}^N \)](image)

Saying that \( \varphi'(t_0) \) is injective is equivalent to saying that \( \varphi \) is an immersion at \( t_0 \).

Recall that \( M \subseteq \mathbb{R}^N \) is a topological space under the subspace topology, and \( U \) is some open subset of \( M \) in the subspace topology, which means that \( U = M \cap W \) for some open subset \( W \) of \( \mathbb{R}^N \). Since \( \varphi: \Omega \rightarrow U \) is a homeomorphism, it has an inverse \( \varphi^{-1}: U \rightarrow \Omega \) that is also a homeomorphism, called a (local) chart. Since \( \Omega \subseteq \mathbb{R}^m \), for every point \( p \in M \) and every parametrization \( \varphi: \Omega \rightarrow U \) of \( M \) at \( p \), we have \( \varphi^{-1}(p) = (z_1, \ldots, z_m) \) for some \( z_i \in \mathbb{R} \), and we call \( z_1, \ldots, z_m \) the local coordinates of \( p \) (w.r.t. \( \varphi^{-1} \)). We often refer to a manifold \( M \) without explicitly specifying its dimension (the integer \( m \)).

Intuitively, a chart provides a “flattened” local map of a region on a manifold. For instance, in the case of surfaces (2-dimensional manifolds), a chart is analogous to a planar...
map of a region on the surface. For a concrete example, consider a map giving a planar representation of a country, a region on the earth, a curved surface.

**Remark:** We could allow $m = 0$ in Definition 4.1. If so, a manifold of dimension 0 is just a set of isolated points, and thus it has the discrete topology. In fact, it can be shown that a discrete subset of $\mathbb{R}^N$ is countable. Such manifolds are not very exciting, but they do correspond to discrete subgroups.

**Example 4.1.** The unit sphere $S^2$ in $\mathbb{R}^3$ defined such that

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

is a smooth 2-manifold, because it can be parametrized using the following two maps $\varphi_1$ and $\varphi_2$:

$$\varphi_1: (u, v) \mapsto \left( \frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1} \right)$$

and

$$\varphi_2: (u, v) \mapsto \left( \frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{1 - u^2 - v^2}{u^2 + v^2 + 1} \right).$$

The map $\varphi_1$ corresponds to the inverse of the stereographic projection from the north pole $N = (0, 0, 1)$ onto the plane $z = 0$, and the map $\varphi_2$ corresponds to the inverse of the stereographic projection from the south pole $S = (0, 0, -1)$ onto the plane $z = 0$, as illustrated in Figure 4.2.

![Figure 4.2: Inverse stereographic projections](image)
We demonstrate the algebraic constructions of $\varphi_1$ and $\varphi_1^{-1}$ leaving the constructions of $\varphi_2$ and $\varphi_2^{-1}$ to the reader. Take $S^2$ and a point $Z = (x_1, x_2, x_3) \in S^2 - \{(0,0,1)\}$ and form $l$, the line connecting $(0,0,1)$ and $Z$. Line $l$ intersects the $xy$-plane at point $(u,v,0)$ and has equation $p + (1-t)\overrightarrow{v}$ where $p = (0,0,1)$ and $\overrightarrow{v} = (u,v,0) - (0,0,1) = (u,v,-1)$. See Figure 4.3.

![Figure 4.3: Line l is in red.](image)

In other words, the line segment on Line $l$ between $(u,v,0)$ and $(0,0,1)$ is parametrized by $((1-t)u,(1-t)v,t)$ for $0 \leq t \leq 1$. The intersection of this line segment and $S^2$ is characterized by the equation

$$(1-t)^2u^2 + (1-t)^2v^2 + t^2 = 1, \quad 0 < t < 1.$$ 

Take this equation, subtract $t^2$, and divide by $1-t$ to obtain

$$(1-t)(u^2 - v^2) = 1 + t.$$ 

Solving this latter equation for $t$ yields

$$t = \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1} \quad \text{and} \quad 1-t = \frac{2}{u^2 + v^2 + 1}.$$ 

By construction we know the intersection of the line segment with $S^2$ is $Z = (x_1,x_2,x_3)$. Hence we conclude that

$$x_1 = (1-t)u = \frac{2u}{u^2 + v^2 + 1}, \quad x_2 = (1-t)v = \frac{2v}{u^2 + v^2 + 1}, \quad x_3 = t = \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}.$$
4.1. INTRODUCTION TO EMBEDDED MANIFOLDS

To calculate $\varphi_1^{-1}$, we parameterize $l$ by $((1-t)x_1, (1-t)x_2, (1-t)(x_3-1)+1)$. The intersection of Line $l$ with the $xy$-plane is characterized by $((1-t)x_1, (1-t)x_2, (1-t)(x_3-1)+1) = (u, v, 0)$ and gives

$$(1-t)(x_3 - 1) + 1 = 0.$$ 

Solving this equation for $x_3$ implies that $t = -\frac{x_3}{1-x_3}$ and $1-t = \frac{1}{1-x_3}$.

Hence $\varphi_1^{-1}(x_1, x_2, x_3) = (u,v)$, where

$$u = (1-t)x_1 = \frac{x_1}{1-x_3}, \quad v = (1-t)x_2 = \frac{x_2}{1-x_3}.$$ 

We leave as an exercise to check that the map $\varphi_1$ parametrizes $S^2 - \{N\}$ and that the map $\varphi_2$ parametrizes $S^2 - \{S\}$ (and that they are smooth, homeomorphisms, etc.). Using $\varphi_1$, the open lower hemisphere is parametrized by the open disk of center $O$ and radius 1 contained in the plane $z = 0$.

The chart $\varphi_1^{-1}$ assigns local coordinates to the points in the open lower hemisphere. If we draw a grid of coordinate lines parallel to the $x$ and $y$ axes inside the open unit disk and map these lines onto the lower hemisphere using $\varphi_1$, we get curved lines on the lower hemisphere. These “coordinate lines” on the lower hemisphere provide local coordinates for every point on the lower hemisphere. For this reason, older books often talk about curvilinear coordinate systems to mean the coordinate lines on a surface induced by a chart. We urge our readers to define a manifold structure on a torus. This can be done using four charts.

Every open subset of $\mathbb{R}^N$ is a manifold in a trivial way. Indeed, we can use the inclusion map as a parametrization. In particular, $\text{GL}(n, \mathbb{R})$ is an open subset of $\mathbb{R}^{n^2}$, since its complement is closed (the set of invertible matrices is the inverse image of the determinant function, which is continuous). Thus, $\text{GL}(n, \mathbb{R})$ is a manifold. We can view $\text{GL}(n, \mathbb{C})$ as a subset of $\mathbb{R}^{(2n)^2}$ using the embedding defined as follows: For every complex $n \times n$ matrix $A$, construct the real $2n \times 2n$ matrix such that every entry $a + ib$ in $A$ is replaced by the $2 \times 2$ block

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

where $a, b \in \mathbb{R}$. It is immediately verified that this map is in fact a group isomorphism. Thus, we can view $\text{GL}(n, \mathbb{C})$ as a subgroup of $\text{GL}(2n, \mathbb{R})$, and as a manifold in $\mathbb{R}^{(2n)^2}$.

A 1-manifold is called a (smooth) curve, and a 2-manifold is called a (smooth) surface (although some authors require that they also be connected).

The following two lemmas provide the link with the definition of an abstract manifold. The first lemma is shown using Proposition 4.4 and is Condition (2) of Theorem 4.6; see below.
Lemma 4.1. Given an $m$-dimensional manifold $M$ in $\mathbb{R}^N$, for every $p \in M$ there are two open sets $O, W \subseteq \mathbb{R}^N$ with $0_N \in O$ and $p \in M \cap W$, and a smooth diffeomorphism $\varphi: O \to W$, such that $\varphi(0_N) = p$ and

$$\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.$$ 

There is an open subset $\Omega$ of $\mathbb{R}^m$ such that

$$O \cap (\mathbb{R}^m \times \{0_{N-m}\}) = \Omega \times \{0_{N-m}\},$$

and the map $\psi: \Omega \to \mathbb{R}^N$ given by

$$\psi(x) = \varphi(x, 0_{N-m})$$

is an immersion and a homeomorphism onto $U = W \cap M$; so $\psi$ is a parametrization of $M$ at $p$. We can think of $\varphi$ as a promoted version of $\psi$ which is actually a diffeomorphism between open subsets of $\mathbb{R}^N$; see Figure 4.4.

Figure 4.4: An illustration of Lemma 4.1, where $M$ is a surface embedded in $\mathbb{R}^3$, namely $m = 2$ and $N = 3$.

The next lemma is easily shown from Lemma 4.1 (see Berger and Gostiaux [20], Theorem 2.1.9 or DoCarmo [60], Chapter 0, Section 4). It is a key technical result used to show that interesting properties of maps between manifolds do not depend on parametrizations.
Lemma 4.2. Given an $m$-dimensional manifold $M$ in $\mathbb{R}^N$, for every $p \in M$ and any two parametrizations $\varphi_1: \Omega_1 \to U_1$ and $\varphi_2: \Omega_2 \to U_2$ of $M$ at $p$, if $U_1 \cap U_2 \neq \emptyset$, the map $\varphi_2^{-1} \circ \varphi_1: \varphi_1^{-1}(U_1 \cap U_2) \to \varphi_2^{-1}(U_1 \cap U_2)$ is a smooth diffeomorphism.

The maps $\varphi_2^{-1} \circ \varphi_1: \varphi_1^{-1}(U_1 \cap U_2) \to \varphi_2^{-1}(U_1 \cap U_2)$ are called transition maps. Lemma 4.2 is illustrated in Figure 4.5.

![Figure 4.5: Parametrizations and transition functions](image)

Using Definition 4.1, it may be quite hard to prove that a space is a manifold. Therefore, it is handy to have alternate characterizations such as those given in the next Proposition, which is Condition (3) of Theorem 4.6. An illustration of Proposition 4.3 is given by Figure 4.6.

Proposition 4.3. A subset $M \subseteq \mathbb{R}^{m+k}$ is an $m$-dimensional manifold iff either

1. For every $p \in M$, there is some open subset $W \subseteq \mathbb{R}^{m+k}$ with $p \in W$, and a (smooth) submersion $f: W \to \mathbb{R}^k$, so that $W \cap M = f^{-1}(0)$,

or

2. For every $p \in M$, there is some open subset $W \subseteq \mathbb{R}^{m+k}$ with $p \in W$, and a (smooth) map $f: W \to \mathbb{R}^k$, so that $f'(p)$ is surjective and $W \cap M = f^{-1}(0)$.

Observe that condition (2), although apparently weaker than condition (1), is in fact equivalent to it, but more convenient in practice. This is because to say that $f'(p)$ is surjective means that the Jacobian matrix of $f'(p)$ has rank $k$, which means that some determinant is nonzero, and because the determinant function is continuous this must hold in some open
subset $W_1 \subseteq W$ containing $p$. Consequently, the restriction $f_1$ of $f$ to $W_1$ is indeed a submersion, and $f_1^{-1}(0) = W_1 \cap f^{-1}(0) = W_1 \cap W \cap M = W_1 \cap M$.

Figure 4.6: An illustration of Proposition 4.3, where $M$ is the torus, $m = 2$, and $k = 1$. Note that $f^{-1}(0)$ is the pink patch of the torus, i.e. the zero level set of the open ball $W$.

A proof of Proposition 4.3 can be found in Lafontaine [110] or Berger and Gostiaux [20]. Lemma 4.1 and Proposition 4.3 are actually equivalent to Definition 4.1. This equivalence is also proved in Lafontaine [110] and Berger and Gostiaux [20].

Theorem 4.6, which combines Propositions 4.1 and 4.3, provides four equivalent characterizations of when a subspace of $\mathbb{R}^N$ is a manifold of dimension $m$. Its proof, which is somewhat illuminating, is based on two technical lemmas that are proved using the inverse function theorem (for example, see Guillemin and Pollack [83], Chapter 1, Sections 3 and 4).

**Lemma 4.4.** Let $U \subseteq \mathbb{R}^m$ be an open subset of $\mathbb{R}^m$ and pick some $a \in U$. If $f: U \to \mathbb{R}^n$ is a smooth immersion at $a$, i.e., $df_a$ is injective (so, $m \leq n$), then there is an open set $V \subseteq \mathbb{R}^n$ with $f(a) \in V$, an open subset $U' \subseteq U$ with $a \in U'$ and $f(U') \subseteq V$, an open subset $O \subseteq \mathbb{R}^{n-m}$, and a diffeomorphism $\theta: V \to U' \times O$, so that

$$\theta(f(x_1, \ldots, x_m)) = (x_1, \ldots, x_m, 0, \ldots, 0),$$
for all \((x_1, \ldots, x_m) \in U'\), as illustrated in the diagram below

\[
U' \subseteq U \xrightarrow{f} f(U') \subseteq V
\]

where \(i_{n1}(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0)\); see Figure 4.7.

Figure 4.7: An illustration of Lemma 4.4, where \(m = 2\) and \(n = 3\). Note that \(U'\) is the base of the cylinder and \(\Theta\) is the diffeomorphism between the cylinder and the egged shaped \(V\). The composition \(\Theta \circ f\) maps the pink patch in \(V\) onto \(U'\).

Proof. Since \(f\) is an immersion, its Jacobian matrix \(J(f)\) (an \(n \times m\) matrix) has rank \(m\), and by permuting coordinates if needed, we may assume that the first \(m\) rows of \(J(f)\) are linearly independent and we let

\[
A = \left( \frac{\partial f_i}{\partial x_j}(a) \right)
\]

be this invertible \(m \times m\) matrix. Define the map \(g: U \times \mathbb{R}^{n-m} \to \mathbb{R}^n\) by

\[
g(x, y) = (f_1(x), \ldots, f_m(x), y_1 + f_{m+1}(x), \ldots, y_{n-m} + f_n(x)),
\]
for all \( x \in U \) and all \( y \in \mathbb{R}^{n-m} \). The Jacobian matrix of \( g \) at \((a, 0)\) is of the form

\[
J = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix},
\]

so \( \det(J) = \det(A) \det(I) = \det(A) \neq 0 \), since \( A \) is invertible. By the inverse function theorem, there are some open subsets \( W \subseteq U \times \mathbb{R}^{n-m} \) with \((a, 0) \in W\) and \( V \subseteq \mathbb{R}^n \) such that the restriction of \( g \) to \( W \) is a diffeomorphism between \( W \) and \( V \). Since \( W \subseteq U \times \mathbb{R}^{n-m} \) is an open set, we can find some open subsets \( U' \subseteq U \) and \( O \subseteq \mathbb{R}^{n-m} \) with \( a \in U' \) and \( V' \subseteq \mathbb{R}^n \) such that the restriction of \( g \) to \( U' \times O \) is a diffeomorphism between \( U' \times O \) and \( V' \). If \( \theta: V' \rightarrow U' \times O \) is the inverse of this diffeomorphism, then \( f(U') \subseteq V \) and since \( g(x, 0) = f(x) \),

\[
\theta(g(x, 0)) = \theta(f(x_1, \ldots, x_m)) = (x_1, \ldots, x_m, 0, \ldots, 0),
\]

for all \( x = (x_1, \ldots, x_m) \in U' \). \( \square \)

**Lemma 4.5.** Let \( W \subseteq \mathbb{R}^m \) be an open subset of \( \mathbb{R}^m \) and pick some \( a \in W \). If \( f: W \rightarrow \mathbb{R}^n \) is a smooth submersion at \( a \), i.e., \( df_a \) is surjective (so, \( m \geq n \)), then there is an open set \( V \subseteq W \subseteq \mathbb{R}^m \) with \( a \in V \), and a diffeomorphism \( \psi: O \rightarrow V \) with domain \( O \subseteq \mathbb{R}^m \), so that

\[
f(\psi(x_1, \ldots, x_m)) = (x_1, \ldots, x_n),
\]

for all \((x_1, \ldots, x_m) \in O\), as illustrated in the diagram below

\[
\begin{array}{ccc}
O \subseteq \mathbb{R}^m & \xrightarrow{\psi} & V \subseteq W \subseteq \mathbb{R}^m \\
\pi \downarrow & & \downarrow f \\
\mathbb{R}^n, & & \\
\end{array}
\]

where \( \pi(x_1, \ldots, x_m) = (x_1, \ldots, x_n) \); see Figure 4.8.

**Proof.** Since \( f \) is a submersion, its Jacobian matrix \( J(f) \) (an \( n \times m \) matrix) has rank \( n \), and by permuting coordinates if needed, we may assume that the first \( n \) columns of \( J(f) \) are linearly independent and we let

\[
A = \begin{pmatrix} \frac{\partial f_i}{\partial x_j}(a) \end{pmatrix}
\]

be this invertible \( n \times n \) matrix. Define the map \( g: W \rightarrow \mathbb{R}^m \) by

\[
g(x) = (f(x), x_{n+1}, \ldots, x_m),
\]

for all \( x \in W \). The Jacobian matrix of \( g \) at \( a \) is of the form

\[
J = \begin{pmatrix} A & B \\ 0 & I \end{pmatrix},
\]
4.1. INTRODUCTION TO EMBEDDED MANIFOLDS

Figure 4.8: An illustration of Lemma 4.5, where \( m = 3 \) and \( n = 2 \). Note that \( \psi \) is the diffeomorphism between the 0 and the purple ball \( V \). The composition \( f \circ \psi \) projects \( O \) onto its equatorial pink disk.

so \( \det(J) = \det(A)\det(I) = \det(A) \neq 0 \), since \( A \) is invertible. By the inverse function theorem, there are some open subsets \( V \subseteq W \) with \( a \in V \) and \( O \subseteq \mathbb{R}^m \) such that the restriction of \( g \) to \( V \) is a diffeomorphism between \( V \) and \( O \). Let \( \psi: O \to V \) be the inverse of this diffeomorphism. Because \( g \circ \psi = \text{id} \), we have

\[
(x_1, \ldots, x_m) = g(\psi(x)) = (f(\psi(x)), \psi_{n+1}(x), \ldots, \psi_m(x)),
\]

that is,

\[
f(\psi(x_1, \ldots, x_m)) = (x_1, \ldots, x_n)
\]

for all \((x_1, \ldots, x_m) \in O\), as desired. \( \square \)

Using Lemmas 4.4 and 4.5, we can prove the following theorem which confirms that all our characterizations of a manifold are equivalent.

**Theorem 4.6.** A nonempty subset \( M \subseteq \mathbb{R}^N \) is an \( m \)-manifold (with \( 1 \leq m \leq N \)) iff any of the following conditions hold:

1. For every \( p \in M \), there are two open subsets \( \Omega \subseteq \mathbb{R}^m \) and \( U \subseteq M \) with \( p \in U \), and a smooth function \( \varphi: \Omega \to \mathbb{R}^N \) such that \( \varphi \) is a homeomorphism between \( \Omega \) and \( U = \varphi(\Omega) \), and \( \varphi'(0) \) is injective, where \( p = \varphi(0) \).

2. For every \( p \in M \), there are two open sets \( O, W \subseteq \mathbb{R}^N \) with \( 0_N \in O \) and \( p \in M \cap W \), and a smooth diffeomorphism \( \varphi: O \to W \), such that \( \varphi(0_N) = p \) and

\[
\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.
\]
(3) For every \( p \in M \), there is some open subset \( W \subseteq \mathbb{R}^N \) with \( p \in W \), and a smooth submersion \( f: W \to \mathbb{R}^{N-m} \), so that \( W \cap M = f^{-1}(0) \).

(4) For every \( p \in M \), there is some open subset \( W \subseteq \mathbb{R}^N \) with \( p \in W \), and \( N-m \) smooth functions \( f_i: W \to \mathbb{R} \), so that the linear forms \( df_1(p), \ldots, df_{N-m}(p) \) are linearly independent, and
\[
W \cap M = f_1^{-1}(0) \cap \cdots \cap f_{N-m}^{-1}(0).
\]
See Figure 4.9.

![Figure 4.9: An illustration of Condition (4) in Theorem 4.6, where \( N = 3 \) and \( m = 1 \). The manifold \( M \) is the helix in \( \mathbb{R}^3 \). The dark green portion of \( M \) is magnified in order to show that it is the intersection of the pink surface, \( f_1^{-1}(0) \), and the blue surface, \( f_2^{-1}(0) \).](image)

**Proof.** If (1) holds, then by Lemma 4.4, replacing \( \Omega \) by a smaller open subset \( \Omega' \subseteq \Omega \) if necessary, there is some open subset \( V \subseteq \mathbb{R}^N \) with \( p \in V \) and \( \varphi(\Omega') \subseteq V \), an open subset \( O' \subseteq \mathbb{R}^{N-m} \), and some diffeomorphism \( \theta: V \to \Omega' \times O' \), so that
\[
(\theta \circ \varphi)(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0),
\]
for all \( (x_1, \ldots, x_m) \in \Omega' \). Observe that the above condition implies that
\[
(\theta \circ \varphi)(\Omega') = \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).
\]
Since \( \varphi \) is a homeomorphism between \( \Omega \) and its image in \( M \) and since \( \Omega' \subseteq \Omega \) is an open subset, \( \varphi(\Omega') = M \cap W' \) for some open subset \( W' \subseteq \mathbb{R}^N \), so if we let \( W = V \cap W' \), because \( \varphi(\Omega') \subseteq V \), it follows that \( \varphi(\Omega') = M \cap W \) and
\[
\theta(W \cap M) = \theta(\varphi(\Omega')) = \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).
\]
4.1. INTRODUCTION TO EMBEDDED MANIFOLDS

However, \( \theta \) is injective and \( \theta(W \cap M) \subseteq \theta(W) \), so

\[
\theta(W \cap M) = \theta(W) \cap \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}) = \theta(W) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}) = \theta(W) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).
\]

If we let \( O = \theta(W) \), we get

\[
\theta^{-1}(O \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\})) = M \cap W,
\]

which is (2).

If (2) holds, we can write \( \varphi^{-1} = (f_1, \ldots, f_N) \) and because \( \varphi^{-1}: W \rightarrow O \) is a diffeomorphism, \( df_1(q), \ldots, df_N(q) \) are linearly independent for all \( q \in W \), so the map

\[
f = (f_{m+1}, \ldots, f_N)
\]
is a submersion \( f: W \rightarrow \mathbb{R}^{N-m} \), and we have \( f(x) = 0 \) iff \( f_{m+1}(x) = \cdots = f_N(x) = 0 \) iff

\[
\varphi^{-1}(x) = (f_1(x), \ldots, f_m(x), 0, \ldots, 0)
\]

iff \( \varphi^{-1}(x) \in O \cap (\mathbb{R}^m \times \{0_{N-m}\}) \) iff \( x \in \varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W \), because

\[
\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.
\]

Thus, \( M \cap W = f^{-1}(0) \), which is (3).

The proof that (3) implies (2) uses Lemma 4.5 instead of Lemma 4.4. If \( f: W \rightarrow \mathbb{R}^{N-m} \) is the submersion such that \( M \cap W = f^{-1}(0) \) given by (3), then by Lemma 4.5, there are open subsets \( V \subseteq W \), \( O \subseteq \mathbb{R}^N \) and a diffeomorphism \( \psi: O \rightarrow V \), so that

\[
f(\psi(x_1, \ldots, x_N)) = (x_1, \ldots, x_{N-m})
\]

for all \( (x_1, \ldots, x_N) \in O \). If \( \sigma \) is the permutation of variables given by

\[
\sigma(x_1, \ldots, x_m, x_{m+1}, \ldots, x_N) = (x_{m+1}, \ldots, x_N, x_1, \ldots, x_m),
\]

then \( \varphi = \psi \circ \sigma \) is a diffeomorphism such that

\[
f(\varphi(x_1, \ldots, x_N)) = (x_{m+1}, \ldots, x_N)
\]

for all \( (x_1, \ldots, x_N) \in O \). If we denote the restriction of \( f \) to \( V \) by \( g \), it is clear that

\[
M \cap V = g^{-1}(0),
\]

and because \( g(\varphi(x_1, \ldots, x_N)) = 0 \) iff \( (x_{m+1}, \ldots, x_N) = 0_{N-m} \) and \( \varphi \) is a bijection,

\[
M \cap V = \{(y_1, \ldots, y_N) \in V \mid g(y_1, \ldots, y_N) = 0\} = \varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})).
\]
which is (2).

If (2) holds, then \( \varphi: O \to W \) is a diffeomorphism,

\[ O \cap (\mathbb{R}^m \times \{0_{N-m}\}) = \Omega \times \{0_{N-m}\} \]

for some open subset \( \Omega \subseteq \mathbb{R}^m \), and the map \( \psi: \Omega \to \mathbb{R}^N \) given by

\[ \psi(x) = \varphi(x, 0_{N-m}) \]

is an immersion on \( \Omega \) and a homeomorphism onto \( W \cap M \), which implies (1).

If (3) holds, then if we write \( f = (f_1, \ldots, f_{N-m}) \), with \( f_i: W \to \mathbb{R} \), then the fact that \( df(p) \) is a submersion is equivalent to the fact that the linear forms \( df_1(p), \ldots, df_{N-m}(p) \) are linearly independent and

\[ M \cap W = f^{-1}(0) = f_1^{-1}(0) \cap \cdots \cap f_{N-m}^{-1}(0). \]

Finally, if (4) holds, then if we define \( f: W \to \mathbb{R}^{N-m} \) by

\[ f = (f_1, \ldots, f_{N-m}), \]

because \( df_1(p), \ldots, df_{N-m}(p) \) are linearly independent we get a smooth map which is a submersion at \( p \) such that

\[ M \cap W = f^{-1}(0). \]

Now, \( f \) is a submersion at \( p \) iff \( df(p) \) is surjective, which means that a certain determinant is nonzero, and since the determinant function is continuous, this determinant is nonzero on some open subset \( W' \subseteq W \) containing \( p \), so if we restrict \( f \) to \( W' \), we get an immersion on \( W' \) such that \( M \cap W' = f^{-1}(0) \).

Condition (4) says that locally (that is, in a small open set of \( M \) containing \( p \in M \)), \( M \) is “cut out” by \( N - m \) smooth functions \( f_i: W \to \mathbb{R} \), in the sense that the portion of the manifold \( M \cap W \) is the intersection of the \( N - m \) hypersurfaces \( f_i^{-1}(0) \) (the zero-level sets of the \( f_i \)), and that this intersection is “clean,” which means that the linear forms \( df_1(p), \ldots, df_{N-m}(p) \) are linearly independent.

As an illustration of Theorem 4.6, we can show again that the sphere

\[ S^n = \{ x \in \mathbb{R}^{n+1} \mid \|x\|_2^2 - 1 = 0 \} \]

is an \( n \)-dimensional manifold in \( \mathbb{R}^{n+1} \). Indeed, the map \( f: \mathbb{R}^{n+1} \to \mathbb{R} \) given by \( f(x) = \|x\|_2^2 - 1 \) is a submersion (for \( x \neq 0 \)), since

\[ df(x)(y) = 2 \sum_{k=1}^{n+1} x_k y_k. \]
4.1. INTRODUCTION TO EMBEDDED MANIFOLDS

We can also show that the rotation group $SO(n)$ is an $\frac{n(n-1)}{2}$-dimensional manifold in $\mathbb{R}^{n^2}$.

Indeed, $GL^+(n)$ is an open subset of $\mathbb{R}^{n^2}$ of dimension $n^2$ (recall, $GL^+(n) = \{ A \in GL(n) \mid \det(A) > 0 \}$), and if $f$ is defined by

$$f(A) = A^T A - I,$$

where $A \in GL^+(n)$, then $f(A)$ is symmetric, so $f(A) \in S(n) = \mathbb{R}^{\frac{n(n+1)}{2}}$. We proved in Section 2.2 that

$$df(A)(H) = A^T H + H^T A.$$

But then, $df(A)$ is surjective for all $A \in SO(n)$, because if $S$ is any symmetric matrix, we see that

$$df(A)(AS/2) = A^T \frac{AS}{2} + \left(\frac{AS}{2}\right)^T A = A^T A \frac{S}{2} + \frac{S^T}{2} A^T A = \frac{S}{2} + \frac{S^T}{2} = S.$$

As $SO(n) = f^{-1}(0)$, we conclude that $SO(n)$ is indeed a manifold.

A similar argument proves that $O(n)$ is an $\frac{n(n-1)}{2}$-dimensional manifold.

Using the map $f : GL(n) \to \mathbb{R}$ given by $A \mapsto \det(A)$, we can prove that $SL(n)$ is a manifold of dimension $n^2 - 1$.

Remark: We have $df(A)(B) = \det(A) \text{tr}(A^{-1}B)$ for every $A \in GL(n)$, where $f(A) = \det(A)$.

A class of manifolds generalizing the spheres and the orthogonal groups are the Stiefel manifolds. For any $n \geq 1$ and any $k$ with $1 \leq k \leq n$, let $S(k,n)$ be the set of all orthonormal $k$-frames; that is, of $k$-tuples of orthonormal vectors $(u_1, \ldots, u_k)$ with $u_i \in \mathbb{R}^n$. Obviously $S(1,n) = S^{n-1}$, and $S(n,n) = O(n)$. Every orthonomal $k$-frame $(u_1, \ldots, u_k)$ can be represented by an $n \times k$ matrix $Y$ over the canonical basis of $\mathbb{R}^n$, and such a matrix $Y$ satisfies the equation

$$Y^T Y = I.$$

Thus, $S(k,n)$ can be viewed as a subspace of $M_{n,k}(\mathbb{R})$, where $M_{n,k}(\mathbb{R})$ denotes the vector space of all $n \times k$ matrices with real entries. We claim that $S(k,n)$ is a manifold. Let $W = \{ A \in M_{n,k}(\mathbb{R}) \mid \det(A^T A) > 0 \}$, an open subset of $M_{n,k}(\mathbb{R})$ such that $S(k,n) \subseteq W$ (since if $A \in S(k,n)$, then $A^T A = I$, so $\det(A^T A) = 1$). Generalizing the situation involving $SO(n)$, define the function $f : W \to S(k)$ by

$$f(A) = A^T A - I.$$

Basically the same computation as in the case of $SO(n)$ yields

$$df(A)(H) = A^T H + H^T A.$$
The proof that $df(A)$ is surjective for all $A \in S(k, n)$ is the same as before, because only the equation $A^\top A = I$ is needed. Indeed, given any symmetric matrix $S \in S(k) \approx \mathbb{R}^{\frac{k(k+1)}{2}}$, we have from our previous calculation that
\[ df(A) \left( \frac{AS}{2} \right) = S. \]
As $S(k, n) = f^{-1}(0)$, we conclude that $S(k, n)$ is a smooth manifold of dimension
\[ nk - \frac{k(k + 1)}{2} = k(n - k) + \frac{k(k - 1)}{2}. \]

The third characterization of Theorem 4.6 suggests the following definition.

**Definition 4.2.** Let $f: \mathbb{R}^{m+k} \to \mathbb{R}^k$ be a smooth function. A point $p \in \mathbb{R}^{m+k}$ is called a critical point (of $f$) iff $df_p$ is not surjective, and a point $q \in \mathbb{R}^k$ is called a critical value (of $f$) iff $q = f(p)$ for some critical point $p \in \mathbb{R}^{m+k}$. A point $p \in \mathbb{R}^{m+k}$ is a regular point (of $f$) iff $p$ is not critical, i.e., $df_p$ is surjective, and a point $q \in \mathbb{R}^k$ is a regular value (of $f$) iff it is not a critical value. In particular, any $q \in \mathbb{R}^k - f(\mathbb{R}^{m+k})$ is a regular value, and $q \in f(\mathbb{R}^{m+k})$ is a regular value iff every $p \in f^{-1}(q)$ is a regular point (in contrast, $q$ is a critical value iff some $p \in f^{-1}(q)$ is critical).

Part (3) of Theorem 4.6 implies the following useful proposition:

**Proposition 4.7.** Given any smooth function $f: \mathbb{R}^{m+k} \to \mathbb{R}^k$, for every regular value $q \in f(\mathbb{R}^{m+k})$, the preimage $Z = f^{-1}(q)$ is a manifold of dimension $m$.

Definition 4.2 and Proposition 4.7 can be generalized to manifolds. Regular and critical values of smooth maps play an important role in differential topology. Firstly, given a smooth map $f: \mathbb{R}^{m+k} \to \mathbb{R}^k$, almost every point of $\mathbb{R}^k$ is a regular value of $f$. To make this statement precise, one needs the notion of a set of measure zero. Then, Sard’s theorem says that the set of critical values of a smooth map has measure zero. Secondly, if we consider smooth functions $f: \mathbb{R}^{m+1} \to \mathbb{R}$, a point $p \in \mathbb{R}^{m+1}$ is critical iff $df_p = 0$. Then, we can use second order derivatives to further classify critical points. The Hessian matrix of $f$ (at $p$) is the matrix of second-order partials
\[ H_f(p) = \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(p) \right), \]
and a critical point $p$ is a nondegenerate critical point if $H_f(p)$ is a nonsingular matrix. The remarkable fact is that, at a nondegenerate critical point $p$, the local behavior of $f$ is completely determined, in the sense that after a suitable change of coordinates (given by a smooth diffeomorphism)
\[ f(x) = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{m+1}^2 \]
4.1. INTRODUCTION TO EMBEDDED MANIFOLDS

near \( p \), where \( \lambda \), called the index of \( f \) at \( p \), is an integer which depends only on \( p \) (in fact, \( \lambda \) is the number of negative eigenvalues of \( H_f(p) \)). This result is known as Morse lemma (after Marston Morse, 1892-1977).

Smooth functions whose critical points are all nondegenerate are called Morse functions. It turns out that every smooth function \( f: \mathbb{R}^{m+1} \to \mathbb{R} \) gives rise to a large supply of Morse functions by adding a linear function to it. More precisely, the set of \( a \in \mathbb{R}^{m+1} \) for which the function \( f_a \) given by

\[
f_a(x) = f(x) + a_1x_1 + \cdots + a_{m+1}x_{m+1}
\]

is not a Morse function has measure zero.

Morse functions can be used to study topological properties of manifolds. In a sense to be made precise and under certain technical conditions, a Morse function can be used to reconstruct a manifold by attaching cells, up to homotopy equivalence. However, these results are way beyond the scope of this book. A fairly elementary exposition of nondegenerate critical points and Morse functions can be found in Guillemin and Pollack [83] (Chapter 1, Section 7). Sard’s theorem is proved in Appendix 1 of Guillemin and Pollack [83] and also in Chapter 2 of Milnor [127]. Morse theory (starting with Morse lemma) and much more, is discussed in Milnor [125], widely recognized as a mathematical masterpiece. An excellent and more leisurely introduction to Morse theory is given in Matsumoto [124], where a proof of Morse lemma is also given.

Let us now introduce the definitions of a smooth curve in a manifold and the tangent vector at a point of a curve.

**Definition 4.3.** Let \( M \) be an \( m \)-dimensional manifold in \( \mathbb{R}^N \). A smooth curve \( \gamma \) in \( M \) is any function \( \gamma: I \to M \) where \( I \) is an open interval in \( \mathbb{R} \) and such that for every \( t \in I \), letting \( p = \gamma(t) \), there is some parametrization \( \varphi: \Omega \to U \) of \( M \) at \( p \) and some open interval \((t - \epsilon, t + \epsilon) \subseteq I \) such that the curve \( \varphi^{-1} \circ \gamma: (t - \epsilon, t + \epsilon) \to \mathbb{R}^m \) is smooth.

The notion of a smooth curve is illustrated in Figure 4.10.

Using Lemma 4.2, it is easily shown that Definition 4.3 does not depend on the choice of the parametrization \( \varphi: \Omega \to U \) at \( p \).

Lemma 4.2 also implies that \( \gamma \) viewed as a curve \( \gamma: I \to \mathbb{R}^N \) is smooth. Then the tangent vector to the curve \( \gamma: I \to \mathbb{R}^N \) at \( t \), denoted by \( \gamma'(t) \), is the value of the derivative of \( \gamma \) at \( t \) (a vector in \( \mathbb{R}^N \)) computed as usual:

\[
\gamma'(t) = \lim_{h \to 0} \frac{\gamma(t + h) - \gamma(t)}{h}.
\]

Given any point \( p \in M \), we will show that the set of tangent vectors to all smooth curves in \( M \) through \( p \) is a vector space isomorphic to the vector space \( \mathbb{R}^m \). The tangent vector at \( p \) to a curve \( \gamma \) on a manifold \( M \) is illustrated in Figure 4.11.
Figure 4.10: A smooth curve in a manifold $M$

Figure 4.11: Tangent vector to a curve on a manifold
4.2. LINEAR LIE GROUPS

Given a smooth curve $\gamma: I \to M$, for any $t \in I$, letting $p = \gamma(t)$, since $M$ is a manifold, there is a parametrization $\varphi: \Omega \to U$ such that $\varphi(0_m) = p \in U$ and some open interval $J \subseteq I$ with $t \in J$ and such that the function

$$\varphi^{-1} \circ \gamma: J \to \mathbb{R}^m$$

is a smooth curve, since $\gamma$ is a smooth curve. Letting $\alpha = \varphi^{-1} \circ \gamma$, the derivative $\alpha'(t)$ is well-defined, and it is a vector in $\mathbb{R}^m$. But $\varphi \circ \alpha: J \to M$ is also a smooth curve, which agrees with $\gamma$ on $J$, and by the chain rule,

$$\gamma'(t) = \varphi'(0_m)(\alpha'(t)),$$

since $\alpha(t) = 0_m$ (because $\varphi(0_m) = p$ and $\gamma(t) = p$). See Figure 4.10. Observe that $\gamma'(t)$ is a vector in $\mathbb{R}^N$. Now, for every vector $v \in \mathbb{R}^m$, the curve $\alpha: J \to \mathbb{R}^m$ defined such that $\alpha(u) = (u - t)v$ for all $u \in J$ is clearly smooth, and $\alpha'(t) = v$. This shows that the set of tangent vectors at $t$ to all smooth curves (in $\mathbb{R}^m$) passing through $0_m$ is the entire vector space $\mathbb{R}^m$. Since every smooth curve $\gamma: I \to M$ agrees with a curve of the form $\varphi \circ \alpha: J \to M$ for some smooth curve $\alpha: J \to \mathbb{R}^m$ (with $J \subseteq I$) as explained above, and since it is assumed that $\varphi'(0_m)$ is injective, $\varphi'(0_m)$ maps the vector space $\mathbb{R}^m$ injectively to the set of tangent vectors to $\gamma$ at $p$, as claimed. All this is summarized in the following definition.

**Definition 4.4.** Let $M$ be an $m$-dimensional manifold in $\mathbb{R}^N$. For every point $p \in M$, the tangent space $T_p M$ at $p$ is the set of all vectors in $\mathbb{R}^N$ of the form $\gamma'(0)$, where $\gamma: I \to M$ is any smooth curve in $M$ such that $p = \gamma(0)$. The set $T_p M$ is a vector space isomorphic to $\mathbb{R}^m$. Every vector $v \in T_p M$ is called a tangent vector to $M$ at $p$.

**Remark:** The definition of a tangent vector at $p$ involves smooth curves, where a smooth curve is defined in Definition 4.3. Actually, because of Lemma 4.1, it is only necessary to use curves that are $C^1$ viewed as curves in $\mathbb{R}^N$. The potential problem is that if $\varphi$ is a parametrization at $p$ and $\gamma$ is a $C^1$ curve, it is not obvious that $\varphi^{-1} \circ \gamma$ is $C^1$ in $\mathbb{R}^m$. However, Lemma 4.1 allows us to promote $\varphi$ to a diffeomorphism between open subsets of $\mathbb{R}^N$, and since both $\gamma$ and (this new) $\varphi^{-1}$ are $C^1$, so is $\varphi^{-1} \circ \gamma$. However, in the more general case of an abstract manifold $M$ not assumed to be contained in some $\mathbb{R}^N$, smooth curves have to be defined as in Definition 4.3.

### 4.2 Linear Lie Groups

We can now define Lie groups (postponing defining smooth maps). In general, the difficult part in proving that a subgroup of $\text{GL}(n, \mathbb{R})$ is a Lie group is to prove that it is a manifold. Fortunately, there is a characterization of the linear groups that obviates much of the work.
CHAPTER 4. INTRODUCTION TO MANIFOLDS AND LIE GROUPS

This characterization rests on two theorems. First, a Lie subgroup $H$ of a Lie group $G$ (where $H$ is an embedded submanifold of $G$) is closed in $G$ (see Warner [175], Chapter 3, Theorem 3.21, page 97). Second, a theorem of Von Neumann and Cartan asserts that a closed subgroup of $\text{GL}(n, \mathbb{R})$ is an embedded submanifold, and thus, a Lie group (see Warner [175], Chapter 3, Theorem 3.42, page 110). Thus, a linear Lie group $G$ is a closed subgroup of $\text{GL}(n, \mathbb{R})$. Recall that this means that for every sequence $(A_n)_{n \geq 1}$ of matrices $A_n \in G$, if this sequence converges to a limit $A \in \text{GL}(n, \mathbb{R})$, then actually $A \in G$.

Since our Lie groups are subgroups (or isomorphic to subgroups) of $\text{GL}(n, \mathbb{R})$ for some suitable $n$, it is easy to define the Lie algebra of a Lie group using curves. This approach to define the Lie algebra of a matrix group is followed by a number of authors, such as Curtis [46]. However, Curtis is rather cavalier, since he does not explain why the required curves actually exist, and thus, according to his definition, Lie algebras could be the trivial vector space! with trivial objects.

A small annoying technical problem will arise in our approach, the problem with discrete subgroups. If $A$ is a subset of $\mathbb{R}^N$, recall that $A$ inherits a topology from $\mathbb{R}^N$ called the subspace topology, and defined such that a subset $V$ of $A$ is open if

$$V = A \cap U$$

for some open subset $U$ of $\mathbb{R}^N$. A point $a \in A$ is said to be isolated if there is some open subset $U$ of $\mathbb{R}^N$ such that

$$\{a\} = A \cap U,$$

in other words, if $\{a\}$ is an open set in $A$.

The group $\text{GL}(n, \mathbb{R})$ of real invertible $n \times n$ matrices can be viewed as a subset of $\mathbb{R}^{n^2}$, and as such, it is a topological space under the subspace topology (in fact, a dense open subset of $\mathbb{R}^{n^2}$). One can easily check that multiplication and the inverse operation are continuous, and in fact smooth (i.e., $C^\infty$-continuously differentiable). This makes $\text{GL}(n, \mathbb{R})$ a topological group. Any subgroup $G$ of $\text{GL}(n, \mathbb{R})$ is also a topological space under the subspace topology. A subgroup $G$ is called a discrete subgroup if it has some isolated point. This turns out to be equivalent to the fact that every point of $G$ is isolated, and thus, $G$ has the discrete topology (every subset of $G$ is open). Now, because $\text{GL}(n, \mathbb{R})$ is a topological group, every discrete subgroup of $\text{GL}(n, \mathbb{R})$ is closed (which means that its complement is open); see Proposition 5.5. Moreover, since $\text{GL}(n, \mathbb{R})$ is the union of countably many compact subsets, discrete subgroups of $\text{GL}(n, \mathbb{R})$ must be countable. Thus, discrete subgroups of $\text{GL}(n, \mathbb{R})$ are Lie groups (and countable)! But these are not very interesting Lie groups, and so we will consider only closed subgroups of $\text{GL}(n, \mathbb{R})$ that are not discrete.

**Definition 4.5.** A Lie group is a nonempty subset $G$ of $\mathbb{R}^N$ ($N \geq 1$) satisfying the following conditions:

(a) $G$ is a group.
(b) $G$ is a manifold in $\mathbb{R}^N$.

(c) The group operation $\cdot : G \times G \to G$ and the inverse map $^{-1} : G \to G$ are smooth.

(Smooth maps are defined in Definition 4.8). It is immediately verified that $\text{GL}(n, \mathbb{R})$ is a Lie group. Since all the Lie groups that we are considering are subgroups of $\text{GL}(n, \mathbb{R})$, the following definition is in order.

**Definition 4.6.** A **linear Lie group** is a subgroup $G$ of $\text{GL}(n, \mathbb{R})$ (for some $n \geq 1$) which is a smooth manifold in $\mathbb{R}^{n^2}$.

Let $M_n(\mathbb{R})$ denote the set of all real $n \times n$ matrices (invertible or not). If we recall that the exponential map $\exp : A \mapsto e^A$ is well defined on $M_n(\mathbb{R})$, we have the following crucial theorem due to Von Neumann and Cartan.

**Theorem 4.8.** (Von Neumann and Cartan, 1927) A nondiscrete closed subgroup $G$ of $\text{GL}(n, \mathbb{R})$ is a linear Lie group. Furthermore, the set $g$ defined such that

$$g = \{ X \in M_n(\mathbb{R}) \mid e^{tX} \in G \text{ for all } t \in \mathbb{R} \}$$

is a nontrivial vector space equal to the tangent space $T_I G$ at the identity $I$, and $g$ is closed under the Lie bracket $[-,-]$ defined such that $[A,B] = AB - BA$ for all $A,B \in M_n(\mathbb{R})$.

Theorem 4.8 applies even when $G$ is a discrete subgroup, but in this case, $g$ is trivial (i.e., $g = \{0\}$). For example, the set of nonnull reals $\mathbb{R}^* = \mathbb{R} - \{0\} = \text{GL}(1, \mathbb{R})$ is a Lie group under multiplication, and the subgroup

$$H = \{ 2^n \mid n \in \mathbb{Z} \}$$

is a discrete subgroup of $\mathbb{R}^*$. Thus, $H$ is a Lie group. On the other hand, the set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ of nonnull rational numbers is a multiplicative subgroup of $\mathbb{R}^*$, but it is not closed, since $\mathbb{Q}$ is dense in $\mathbb{R}$.

The first step in proving Theorem 4.8 is to show that if $G$ is a closed and nondiscrete subgroup of $\text{GL}(n, \mathbb{R})$ and if we define $g$ just as $T_I G$ (even though we don’t know yet that $G$ is a manifold), then $g$ is a vector space satisfying the properties of Theorem 4.8. We follow the treatment in Kosmann [108], which we find one of the simplest and clearest.

**Proposition 4.9.** Given any closed subgroup $G$ in $\text{GL}(n, \mathbb{R})$, the set

$$g = \{ X \in M_n(\mathbb{R}) \mid X = \gamma'(0), \gamma : J \to G \text{ is a } C^1 \text{ curve in } M_n(\mathbb{R}) \text{ such that } \gamma(0) = I \}$$

satisfies the following properties:
(1) \( \mathfrak{g} \) is a vector subspace of \( M_n(\mathbb{R}) \).

(2) For every \( X \in M_n(\mathbb{R}) \), we have \( X \in \mathfrak{g} \) iff \( e^{tX} \in G \) for all \( t \in \mathbb{R} \).

(3) For every \( X \in \mathfrak{g} \) and for every \( g \in G \), we have \( gXg^{-1} \in \mathfrak{g} \).

(4) \( \mathfrak{g} \) is closed under the Lie bracket.

Proof. If \( \gamma \) is a \( C^1 \) curve in \( G \) such that \( \gamma(0) = I \) and \( \gamma'(0) = X \), then for any \( \lambda \in \mathbb{R} \), the curve \( \alpha(t) = \gamma(\lambda t) \) passes through \( I \) and \( \alpha'(0) = \lambda X \). If \( \gamma_1 \) and \( \gamma_2 \) are two \( C^1 \) curves in \( G \) such that \( \gamma_1(0) = \gamma_2(0) = I \), \( \gamma_1'(0) = X \), and \( \gamma_2'(0) = Y \), then the curve \( \alpha(t) = \gamma_1(t)\gamma_2(t) \) passes through \( I \) and the product rule implies

\[
\alpha'(0) = (\gamma_1(t)\gamma_2(t))'(0) = X + Y.
\]

Therefore, \( \mathfrak{g} \) is a vector space.

(2) If \( e^{tX} \in G \) for all \( t \in \mathbb{R} \), then \( \gamma \colon t \mapsto e^{tX} \) is a smooth curve through \( I \) in \( G \) such that \( \gamma'(0) = X \), so \( X \in \mathfrak{g} \).

Conversely, if \( X = \gamma'(0) \) for some \( C^1 \) curve in \( G \) such that \( \gamma(0) = I \), using the Taylor expansion of \( \gamma \) near 0, for every \( t \in \mathbb{R} \) and for any positive integer \( k \) large enough \( t/k \) is small enough so that \( \gamma(t/k) \in G \) and we have

\[
\gamma \left( \frac{t}{k} \right) = I + \frac{t}{k}X + \epsilon_1(k) = \exp \left( \frac{t}{k}X + \epsilon_2(k) \right),
\]

where \( \epsilon_1(k) \) is \( O(1/k^2) \), i.e. \( |\epsilon_1(k)| \leq \frac{C}{k^2} \) for some nonnegative \( C \), and \( \epsilon_2(k) \) is also \( O(1/k^2) \). Raising to the \( k \)th power, we deduce that

\[
\gamma \left( \frac{t}{k} \right)^k = \exp \left( tX + \epsilon_3(k) \right),
\]

where \( \epsilon_3(k) \) is \( O(1/k) \), and by the continuity of the exponential, we get

\[
\lim_{k \to \infty} \gamma \left( \frac{t}{k} \right)^k = \exp(tX).
\]

Now, for all \( k \) large enough, since \( G \) is a closed subgroup, \( (\gamma(t/k))^k \in G \) and

\[
\lim_{k \to \infty} \gamma \left( \frac{t}{k} \right)^k \in G,
\]

and thus \( e^{tX} \in G \).

(3) We know by Lemma 1.2 that

\[
e^{tX}g^{-1} = ge^{tX}g^{-1},
\]
and by (2), if $X \in \mathfrak{g}$, then $e^{tx} \in G$ for all $t$, and since $g \in G$, we have $e^{txg^{-1}} = ge^{tx}g^{-1} \in G$. Now, by a familiar computation,

$$(ge^{tx}g^{-1})'(0) = gX^{-1} \in \mathfrak{g}.$$ 

(4) if $X,Y \in \mathfrak{g}$, then by (2), for all $t \in \mathbb{R}$ we have $e^{tx} \in G$, and by (3), $e^{tx}Ye^{-tx} \in \mathfrak{g}$. Again, by a familiar computation, we obtain

$$(e^{tx}Ye^{-tx})'(0) = XY - YX$$

which proves that $\mathfrak{g}$ is a Lie algebra. \hfill \square

The second step in the proof of Theorem 4.8 is to prove that when $G$ is not a discrete subgroup, there is an open subset $\Omega \subseteq M_{n}(\mathbb{R})$ such that $0 \in \Omega$, an open subset $W \subseteq GL(n, \mathbb{R})$ such that $I \in W$, and a diffeomorphism $\Phi : \Omega \to W$ such that

$$\Phi(\Omega \cap \mathfrak{g}) = W \cap G.$$ 

If $G$ is closed and not discrete, we must have $m \geq 1$, and $\mathfrak{g}$ has dimension $m$.

We begin by observing that the exponential map is a diffeomorphism between some open subset of $0$ and some open subset of $I$. This is because $d(\exp)_0 = \text{id}$, which is easy to see since

$$e^X - I = X + \|X\| \epsilon(X)$$

with

$$\epsilon(X) = \frac{1}{\|X\|} \sum_{k=0}^{\infty} \frac{X^{k+2}}{(k+2)!},$$

and so $\lim_{X \to 0} \epsilon(X) = 0$. By the inverse function theorem, $\exp$ is a diffeomorphism between some open subset $U_0$ of $M_{n}(\mathbb{R})$ containing $0$ and some open subset $V_0$ of $GL(n, \mathbb{R})$ containing $I$.

**Proposition 4.10.** Let $G$ be a subgroup of $GL(n, \mathbb{R})$, and assume that $G$ is closed and not discrete. Then $\dim(\mathfrak{g}) \geq 1$, and the exponential map is a diffeomorphism of a neighborhood of $0$ in $\mathfrak{g}$ onto a neighborhood of $I$ in $G$. Furthermore, there is an open subset $\Omega \subseteq M_{n}(\mathbb{R})$ with $0 \in \Omega$, an open subset $W \subseteq GL(n, \mathbb{R})$ with $I \in W$, and a diffeomorphism $\Phi : \Omega \to W$ such that

$$\Phi(\Omega \cap \mathfrak{g}) = W \cap G.$$ 

**Proof.** We follow the proof in Kosmann [108] (Chapter 4, Section 5). A similar proof is given in Helgason [88] (Chapter 2, §2), Mneimné and Testard [130] (Chapter 3, Section 3.4), and in Duistermaat and Kolk [64] (Chapter 1, Section 10). As explained above, by the inverse function theorem, $\exp$ is a diffeomorphism between some open subset $U_0$ of $M_{n}(\mathbb{R})$ containing
0 and some open subset \( V_0 \) of \( \text{GL}(n, \mathbb{R}) \) containing \( I \). Let \( p \) be any subspace of \( M_n(\mathbb{R}) \) such that \( g \) and \( p \) form a direct sum

\[
M_n(\mathbb{R}) = g \oplus p,
\]

and let \( \Phi : g \oplus p \to G \) be the map defined by

\[
\Phi(X + Y) = e^X e^Y.
\]

We claim that \( d\Phi_0 = \text{id} \). One way to prove this is to observe that for \( \|X\| \) and \( \|Y\| \) small,

\[
e^X = I + X + \|X\| \epsilon_1(X) \quad \quad e^Y = I + Y + \|Y\| \epsilon_2(Y),
\]

with \( \lim_{X \to 0} \epsilon_1(X) = 0 \) and \( \lim_{Y \to 0} \epsilon_2(Y) = 0 \), so we get

\[
e^X e^Y = I + X + Y + XY + \|X\| \epsilon_1(X)(I + Y) + \|Y\| \epsilon_2(Y)(I + X) + \|X\| \|Y\| \epsilon_1(X)\epsilon_2(X)
\]

\[
= I + X + Y + \left( \sqrt{\|X\|^2 + \|Y\|^2} \right) \epsilon(X, Y),
\]

with

\[
\epsilon(X, Y) = \frac{\|X\|}{\sqrt{\|X\|^2 + \|Y\|^2}} \epsilon_1(X)(I + Y) + \frac{\|Y\|}{\sqrt{\|X\|^2 + \|Y\|^2}} \epsilon_2(Y)(I + X) + \frac{XY + \|X\| \|Y\| \epsilon_1(X)\epsilon_2(X)}{\sqrt{\|X\|^2 + \|Y\|^2}}.
\]

Since \( \lim_{X \to 0} \epsilon_1(X) = 0 \) and \( \lim_{Y \to 0} \epsilon_2(Y) = 0 \), the first two terms go to 0 when \( X \) and \( Y \) go to 0, and since

\[
\|XY + \|X\| \|Y\| \epsilon_1(X)\epsilon_2(X)\| \leq \|X\| \|Y\| \left( 1 + \|\epsilon_1(X)\epsilon_2(X)\| \right)
\]

\[
\leq \frac{1}{2} (\|X\|^2 + \|Y\|^2) (1 + \|\epsilon_1(X)\epsilon_2(X)\|),
\]

we have

\[
\left| \frac{XY + \|X\| \|Y\| \epsilon_1(X)\epsilon_2(X)}{\sqrt{\|X\|^2 + \|Y\|^2}} \right| \leq \frac{1}{2} \left( \sqrt{\|X\|^2 + \|Y\|^2} \right) (1 + \|\epsilon_1(X)\epsilon_2(X)\|),
\]

so the third term also goes to 0 when \( X \) and \( Y \) to 0. Therefore, \( \lim_{X \to 0, Y \to 0} \epsilon(X, Y) = 0 \), and \( d\Phi_0(X + Y) = X + Y \), as claimed.

By the inverse function theorem, there exists an open subset of \( M_n(\mathbb{R}) \) containing 0 of the form \( U' + U'' \) with \( U' \subseteq g \) and \( U'' \subseteq p \) and some open subset \( W' \) of \( \text{GL}(n, \mathbb{R}) \) such that \( \Phi \) is a diffeomorphism of \( U' + U'' \) onto \( W' \). By considering \( U_0 \cap (U' + U'') \), we may assume that \( U_0 = U' + U'' \), and write \( W' = \Phi(U_0) \); the maps \( \exp \) and \( \Phi \) are diffeomorphisms on \( U_0 \).
For any sequence \(Z\), we would like to prove that

\[
\text{Proof.}
\]

Lemma 4.11. Let \(G\) be a closed subgroup of \(\text{GL}(n, \mathbb{R})\) and let \(m\) be any subspace of \(M_n(\mathbb{R})\). For any sequence \((X_n)\) of nonzero matrices in \(m\), if \(e^{X_n} \in G\) for all \(n\), if \((X_n)\) converges to \(0\), and if the sequence \((Z_n)\) given by

\[
Z_n = \frac{X_n}{\|X_n\|},
\]

converges to a limit \(Z\) (necessarily in \(m\) and with \(\|Z\| = 1\)), then \(Z \in g\).

Proof. We would like to prove that \(e^{tZ} \in G\) for all \(t \in \mathbb{R}\), because then, by Proposition 4.9(2), \(Z \in g\). For any \(t \in \mathbb{R}\), write

\[
\frac{t}{\|X_n\|} = p_n(t) + u_n(t), \quad \text{with } p_n(t) \in \mathbb{Z} \text{ and } u_n(t) \in [0, 1).
\]
Then, we have
\[ e^{tZ_n} = e^{\left( \frac{t}{\|X_n\|} X_n \right)} = (e^{X_n})^{p_n(t)} e^{u_n(t)X_n}. \]
Since \( u_n(t) \in [0, 1] \) and since the sequence \((X_n)\) converges to 0, the sequence \((u_n(t)X_n)\) also converges to 0, so the sequence \(e^{u_n(t)X_n}\) converges to \(I\). Furthermore, since \(p_n(t)\) is an integer, \(e^{X_n} \in G\), and \(G\) is a group, we have \((e^{X_n})^{p_n(t)} \in G\). Since \(G\) is closed, the limit of the sequence \(e^{tZ_n} = (e^{X_n})^{p_n(t)} e^{u_n(t)X_n}\) belongs to \(G\), and since \(\lim_{n \to \infty} Z_n = Z\), by the continuity of the exponential, we conclude that \(e^{tZ} \in G\). Since this holds for all \(t \in \mathbb{R}\), we have \(Z \in \mathfrak{g}\).

Applying Lemma 4.11 to \(m = p\), we deduce that \(Z \in \mathfrak{g} \cap p = (0)\), so \(Z = 0\), contradicting the fact that \(\|Z\| = 1\). Therefore, the claim holds.

It remains to prove that \(\mathfrak{g}\) is nontrivial. This is where the assumption that \(G\) is not discrete is needed. Indeed, if \(G\) is not discrete, we can find a sequence \((g_n)\) of elements of \(G\) such that \(g_n \neq I\) and the sequence converges to \(I\). Since the exponential is a diffeomorphism between a neighborhood of \(0\) and a neighborhood of \(I\), we may assume by dropping some initial segment of the sequence that \(g_n = e^{X_n}\) for some nonzero matrices \(X_n\), and that the sequence \((X_n)\) converges to 0. For \(n\) large enough, the sequence
\[ Z_n = \frac{X_n}{\|X_n\|} \]
makes sense and belongs to the unit sphere. By compactness of the unit sphere, \((Z_n)\) has some subsequence that converges to some matrix \(Z\) with \(\|Z\| = 1\). The corresponding subsequence of \(X_n\) still consists of nonzero matrices and converges to 0. Now, we can apply Lemma 4.11 to \(m = M_n(\mathbb{R})\) and to the converging subsequences of \((X_n)\) and \((Z_n)\) to conclude that \(Z \in \mathfrak{g}\), with \(Z \neq 0\). This proves that \(\dim(\mathfrak{g}) \geq 1\), and completes the proof of Proposition 4.10.

Remark: The first part of Proposition 4.10 shows that \(\exp\) is a diffeomorphism of an open subset \(U' \subseteq \mathfrak{g}\) containing 0 onto \(W \cap G\), which is condition (1) of Theorem 4.6; that is, the restriction of \(\exp\) to \(U'\) is a parametrization of \(G\).

Theorem 4.8 now follows immediately from Propositions 4.9 and 4.10.

Proof of Theorem 4.8. Proposition 4.9 shows that \(\mathfrak{g} = T_I G\) and that it is a Lie algebra. Proposition 4.10 shows that condition (2) of Theorem 4.6 holds; that is, there is an open subset \(\Omega \subseteq M_n(\mathbb{R})\) with \(0 \in \Omega\), an open subset \(W \subseteq \text{GL}(n, \mathbb{R})\) with \(I \in W\), and a diffeomorphism \(\Phi: \Omega \to W\) such that
\[ \Phi(\Omega \cap \mathfrak{g}) = W \cap G. \]
To prove that this condition holds for every \(g \in G\) besides \(I\) is easy. Indeed, \(L_g: G \to G\) is a diffeomorphism, so \(L_g \circ \Phi: \Omega \to L_g(W)\) is a diffeomorphism such that
\[ (L_g \circ \Phi)(\Omega \cap \mathfrak{g}) = L_g(W) \cap G, \]
4.2. LINEAR LIE GROUPS

which shows that condition (2) of Theorem 4.6 also holds for any $g \in G$, and thus $G$ is a manifold. 

It should be noted that the assumption that $G$ is closed is crucial, as shown by the following example from Tapp [167].

Pick any irrational multiple $\lambda$ of $2\pi$, and define

$$G = \left\{ g_t = \begin{pmatrix} e^{it} & 0 \\ 0 & e^{\lambda t} \end{pmatrix} \mid t \in \mathbb{R} \right\}.$$ 

It is clear that $G$ is a subgroup of $\text{GL}(2, \mathbb{C})$. We leave it as an exercise to prove that the map $\varphi : t \mapsto g_t$ is a continuous isomorphism of $(\mathbb{R}, +)$ onto $G$, but that $\varphi^{-1}$ is not continuous. Geometrically, $\varphi$ is a curve embedded in $\mathbb{R}^4$ (by viewing $\mathbb{C}^2$ as $\mathbb{R}^4$). It is easy to check that $\mathfrak{g}$ (as defined in Proposition 4.9) is the one dimensional vector space spanned by

$$W = \begin{pmatrix} i \\ 0 \\ 0 \\ \lambda i \end{pmatrix},$$

and that $e^{tW} = g_t$ for all $t \in \mathbb{R}$. Now, for every $r > 0$ ($r \in \mathbb{R}$), we leave it as an exercise to prove that

$$\exp(\{tW \mid t \in (-r, r)\}) = \{g_t \mid t \in (-r, r)\}$$

is not a neighborhood of $I$ in $G$. The problem is that there are elements of $G$ of the form $g_{2\pi n}$ for some large $n$ that are arbitrarily close to $I$, so they are exponential images of very short vectors in $\text{M}_2(\mathbb{C})$, but they are exponential images only of very long vectors in $\mathfrak{g}$. The reader should prove that the closure of the group $G$ is the group

$$\overline{G} = \left\{ \begin{pmatrix} e^{it} & 0 \\ 0 & e^{si} \end{pmatrix} \mid t, s \in \mathbb{R} \right\},$$

and that $G$ is dense in $\overline{G}$. Geometrically, $G$ is a curve in $\mathbb{R}^4$ and $\overline{G}$ is the product of two circles, that is, a torus (in $\mathbb{R}^4$). Due to the the irrationality of $\lambda$, the curve $G$ winds around the torus and forms a dense subset.

With the help of Theorem 4.8 it is now very easy to prove that $\text{SL}(n)$, $\text{O}(n)$, $\text{SO}(n)$, $\text{SL}(n, \mathbb{C})$, $\text{U}(n)$, and $\text{SU}(n)$ are Lie groups and to figure out what are their Lie algebras. (Of course, $\text{GL}(n, \mathbb{R})$ is a Lie group, as we already know.) It suffices to show that these subgroups of $\text{GL}(n, \mathbb{R})$ are closed, which is easy to show since these groups are zero sets of simple continuous functions. For example, $\text{SL}(n)$ is the zero set of the function $A \mapsto \det(A) - 1$, $\text{O}(n)$ is the zero set of the function $R \mapsto R^\top R - I$, $\text{SO}(n) = \text{SL}(n) \cap \text{O}(n)$, etc.

For example, if $G = \text{GL}(n, \mathbb{R})$, as $e^{tA}$ is invertible for every matrix $A \in \text{M}_n(\mathbb{R})$, we deduce that the Lie algebra $\mathfrak{gl}(n, \mathbb{R})$ of $\text{GL}(n, \mathbb{R})$ is equal to $\text{M}_n(\mathbb{R})$. We also claim that the Lie algebra $\mathfrak{sl}(n, \mathbb{R})$ of $\text{SL}(n, \mathbb{R})$ is the set of all matrices with zero trace. Indeed, $\mathfrak{sl}(n, \mathbb{R})$ is the subalgebra of $\mathfrak{gl}(n, \mathbb{R})$ consisting of all matrices $X \in \mathfrak{gl}(n, \mathbb{R})$ such that

$$\det(e^{tX}) = 1$$
for all \( t \in \mathbb{R} \), and because \( \det(e^{tX}) = e^{\operatorname{tr}(tX)} \), for \( t = 1 \), we get \( \operatorname{tr}(X) = 0 \), as claimed.

We can also prove that \( \operatorname{SE}(n) \) is a Lie group as follows. Recall that we can view every element of \( \operatorname{SE}(n) \) as a real \((n + 1) \times (n + 1)\) matrix
\[
\begin{pmatrix}
R & U \\
0 & 1
\end{pmatrix}
\]
where \( R \in \operatorname{SO}(n) \) and \( U \in \mathbb{R}^n \). In fact, such matrices belong to \( \operatorname{SL}(n + 1) \). This embedding of \( \operatorname{SE}(n) \) into \( \operatorname{SL}(n + 1) \) is a group homomorphism, since the group operation on \( \operatorname{SE}(n) \) corresponds to multiplication in \( \operatorname{SL}(n + 1) \):
\[
\begin{pmatrix}
RS & RV + U \\
0 & 1
\end{pmatrix} =
\begin{pmatrix}
R & U \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
S & V \\
0 & 1
\end{pmatrix}.
\]
Note that the inverse of \( \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \) is given by
\[
\begin{pmatrix}
R^{-1} & -R^{-1}U \\
0 & 1
\end{pmatrix} = \begin{pmatrix} R^\top & -R^\top U \\ 0 & 1 \end{pmatrix}.
\]

It is easy to show that \( \operatorname{SE}(n) \) is a closed subgroup of \( \operatorname{GL}(n + 1, \mathbb{R}) \) (because \( \operatorname{SO}(n) \) and \( \mathbb{R}^n \) are closed). Also note that the embedding shows that, as a manifold, \( \operatorname{SE}(n) \) is diffeomorphic to \( \operatorname{SO}(n) \times \mathbb{R}^n \) (given a manifold \( M_1 \) of dimension \( m_1 \) and a manifold \( M_2 \) of dimension \( m_2 \), the product \( M_1 \times M_2 \) can be given the structure of a manifold of dimension \( m_1 + m_2 \) in a natural way). Thus, \( \operatorname{SE}(n) \) is a Lie group with underlying manifold \( \operatorname{SO}(n) \times \mathbb{R}^n \), and in fact, a closed subgroup of \( \operatorname{SL}(n + 1) \).

Even though \( \operatorname{SE}(n) \) is diffeomorphic to \( \operatorname{SO}(n) \times \mathbb{R}^n \) as a manifold, it is not isomorphic to \( \operatorname{SO}(n) \times \mathbb{R}^n \) as a group, because the group multiplication on \( \operatorname{SE}(n) \) is not the multiplication on \( \operatorname{SO}(n) \times \mathbb{R}^n \). Instead, \( \operatorname{SE}(n) \) is a semidirect product of \( \operatorname{SO}(n) \) by \( \mathbb{R}^n \); see Section 15.5 or Gallier [72] (Chapter 2, Problem 2.19).

Returning to Theorem 4.8, the vector space \( \mathfrak{g} \) is called the Lie algebra of the Lie group \( G \). Lie algebras are defined as follows.

**Definition 4.7.** A (real) Lie algebra \( \mathcal{A} \) is a real vector space together with a bilinear map \([\cdot, \cdot]: \mathcal{A} \times \mathcal{A} \rightarrow \mathcal{A}\) called the Lie bracket on \( \mathcal{A} \) such that the following two identities hold for all \( a, b, c \in \mathcal{A} \):
\[
[a, a] = 0,
\]
and the so-called Jacobi identity
\[
[a, [b, c]] + [c, [a, b]] + [b, [c, a]] = 0.
\]
By using the Jacobi identity, it is readily verified that \([b, a] = -[a, b]\).
4.2. LINEAR LIE GROUPS

In view of Theorem 4.8, the vector space $\mathfrak{g} = T_I G$ associated with a Lie group $G$ is indeed a Lie algebra. Furthermore, the exponential map $\exp: \mathfrak{g} \rightarrow G$ is well-defined. In general, $\exp$ is neither injective nor surjective, as we observed earlier. Theorem 4.8 also provides a kind of recipe for “computing” the Lie algebra $\mathfrak{g} = T_I G$ of a Lie group $G$. Indeed, $\mathfrak{g}$ is the tangent space to $G$ at $I$, and thus we can use curves to compute tangent vectors. Actually, for every $X \in T_I G$, the map

$$\gamma_X: t \mapsto e^{tx}$$

is a smooth curve in $G$, and it is easily shown that $\gamma'_X(0) = X$. Thus, we can use these curves. As an illustration, we show that the Lie algebras of $\text{SL}(n)$ and $\text{SO}(n)$ are the matrices with null trace and the skew symmetric matrices.

Let $t \mapsto R(t)$ be a smooth curve in $\text{SL}(n)$ such that $R(0) = I$. We have $\det(R(t)) = 1$ for all $t \in (-\epsilon, \epsilon)$. Using the chain rule, we can compute the derivative of the function

$$t \mapsto \det(R(t))$$

at $t = 0$, and since $\det(R(t)) = 1$ we get

$$\det'(R'(0)) = 0.$$ 

We leave it as an exercise for the reader to prove that

$$\det'(X) = \text{tr}(X),$$

and thus $\text{tr}(R'(0)) = 0$, which says that the tangent vector $X = R'(0)$ has null trace. Clearly, $\mathfrak{sl}(n, \mathbb{R})$ has dimension $n^2 - 1$.

Let $t \mapsto R(t)$ be a smooth curve in $\text{SO}(n)$ such that $R(0) = I$. Since each $R(t)$ is orthogonal, we have

$$R(t) R(t)^\top = I$$

for all $t \in (-\epsilon, \epsilon)$. By using the product rule and taking the derivative at $t = 0$, we get

$$R'(0) R(0)^\top + R(0) R'(0)^\top = 0,$$

but since $R(0) = I = R(0)^\top$, we get

$$R'(0) + R'(0)^\top = 0,$$

which says that the tangent vector $X = R'(0)$ is skew symmetric. Since the diagonal elements of a skew symmetric matrix are null, the trace is automatically null, and the condition $\det(R) = 1$ yields nothing new. This shows that $\mathfrak{o}(n) = \mathfrak{so}(n)$. It is easily shown that $\mathfrak{so}(n)$ has dimension $n(n - 1)/2$.

As a concrete example, the Lie algebra $\mathfrak{so}(3)$ of $\text{SO}(3)$ is the real vector space consisting of all $3 \times 3$ real skew symmetric matrices. Every such matrix is of the form

$$\begin{pmatrix} 0 & -b & c \\ b & 0 & -a \\ -c & a & 0 \end{pmatrix}$$

where $a, b, c \in \mathbb{R}$.
where \( b, c, d \in \mathbb{R} \). The Lie bracket \([A, B]\) in \( \mathfrak{so}(3) \) is also given by the usual commutator, \([A, B] = AB - BA\).

Let \( \times \) represent the cross product of two vectors in \( \mathbb{R}^3 \) where for \( u = (u_1, u_2, u_3) \) and \( v = (v_1, v_2, v_3) \), we have

\[
\quad u \times v = -v \times u = (u_2v_3 - u_3v_2, -u_1v_3 + u_3v_1, u_1v_2 - u_2v_1).
\]

It is easily checked that the vector space \( \mathbb{R}^3 \) is a Lie algebra if we define the Lie bracket on \( \mathbb{R}^3 \) as the usual cross product \( u \times v \) of vectors. We can define an isomorphism of Lie algebras \( \psi : (\mathbb{R}^3, \times) \to \mathfrak{so}(3) \) by the formula

\[
\psi(b, c, d) = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.
\]

A basic algebraic computation verifies that

\[
\psi(u \times v) = [\psi(u), \psi(v)].
\]

It is also verified that for any two vectors \( u = (b, c, d) \) and \( v = (b', c', d') \) in \( \mathbb{R}^3 \)

\[
\psi(u)(v) = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix} \begin{pmatrix} b' \\ c' \\ d' \end{pmatrix} = \begin{pmatrix} -dc' + cd' \\ db' - bd' \\ -cb' + bc' \end{pmatrix} = u \times v.
\]

In robotics and in computer vision, \( \psi(u) \) is often denoted by \( u_\times \).

The exponential map \( \exp : \mathfrak{so}(3) \to \text{SO}(3) \) is given by Rodrigues’s formula (see Lemma 1.7):

\[
e^A = \cos \theta I_3 + \frac{\sin \theta}{\theta} A + \frac{1 - \cos \theta}{\theta^2} B,
\]

or equivalently by

\[
e^A = I_3 + \frac{\sin \theta}{\theta} A + \frac{(1 - \cos \theta)}{\theta^2} A^2
\]

if \( \theta \neq 0 \), where

\[
A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix},
\]

\[
\theta = \sqrt{b^2 + c^2 + d^2}, \quad B = A^2 + \theta^2 I_3, \quad \text{and with } e^0 = I_3.
Using the above methods, it is easy to verify that the Lie algebras \( \mathfrak{gl}(n, \mathbb{R}) \), \( \mathfrak{sl}(n, \mathbb{R}) \), \( \mathfrak{o}(n) \), and \( \mathfrak{so}(n) \), are respectively \( M_n(\mathbb{R}) \), the set of matrices with null trace, and the set of skew symmetric matrices (in the last two cases). A similar computation can be done for \( \mathfrak{gl}(n, \mathbb{C}) \), \( \mathfrak{sl}(n, \mathbb{C}) \), \( \mathfrak{u}(n) \), and \( \mathfrak{su}(n) \), confirming the claims of Section 1.4. It is easy to show that \( \mathfrak{gl}(n, \mathbb{C}) \) has dimension \( 2n^2 \), \( \mathfrak{sl}(n, \mathbb{C}) \) has dimension \( 2(n^2 - 1) \), \( \mathfrak{u}(n) \) has dimension \( n^2 \), and \( \mathfrak{su}(n) \) has dimension \( n^2 - 1 \).

For example, the Lie algebra \( \mathfrak{su}(2) \) of \( \text{SU}(2) \) (or \( S^3 \)) is the real vector space consisting of all \( 2 \times 2 \) (complex) skew Hermitian matrices of null trace. Every such matrix is of the form

\[
i(d\sigma_1 + c\sigma_2 + b\sigma_3) = \begin{pmatrix} ib & c+id \\ -c+ id & -ib \end{pmatrix},
\]

where \( b, c, d \in \mathbb{R} \), and \( \sigma_1, \sigma_2, \sigma_3 \) are the Pauli spin matrices

\[
\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},
\]

and thus the matrices \( i\sigma_1, i\sigma_2, i\sigma_3 \) form a basis of the Lie algebra \( \mathfrak{su}(2) \). The Lie bracket \([A, B]\) in \( \mathfrak{su}(2) \) is given by the usual commutator, \([A, B] = AB - BA\).

Let \( \times \) represent the cross product of two vectors in \( \mathbb{R}^3 \). Then we can define an isomorphism of Lie algebras \( \varphi : (\mathbb{R}^3, \times) \to \mathfrak{su}(2) \) by the formula

\[
\varphi(b, c, d) = \frac{i}{2}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{2} \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}.
\]

A tedious but basic algebraic computation verifies that

\[
\varphi(u \times v) = [\varphi(u), \varphi(v)].
\]

Returning to \( \mathfrak{su}(2) \), letting \( \theta = \sqrt{b^2 + c^2 + d^2} \), we can write

\[
d\sigma_1 + c\sigma_2 + b\sigma_3 = \begin{pmatrix} b & -ic + d \\ ic + d & -b \end{pmatrix} = \theta A,
\]

where

\[
A = \frac{1}{\theta}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{\theta} \begin{pmatrix} b & -ic + d \\ ic + d & -b \end{pmatrix},
\]

so that \( A^2 = I \), and it can be shown that the exponential map \( \exp : \mathfrak{su}(2) \to \text{SU}(2) \) is given by

\[
\exp(i\theta A) = \cos \theta I + i \sin \theta A.
\]

In view of the isomorphism \( \varphi : (\mathbb{R}^3, \times) \to \mathfrak{su}(2) \), where

\[
\varphi(b, c, d) = \frac{1}{2} \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix} = \frac{\theta}{2} A,
\]
the exponential map can be viewed as a map \( \exp: (\mathbb{R}^3, \times) \rightarrow SU(2) \) given by the formula

\[
\exp(\theta v) = \begin{bmatrix}
\cos \frac{\theta}{2} & \sin \frac{\theta}{2} \\
0 & 0
\end{bmatrix} v,
\]

for every vector \( \theta v \), where \( v \) is a unit vector in \( \mathbb{R}^3 \) and \( \theta \in \mathbb{R} \). Recall that \([a, (b, c, d)]\) is another way of denoting the quaternion \( a1 + bi + cj + dk \); see Section 30.1 for the definition of the quaternions. In this form, \( \exp(\theta v) \) is a unit quaternion corresponding to a rotation of axis \( v \) and angle \( \theta \).

As we showed, \( SE(n) \) is a Lie group, and its Lie algebra \( \mathfrak{se}(n) \) described in Section 1.6 is easily determined as the subalgebra of \( \mathfrak{sl}(n+1) \) consisting of all matrices of the form

\[
\begin{pmatrix}
B & U \\
0 & 0
\end{pmatrix}
\]

where \( B \in \mathfrak{so}(n) \) and \( U \in \mathbb{R}^n \). Thus, \( \mathfrak{se}(n) \) has dimension \( n(n+1)/2 \). The Lie bracket is given by

\[
\begin{pmatrix}
B & U \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
C & V \\
0 & 0
\end{pmatrix}
- \begin{pmatrix}
C & V \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
B & U \\
0 & 0
\end{pmatrix}
= \begin{pmatrix}
BC-CB & BV-CU \\
0 & 0
\end{pmatrix}.
\]

We conclude by indicating the relationship between homomorphisms of Lie groups and homomorphisms of Lie algebras. First, we need to explain what is meant by a smooth map between manifolds.

**Definition 4.8.** Let \( M_1 \) (\( m_1 \)-dimensional) and \( M_2 \) (\( m_2 \)-dimensional) be manifolds in \( \mathbb{R}^N \). A function \( f: M_1 \rightarrow M_2 \) is smooth if for every \( p \in M_1 \) there are parametrizations \( \varphi: \Omega_1 \rightarrow U_1 \) of \( M_1 \) at \( p \) and \( \psi: \Omega_2 \rightarrow U_2 \) of \( M_2 \) at \( f(p) \) such that \( f(U_1) \subseteq U_2 \) and

\[
\psi^{-1} \circ f \circ \varphi: \Omega_1 \rightarrow \mathbb{R}^{m_2}
\]

is smooth; see Figure 4.12.

Using Lemma 4.2, it is easily shown that Definition 4.8 does not depend on the choice of the parametrizations \( \varphi: \Omega_1 \rightarrow U_1 \) and \( \psi: \Omega_2 \rightarrow U_2 \). A smooth map \( f \) between manifolds is a smooth diffeomorphism if \( f \) is bijective and both \( f \) and \( f^{-1} \) are smooth maps.

We now define the derivative of a smooth map between manifolds.

**Definition 4.9.** Let \( M_1 \) (\( m_1 \)-dimensional) and \( M_2 \) (\( m_2 \)-dimensional) be manifolds in \( \mathbb{R}^N \). For any smooth function \( f: M_1 \rightarrow M_2 \) and any \( p \in M_1 \), the function \( f'_p: T_pM_1 \rightarrow T_{f(p)}M_2 \), called the tangent map of \( f \) at \( p \), or derivative of \( f \) at \( p \), or differential of \( f \) at \( p \), is defined as follows: For every \( v \in T_pM_1 \) and every smooth curve \( \gamma: I \rightarrow M_1 \) such that \( \gamma(0) = p \) and \( \gamma'(0) = v \),

\[
f'_p(v) = (f \circ \gamma)'(0).
\]

See Figure 4.13.
4.2. LINEAR LIE GROUPS

Figure 4.12: An illustration of a smooth map from the torus, $M_1$, to the solid ellipsoid $M_2$. The pink patch on $M_1$ is mapped into interior pink ellipsoid of $M_2$.

Figure 4.13: An illustration of the tangent map from $T_p M_1$ to $T_{f(p)} M_2$
The map $f'_p$ is also denoted by $df_p$ or $T_pf$. Doing a few calculations involving the facts that
\[ f \circ \gamma = (f \circ \varphi) \circ (\varphi^{-1} \circ \gamma) \quad \text{and} \quad \gamma = \varphi \circ (\varphi^{-1} \circ \gamma) \]
and using Lemma 4.2, it is not hard to show that $f'_p(v)$ does not depend on the choice of the curve $\gamma$. It is easily shown that $f'_p$ is a linear map.

Given a linear Lie group $G$, since $L_a$ and $R_a$ are diffeomorphisms for every $a \in G$, the maps $d(L_a)_b : g \to T_aG$ and $d(R_a)_b : g \to T_aG$ are linear isomorphisms between the Lie algebra $g$ and the tangent space $T_aG$ to $G$ at $a$. Since $G$ is a linear group, both $L_a$ and $R_a$ are linear, we have $(dL_a)_b = L_a$ and $(dR_a)_b = R_a$ for all $b \in G$, and so
\[ T_aG = a g = \{aX \mid X \in g\} = \{Xa \mid X \in g\} = ga. \]

Finally, we define homomorphisms of Lie groups and Lie algebras and see how they are related.

**Definition 4.10.** Given two Lie groups $G_1$ and $G_2$, a homomorphism (or map) of Lie groups is a function $f : G_1 \to G_2$ that is a homomorphism of groups and a smooth map (between the manifolds $G_1$ and $G_2$). Given two Lie algebras $A_1$ and $A_2$, a homomorphism (or map) of Lie algebras is a function $f : A_1 \to A_2$ that is a linear map between the vector spaces $A_1$ and $A_2$ and that preserves Lie brackets, i.e.,
\[ f([A,B]) = [f(A),f(B)] \]
for all $A,B \in A_1$.

An isomorphism of Lie groups is a bijective function $f$ such that both $f$ and $f^{-1}$ are homomorphisms of Lie groups, and an isomorphism of Lie algebras is a bijective function $f$ such that both $f$ and $f^{-1}$ are maps of Lie algebras. If $f : G_1 \to G_2$ is a homomorphism of Lie groups, then $f'_1 : g_1 \to g_2$ is a homomorphism of Lie algebras, but in order to prove this, we need the adjoint representation $Ad$, so we postpone the proof.

The notion of a one-parameter group plays a crucial role in Lie group theory.

**Definition 4.11.** A smooth homomorphism $h : (\mathbb{R},+) \to G$ from the additive group $\mathbb{R}$ to a Lie group $G$ is called a one-parameter group in $G$.

All one-parameter groups of a linear Lie group can be determined explicitly.

**Proposition 4.12.** Let $G$ be any linear Lie group.

1. For every $X \in g$, the map $h(t) = e^{tX}$ is a one-parameter group in $G$.
2. Every one-parameter group $h : \mathbb{R} \to G$ is of the form $h(t) = e^{tZ}$, with $Z = h'(0)$.

In summary, for every $Z \in g$, there is a unique one-parameter group $h$ such that $h'(0) = Z$ given by $h(t) = e^{2t}$. 


4.2. LINEAR LIE GROUPS

Proof. The proof of (1) is easy and left as an exercise. To prove (2), since \( h \) is a homomorphism, for all \( s, t \in \mathbb{R} \), we have

\[
h(s + t) = h(s)h(t).
\]

Taking the derivative with respect to \( s \) for \( s = 0 \) and holding \( t \) constant, the product rule implies that

\[
h'(t) = h'(0)h(t).
\]

If we write \( Z = h'(0) \) we we have

\[
h'(t) = Zh(t) = X_Z(h(t)) \quad \text{for all } t \in \mathbb{R}.
\]

This means that \( h(t) \) is an integral curve for all \( t \) passing through \( I \) for the linear vector field \( X_Z \), and by Proposition 2.25, it must be equal to \( e^{tZ} \).

The exponential map is natural in the following sense:

**Proposition 4.13.** Given any two linear Lie groups \( G \) and \( H \), for every Lie group homomorphism \( f: G \to H \), the following diagram commutes:

\[
\begin{array}{ccc}
G & \xrightarrow{f} & H \\
\exp \downarrow & & \exp \downarrow \\
\mathfrak{g} & \xrightarrow{df_I} & \mathfrak{h}
\end{array}
\]

Proof. Observe that for every \( v \in \mathfrak{g} \), the map \( h: t \mapsto f(e^{tv}) \) is a homomorphism from \((\mathbb{R}, +)\) to \( G \) such that \( h'(0) = df_I(v) \). On the other hand, by Proposition 4.12 the map \( t \mapsto e^{tdf_I(v)} \) is the unique one-parameter group whose tangent vector at 0 is \( df_I(v) \), so \( f(e^v) = e^{df_I(v)} \).

Alert readers must have noticed that in Theorem 4.8 we only defined the Lie algebra of a linear group. In the more general case, we can still define the Lie algebra \( \mathfrak{g} \) of a Lie group \( G \) as the tangent space \( T_I G \) at the identity \( I \). The tangent space \( \mathfrak{g} = T_I G \) is a vector space, but we need to define the Lie bracket. This can be done in several ways. We explain briefly how this can be done in terms of so-called adjoint representations. This has the advantage of not requiring the definition of left-invariant vector fields, but it is still a little bizarre!

Given a Lie group \( G \), for every \( a \in G \) we define left translation as the map \( L_a: G \to G \) such that \( L_a(b) = ab \) for all \( b \in G \), and right translation as the map \( R_a: G \to G \) such that \( R_a(b) = ba \) for all \( b \in G \). The maps \( L_a \) and \( R_a \) are diffeomorphisms, and their derivatives play an important role.

The inner automorphisms \( \text{Ad}_a : G \to G \) defined by \( \text{Ad}_a = R_{a^{-1}} \circ L_a = R_{a^{-1}}L_a \) also play an important role. Note that

\[
\text{Ad}_a(b) = aba^{-1}.
\]
The derivative

\[(\text{Ad}_a)'_I : T_I G \to T_I G\]

of \(\text{Ad}_a\) at \(I\) is an isomorphism of Lie algebras, and since \(T_I G = \mathfrak{g}\), if we denote \((\text{Ad}_a)'_I\) by \(\text{Ad}_a\), we get a map

\[\text{Ad}_a : \mathfrak{g} \to \mathfrak{g}.\]

The map \(a \mapsto \text{Ad}_a\) is a map of Lie groups

\[\text{Ad} : G \to \text{GL}(\mathfrak{g}),\]

called the *adjoint representation of* \(G\) (where \(\text{GL}(\mathfrak{g})\) denotes the Lie group of all bijective linear maps on \(\mathfrak{g}\)).

In the case of a linear group, we have

\[\text{Ad}(a)(X) = aXa^{-1}\]

for all \(a \in G\) and all \(X \in \mathfrak{g}\). Indeed, for any \(X \in \mathfrak{g}\), the curve \(\gamma(t) = e^{tX}\) is a curve in \(G\) such that \(\gamma(0) = I\) and \(\gamma'(0) = X\). Then, by the definition of the tangent map, we have

\[d(\text{Ad}_a)_I(X) = (\text{Ad}_a(\gamma(t)))'(0)\]
\[= (ae^{tX}a^{-1})'(0)\]
\[= aXa^{-1}.\]

We are now almost ready to prove that if \(f : G_1 \to G_2\) is a homomorphism of linear Lie groups, then \(f'_I : \mathfrak{g}_1 \to \mathfrak{g}_2\) is a homomorphism of Lie algebras. What we need is to express the Lie bracket \([A, B]\) in terms of the derivative of an expression involving the adjoint representation \(\text{Ad}\). For any \(A, B \in \mathfrak{g}\), we have

\[(\text{Ad}_{e^{tA}}(B))'(0) = (e^{tA}Be^{-tA})'(0) = AB - BA = [A, B].\]

**Proposition 4.14.** If \(f : G_1 \to G_2\) is a homomorphism of linear Lie groups, then the linear map \(df_I : \mathfrak{g}_1 \to \mathfrak{g}_2\) satisfies the equation

\[df_I(\text{Ad}_a(X)) = \text{Ad}_{f(a)}(df_I(X)),\]

for all \(a \in G\) and all \(X \in \mathfrak{g}_1\),

*that is, the following diagram commutes*

\[
\begin{array}{cc}
\mathfrak{g}_1 & \mathfrak{g}_2 \\
\downarrow \text{Ad}_a & \downarrow \text{Ad}_{f(a)} \\
\mathfrak{g}_1 & \mathfrak{g}_2 \\
\downarrow df_I & \\
\end{array}
\]

Furthermore, \(df_I\) is a homomorphism of Lie algebras.
4.2. LINEAR LIE GROUPS

Proof. Since \( f \) is a group homomorphism, for all \( X \in g_1 \), we have
\[
f(ae^{tX}a^{-1}) = f(a)f(e^{tX})f(a^{-1}) = f(a)f(e^{tX})f(a)^{-1}.
\]
The curve \( \alpha \) given by \( \alpha(t) = ae^{tX}a^{-1} \) passes through \( I \) and \( \alpha'(0) = aXa^{-1} = \text{Ad}_a(X) \), so we have
\[
df_I(\text{Ad}_a(X)) = (f(\alpha(t)))'(0)
\]
\[
= (f(ae^{tX}a^{-1}))'(0)
\]
\[
= (f(a)f(e^{tX})f(a)^{-1})'(0)
\]
\[
= \text{Ad}_{f(a)}(df_I(X)),
\]
as claimed. Now, pick any \( X, Y \in g_1 \). The plan is to use the identity we just proved with \( a = e^{tX} \) and \( X = Y \), namely
\[
df_I(\text{Ad}_{e^{tX}}(Y)) = \text{Ad}_{f(e^{tX})}(df_I(Y)),
\]
and to take the derivative of both sides for \( t = 0 \). We make use of the fact that since \( df_I : g \to g \) is linear, for any \( Z \in g_1 \), we have
\[
d(df_I)_Z = df_I.
\]
Then, if we write \( \beta(t) = \text{Ad}_{e^{tX}}Y \), we have \( df_I(\text{Ad}_{e^{tX}}Y) = df_I(\beta(t)) \), and as \( df_I \) is linear, the derivative of the left hand side of (*) is
\[
(df_I(\beta(0)))' = d(df_I)_{\beta(0)}(\beta'(0)) = df_I(\beta'(0)).
\]
On the other hand, by the fact proven just before stating Proposition 4.14,
\[
\beta'(0) = (\text{Ad}_{e^{tX}}Y)'(0) = [X,Y],
\]
so the the derivative of the left hand side of (*) is equal to \( df_I(\beta'(0)) = df_I([X,Y]) \). When we take the derivative of the right hand side, since \( f \) is a group homomorphism, we get
\[
(\text{Ad}_{f(e^{tX})}(df_I(Y)))' = (f(e^{tX})df_I(Y)(f(e^{tX}))^{-1})'(0)
\]
\[
= (f(e^{tX})df_I(Y)f(e^{-tX}))'(0) = [df_I(X), df_I(Y)],
\]
and we conclude that
\[
df_I([X,Y]) = [df_I(X), df_I(Y)];
\]
that is, \( f_I \) is a Lie algebra homomorphism. \( \square \)

If some additional assumptions are made about \( G_1 \) and \( G_2 \) (for example, connected, simply connected), it can be shown that \( f \) is pretty much determined by \( f'_I \).
The derivative
\[ \text{Ad}_I' : \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}) \]
of \( \text{Ad} : G \to \text{GL}(\mathfrak{g}) \) at \( I \) is map of Lie algebras, and if we denote \( \text{Ad}_I' \) by \( \text{ad} \), it is a map
\[ \text{ad} : \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}), \]
called the adjoint representation of \( \mathfrak{g} \). (Recall that Theorem 4.8 immediately implies that the Lie algebra \( \mathfrak{gl}(\mathfrak{g}) \) of \( \text{GL}(\mathfrak{g}) \) is the vector space \( \text{Hom}(\mathfrak{g}, \mathfrak{g}) \) of all linear maps on \( \mathfrak{g} \)).

In the case of linear Lie groups, if we apply Proposition 4.13 to \( \text{Ad} : G \to \text{GL}(\mathfrak{g}) \), we obtain the equation
\[ \text{Ad}_{e^{tA}} = e^{\text{ad}_A} \quad \text{for all} \quad A \in \mathfrak{g}, \]
or equivalently
\[ \begin{array}{ccc}
G & \xrightarrow{\text{Ad}} & \text{GL}(\mathfrak{g}) \\
\exp & & \exp \\
\mathfrak{g} & \xrightarrow{\text{ad}} & \mathfrak{gl}(\mathfrak{g})
\end{array} \]
which is a generalization of the identity of Proposition 2.26.

In the case of a linear group, we have
\[ \text{ad}(A)(B) = [A, B] \]
for all \( A, B \in \mathfrak{g} \). This can be shown as follows.

Proof. For any \( A, B \in \mathfrak{g} \), the curve \( \gamma(t) = e^{tA} \) is a curve in \( G \) passing through \( I \) and such that \( \gamma'(0) = A \), so we have
\[
\text{ad}_A(B) = ((\text{Ad}_{e^{tA}})'(0))(B) \\
= ((\text{Ad}_{e^{tA}})(B))'(0) \\
= (e^{tA}Be^{-tA})'(0) \\
= AB - BA,
\]
which proves our result. \( \square \)

Remark: The equation
\[ ((\text{Ad}_{e^{tA}})'(0))(B) = ((\text{Ad}_{e^{tA}})(B))'(0) \]
requires some justification. Define \( \text{eval}_B : \text{Hom}(\mathfrak{g}, \mathfrak{g}) \to \mathfrak{g} \) by \( \text{eval}_B(f) = f(B) \) for any \( f \in \text{Hom}(\mathfrak{g}, \mathfrak{g}) \). Note that \( \text{eval}_B \) is a linear map, and hence \( d(\text{eval})_B)_f = \text{eval}_B \) for all \( f \in \mathfrak{g} \).
4.2. LINEAR LIE GROUPS

Hom(\mathfrak{g}, \mathfrak{g})$. By definition $\text{Ad}_{e^{tA}}(B) = \text{eval}_B(\text{Ad}_{e^{tA}})$, and an application of the chain rule implies that

$$(\text{Ad}_{e^{tA}}(B))'(0) = (\text{eval}_B(\text{Ad}_{e^{tA}}))'(0) = d(\text{eval}_B) \circ (\text{Ad}_{e^{tA}})'(0) = \text{eval}_B(\text{Ad}_{e^{tA}})'(0) = ((\text{Ad}_{e^{tA}})'(0))(B).$$

Another proof of the fact that $\text{ad}_A(B) = [A,B]$ can be given using Propositions 2.26 and 4.13. To avoid confusion, let us temporarily write $\text{ad}_A(B) = [A,B]$ to distinguish it from $\text{ad}_A(B) = (d(\text{Ad})_t(A))(B)$. Both $\text{ad}$ and $\text{ad}$ are linear. For any fixed $t \in \mathbb{R}$, by Proposition 2.26 we have

$$\text{Ad}_{e^{tA}} = e^{\text{ad}_A} = e^{t\text{ad}_A},$$

and By Proposition 4.13 applied to $\text{Ad}$, we have

$$\text{Ad}_{e^{tA}} = e^{\text{ad}_A} = e^{t\text{ad}_A}.$$

It follows that

$$e^{t\text{ad}_A} = e^{t\text{ad}_A} \text{ for all } t \in \mathbb{R},$$

and by taking the derivative at $t = 0$, we get $\text{ad}_A = \text{ad}_A$.

One can also check that the Jacobi identity on $\mathfrak{g}$ is equivalent to the fact that $\text{ad}$ preserves Lie brackets, i.e., $\text{ad}$ is a map of Lie algebras:

$$\text{ad}([A, B]) = [\text{ad}(A), \text{ad}(B)]$$

for all $A, B \in \mathfrak{g}$ (where on the right, the Lie bracket is the commutator of linear maps on $\mathfrak{g}$). Thus, we recover the Lie bracket from $\text{ad}$.

This is the key to the definition of the Lie bracket in the case of a general Lie group (not just a linear Lie group). We define the Lie bracket on $\mathfrak{g}$ as

$$[A, B] = \text{ad}(A)(B).$$

To be complete, we have to define the exponential map $\exp: \mathfrak{g} \to G$ for a general Lie group. For this we need to introduce some left-invariant vector fields induced by the derivatives of the left translations, and integral curves associated with such vector fields. We will do this in Chapter 15 but for this we will need a deeper study of manifolds (see Chapters 7 and 8).

We conclude this section by computing explicitly the adjoint representations $\text{ad}$ of $\mathfrak{so}(3)$ and $\text{Ad}$ of $\text{SO}(3)$. Recall that for every $X \in \mathfrak{so}(3)$, $\text{ad}_X$ is a linear map $\text{ad}_X: \mathfrak{so}(3) \to \mathfrak{so}(3)$. Also, for every $R \in \text{SO}(3)$, the map $\text{Ad}_R: \mathfrak{so}(3) \to \mathfrak{so}(3)$ is an invertible linear map of $\mathfrak{so}(3)$. Now, as we saw earlier, $\mathfrak{so}(3)$ is isomorphic to $(\mathbb{R}^3, \times)$, where $\times$ is the cross-product on $\mathbb{R}^3$, via the isomorphism $\psi: (\mathbb{R}^3, \times) \to \mathfrak{so}(3)$ given by the formula

$$\psi(a, b, c) = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}.$$
In robotics and in computer vision, $\psi(u)$ is often denoted by $u_x$. Recall that
\[ \psi(u)v = u_xv = u \times v \quad \text{for all } u, v \in \mathbb{R}^3. \]
The image of the canonical basis $(e_1, e_2, e_3)$ of $\mathbb{R}^3$ is the following basis of $\mathfrak{so}(3)$:
\[
\begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & -1 \\
0 & 1 & 0
\end{pmatrix},
\begin{pmatrix}
0 & 0 & 1 \\
0 & 0 & 0 \\
-1 & 0 & 0
\end{pmatrix},
\begin{pmatrix}
0 & -1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 0
\end{pmatrix}.
\]
Observe that
\[ [E_1, E_2] = E_3, \quad [E_2, E_3] = E_1, \quad [E_3, E_1] = E_2. \]
Using the isomorphism $\psi$, we obtain an isomorphism $\Psi$ between $\text{Hom}(\mathfrak{so}(3), \mathfrak{so}(3))$ and $M_3(\mathbb{R}) = \text{gl}(3, \mathbb{R})$ such that every linear map $f : \mathfrak{so}(3) \to \mathfrak{so}(3)$ corresponds to the matrix of the linear map $\Psi(f) = \psi^{-1} \circ f \circ \psi$ in the basis $(e_1, e_2, e_3)$. By restricting $\Psi$ to $\text{GL}(\mathfrak{so}(3))$, we obtain an isomorphism between $\text{GL}(\mathfrak{so}(3))$ and $\text{GL}(3, \mathbb{R})$. It turns out that if we use the basis $(E_1, E_2, E_3)$ in $\mathfrak{so}(3)$, for every $X \in \mathfrak{so}(3)$, the matrix representing $\text{ad}_X \in \text{Hom}(\mathfrak{so}(3), \mathfrak{so}(3))$ is $X$ itself, and for every $R \in \text{SO}(3)$, the matrix representing $\text{Ad}_R \in \text{GL}(\mathfrak{so}(3))$ is $R$ itself.

**Proposition 4.15.** For all $X \in \mathfrak{so}(3)$ and all $R \in \text{SO}(3)$, we have
\[ \Psi(\text{ad}_X) = X, \quad \Psi(\text{Ad}_R) = R, \]
which means that $\Psi \circ \text{ad}$ is the inclusion map from $\mathfrak{so}(3) \to M_3(\mathbb{R}) = \text{gl}(3, \mathbb{R})$, and that $\Psi \circ \text{Ad}$ is the inclusion map from $\text{SO}(3)$ to $\text{GL}(3, \mathbb{R})$. Equivalently, for all $u \in \mathbb{R}^3$, we have
\[ \text{ad}_X(\psi(u)) = \psi(Xu), \quad \text{Ad}_R(\psi(u)) = \psi(Ru). \]
These equations can also be written as
\[ [X, u_x] = (Xu)_x, \quad Ru_xR^{-1} = (Ru)_x. \]

**Proof.** Since $\text{ad}$ is linear, it suffices to prove the equation for the basis $(E_1, E_2, E_3)$. For $E_1$, since $\psi(e_i) = E_i$, we have
\[ \text{ad}_{E_1}(\psi(e_i)) = [E_1, \psi(e_i)] = \begin{cases} 0 & \text{if } i = 1 \\ E_3 & \text{if } i = 2 \\ -E_2 & \text{if } i = 3. \end{cases} \]
Since
\[ E_1e_1 = 0, \quad E_1e_2 = e_3, \quad E_1e_3 = -e_2, \quad \psi(0) = 0, \quad \psi(e_3) = E_3, \quad \psi(e_2) = E_2, \]
we proved that
\[ \text{ad}_{E_1}(\psi(e_i)) = \psi(E_1 e_i), \quad i = 1, 2, 3. \]

Similarly, the reader should check that
\[ \text{ad}_{E_j}(\psi(e_i)) = \psi(E_j e_i), \quad j = 2, 3, \quad i = 1, 2, 3, \]

and so,
\[ \text{ad}_X(\psi(u)) = \psi(Xu) \quad \text{for all } X \in \mathfrak{so}(3) \text{ and all } u \in \mathbb{R}^3, \]

or equivalently
\[ \psi^{-1}(\text{ad}_X(\psi(u))) = X(u) \quad \text{for all } X \in \mathfrak{so}(3) \text{ and all } u \in \mathbb{R}^3; \]

that is, \( \Psi \circ \text{ad} \) is the inclusion map from \( \mathfrak{so}(3) \) to \( M_3(\mathbb{R}) = \mathfrak{gl}(3, \mathbb{R}) \).

Since every one-parameter group in \( \text{SO}(3) \) is of the form \( t \mapsto e^{tX} \) for some \( X \in \mathfrak{so}(3) \) and since \( \Psi \circ \text{ad} \) is the inclusion map from \( \mathfrak{so}(3) \) to \( M_3(\mathbb{R}) = \mathfrak{gl}(3, \mathbb{R}) \), the map \( \Psi \circ \text{Ad} \) maps every one-parameter group in \( \text{SO}(3) \) to itself in \( \text{GL}(3, \mathbb{R}) \). Since the exponential map \( \exp: \mathfrak{so}(3) \to \text{SO}(3) \) is surjective, every \( R \in \text{SO}(3) \) is of the form \( R = e^X \) for some \( X \in \mathfrak{so}(3) \), so \( R \) is contained in some one-parameter group, and thus \( R \) is mapped to itself by \( \Psi \circ \text{Ad} \).

Readers who wish to learn more about Lie groups and Lie algebras should consult (more or less listed in order of difficulty) Tapp [167], Rossmann [146], Kosmann [108], Curtis [46], Sattinger and Weaver [154], Hall [84], and Marsden and Ratiu [121]. The excellent lecture notes by Carter, Segal, and Macdonald [38] constitute a very efficient (although somewhat terse) introduction to Lie algebras and Lie groups. Classics such as Weyl [179] and Chevalley [41] are definitely worth consulting, although the presentation and the terminology may seem a bit old fashioned. For more advanced texts, one may consult Abraham and Marsden [1], Warner [175], Sternberg [166], Bröcker and tom Dieck [31], and Knapp [106]. For those who read French, Mneimné and Testard [130] is very clear and quite thorough, and uses very little differential geometry, although it is more advanced than Curtis. Chapter 1, by Bryant, in Freed and Uhlenbeck [32] is also worth reading, but the pace is fast.
Chapter 5

Groups and Group Actions

This chapter provides the foundations for deriving a class of manifolds known as homogeneous spaces. It begins with a short review of group theory, introduces the concept of a group acting on a set, and defines the Grassmanians and Stiefel manifolds as homogenous manifolds arising from group actions of Lie groups. The last section provides an overview of topological groups, of which Lie groups are a special example, and contains more advanced material that may be skipped upon first reading.

5.1 Basic Concepts of Groups

We begin with a brief review of the group theory necessary for understanding the concept of a group acting on a set. Readers familiar with this material may proceed to the next section.

Definition 5.1. A group is a set $G$ equipped with a binary operation $\cdot \colon G \times G \to G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in $G$ is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. (associativity)

(G2) $a \cdot e = e \cdot a = a$. (identity)

(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. (inverse)

A group $G$ is abelian (or commutative) if

\[ a \cdot b = b \cdot a \quad \text{for all} \ a, b \in G. \]

A set $M$ together with an operation $\cdot \colon M \times M \to M$ and an element $e$ satisfying only conditions (G1) and (G2) is called a monoid. For example, the set $\mathbb{N} = \{0, 1, \ldots, n, \ldots\}$ of natural numbers is a (commutative) monoid under addition. However, it is not a group.

Some examples of groups are given below.
Example 5.1.

1. The set \( Z = \{ \ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots \} \) of integers is an abelian group under addition, with identity element 0. However, \( Z^* = Z - \{ 0 \} \) is not a group under multiplication, but rather a commutative monoid.

2. The set \( Q \) of rational numbers (fractions \( p/q \) with \( p, q \in Z \) and \( q \neq 0 \)) is an abelian group under addition, with identity element 0. The set \( Q^* = Q - \{ 0 \} \) is also an abelian group under multiplication, with identity element 1.

3. Similarly, the sets \( R \) of real numbers and \( C \) of complex numbers are abelian groups under addition (with identity element 0), and \( R^* = R - \{ 0 \} \) and \( C^* = C - \{ 0 \} \) are abelian groups under multiplication (with identity element 1).

4. The sets \( R^n \) and \( C^n \) of \( n \)-tuples of real or complex numbers are groups under componentwise addition:

\[
(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),
\]

with identity element \((0, \ldots, 0)\). All these groups are abelian.

5. Given any nonempty set \( S \), the set of bijections \( f: S \to S \), also called permutations of \( S \), is a group under function composition (i.e., the multiplication of \( f \) and \( g \) is the composition \( g \circ f \)), with identity element the identity function \( \text{id}_S \). This group is not abelian as soon as \( S \) has more than two elements.

6. The set of \( n \times n \) matrices with real (or complex) coefficients is an abelian group under addition of matrices, with identity element the null matrix. It is denoted by \( M_n(\mathbb{R}) \) (or \( M_n(\mathbb{C}) \)).

7. The set \( \mathbb{R}[X] \) of all polynomials in one variable with real coefficients is an abelian group under addition of polynomials.

8. The set of \( n \times n \) invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix \( I_n \). This group is called the general linear group and is usually denoted by \( \text{GL}(n, \mathbb{R}) \) (or \( \text{GL}(n, \mathbb{C}) \)).

9. The set of \( n \times n \) invertible matrices with real (or complex) coefficients and determinant \(+1\) is a group under matrix multiplication, with identity element the identity matrix \( I_n \). This group is called the special linear group and is usually denoted by \( \text{SL}(n, \mathbb{R}) \) (or \( \text{SL}(n, \mathbb{C}) \)).

10. The set of \( n \times n \) invertible matrices with real coefficients such that \( RR^T = I_n \) and of determinant \(+1\) is a group called the orthogonal group and is usually denoted by \( \text{SO}(n) \) (where \( R^T \) is the transpose of the matrix \( R \), i.e., the rows of \( R^T \) are the columns of \( R \)). It corresponds to the rotations in \( \mathbb{R}^n \).
11. Given an open interval \((a, b)\), the set \(C((a, b))\) of continuous functions \(f: (a, b) \rightarrow \mathbb{R}\) is an abelian group under the operation \(f + g\) defined such that

\[
(f + g)(x) = f(x) + g(x)
\]

for all \(x \in (a, b)\).

It is customary to denote the operation of an abelian group \(G\) by \(+\), in which case the inverse \(a^{-1}\) of an element \(a \in G\) is denoted by \(-a\).

The identity element of a group is unique. In fact, we can prove a more general fact:

**Fact 1.** If a binary operation \(\cdot: M \times M \rightarrow M\) is associative and if \(e' \in M\) is a left identity and \(e'' \in M\) is a right identity, which means that

\[
e' \cdot a = a \quad \text{for all} \quad a \in M \quad \text{(G2l)}
\]

and

\[
a \cdot e'' = a \quad \text{for all} \quad a \in M, \quad \text{(G2r)}
\]

then \(e' = e''\).

**Proof.** If we let \(a = e''\) in equation (G2l), we get

\[e' \cdot e'' = e'',\]

and if we let \(a = e'\) in equation (G2r), we get

\[e' \cdot e'' = e',\]

and thus

\[e' = e' \cdot e'' = e'',\]

as claimed. \(\square\)

Fact 1 implies that the identity element of a monoid is unique, and since every group is a monoid, the identity element of a group is unique. Furthermore, every element in a group has a unique inverse. This is a consequence of a slightly more general fact:

**Fact 2.** In a monoid \(M\) with identity element \(e\), if some element \(a \in M\) has some left inverse \(a' \in M\) and some right inverse \(a'' \in M\), which means that

\[
a' \cdot a = e \quad \text{(G3l)}
\]

and

\[
a \cdot a'' = e, \quad \text{(G3r)}
\]

then \(a' = a''\).
Proof. Using (G3l) and the fact that \( e \) is an identity element, we have
\[(a' \cdot a) \cdot a'' = e \cdot a'' = a''\].
Similarly, Using (G3r) and the fact that \( e \) is an identity element, we have
\[a' \cdot (a \cdot a'') = a' \cdot e = a'.\]
However, since \( M \) is monoid, the operation \( \cdot \) is associative, so
\[a' = a' \cdot (a \cdot a'') = (a' \cdot a) \cdot a'' = a'',\]
as claimed. \(\square\)

Remark: Axioms (G2) and (G3) can be weakened a bit by requiring only (G2r) (the existence of a right identity) and (G3r) (the existence of a right inverse for every element) (or (G2l) and (G3l)). It is a good exercise to prove that the group axioms (G2) and (G3) follow from (G2r) and (G3r).

Given a group \( G \), for any two subsets \( R, S \subseteq G \), we let
\[RS = \{r \cdot s \mid r \in R, s \in S\}\].
In particular, for any \( g \in G \), if \( R = \{g\} \), we write
\[gS = \{g \cdot s \mid s \in S\},\]
and similarly, if \( S = \{g\} \), we write
\[Rg = \{r \cdot g \mid r \in R\}.
From now on, we will drop the multiplication sign and write \( g_1g_2 \) for \( g_1 \cdot g_2 \).

Definition 5.2. Given a group \( G \), a subset \( H \) of \( G \) is a subgroup of \( G \) iff
\begin{enumerate}
  \item The identity element \( e \) of \( G \) also belongs to \( H \) (\( e \in H \));
  \item For all \( h_1, h_2 \in H \), we have \( h_1h_2 \in H \);
  \item For all \( h \in H \), we have \( h^{-1} \in H \).
\end{enumerate}
It is easily checked that a subset \( H \subseteq G \) is a subgroup of \( G \) iff \( H \) is nonempty and whenever \( h_1, h_2 \in H \), then \( h_1h_2^{-1} \in H \).

Definition 5.3. If \( H \) is a subgroup of \( G \) and \( g \in G \) is any element, the sets of the form \( gH \) are called left cosets of \( H \) in \( G \) and the sets of the form \( Hg \) are called right cosets of \( H \) in \( G \).
The left cosets (resp. right cosets) of $H$ induce an equivalence relation $\sim$ defined as follows: For all $g_1, g_2 \in G$,

$$g_1 \sim g_2 \iff g_1 H = g_2 H$$

(resp. $g_1 \sim g_2$ iff $Hg_1 = Hg_2$).

Obviously, $\sim$ is an equivalence relation. Now, it is easy to see that $g_1 H = g_2 H$ iff $g_2^{-1} g_1 \in H$, so the equivalence class of an element $g \in G$ is the coset $g H$ (resp. $Hg$). The set of left cosets of $H$ in $G$ (which, in general, is not a group) is denoted $G/H$. The “points” of $G/H$ are obtained by “collapsing” all the elements in a coset into a single element. This is the same intuition used for constructing the quotient space topology. The set of right cosets is denoted by $H \backslash G$.

It is tempting to define a multiplication operation on left cosets (or right cosets) by setting

$$(g_1 H)(g_2 H) = (g_1 g_2) H,$$

but this operation is not well defined in general, unless the subgroup $H$ possesses a special property. This property is typical of the kernels of group homomorphisms, so we are led to

**Definition 5.4.** Given any two groups $G$ and $G'$, a function $\varphi : G \to G'$ is a **homomorphism** iff

$$\varphi(g_1 g_2) = \varphi(g_1) \varphi(g_2), \quad \text{for all } g_1, g_2 \in G.$$

Taking $g_1 = g_2 = e$ (in $G$), we see that

$$\varphi(e) = e',$$

and taking $g_1 = g$ and $g_2 = g^{-1}$, we see that

$$\varphi(g^{-1}) = \varphi(g)^{-1}.$$

If $\varphi : G \to G'$ and $\psi : G' \to G''$ are group homomorphisms, then $\psi \circ \varphi : G \to G''$ is also a homomorphism. If $\varphi : G \to G'$ is a homomorphism of groups, and $H \subseteq G$, $H' \subseteq G'$ are two subgroups, then it is easily checked that

$$\text{Im } H = \varphi(H) = \{ \varphi(g) \mid g \in H \}$$

is a subgroup of $G'$ called the *image of $H$ by $\varphi$*, and

$$\varphi^{-1}(H') = \{ g \in G \mid \varphi(g) \in H' \}$$

is a subgroup of $G$. In particular, when $H' = \{ e' \}$, we obtain the *kernel* $\text{Ker } \varphi$ of $\varphi$. Thus,

$$\text{Ker } \varphi = \{ g \in G \mid \varphi(g) = e' \}.$$
It is immediately verified that \( \varphi : G \to G' \) is injective iff \( \text{Ker } \varphi = \{ e \} \). (We also write \( \text{Ker } \varphi = (0) \).) We say that \( \varphi \) is an isomorphism if there is a homomorphism \( \psi : G' \to G \), so that
\[
\psi \circ \varphi = \text{id}_G \quad \text{and} \quad \varphi \circ \psi = \text{id}_{G'}.
\]
In this case, \( \psi \) is unique and it is denoted \( \varphi^{-1} \). When \( \varphi \) is an isomorphism, we say the the groups \( G \) and \( G' \) are isomorphic and we write \( G \cong G' \) (or \( G \approx G' \)). When \( G' = G \), a group isomorphism is called an automorphism.

We claim that \( H = \text{Ker } \varphi \) satisfies the following property:
\[
gH = Hg, \quad \text{for all } g \in G. \tag{*}
\]
First, note that \( (*) \) is equivalent to
\[
gHg^{-1} = H, \quad \text{for all } g \in G,
\]
and the above is equivalent to
\[
gHg^{-1} \subseteq H, \quad \text{for all } g \in G. \tag{**}
\]
This is because \( gHg^{-1} \subseteq H \) implies \( H \subseteq g^{-1}Hg \), and this for all \( g \in G \). But
\[
\varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e',
\]
for all \( h \in H = \text{Ker } \varphi \) and all \( g \in G \). Thus, by definition of \( H = \text{Ker } \varphi \), we have \( gHg^{-1} \subseteq H \).

**Definition 5.5.** For any group \( G \), a subgroup \( N \) of \( G \) is a normal subgroup of \( G \) iff
\[
gNg^{-1} = N, \quad \text{for all } g \in G.
\]
This is denoted by \( N \triangleleft G \).

If \( N \) is a normal subgroup of \( G \), the equivalence relation induced by left cosets is the same as the equivalence induced by right cosets. Furthermore, this equivalence relation \( \sim \) is a congruence, which means that: For all \( g_1, g_2, g'_1, g'_2 \in G \),
\[
\begin{align*}
(1) & \quad \text{If } g_1N = g'_1N \text{ and } g_2N = g'_2N, \text{ then } g_1g_2N = g'_1g'_2N, \text{ and} \\
(2) & \quad \text{If } g_1N = g_2N, \text{ then } g_1^{-1}N = g_2^{-1}N.
\end{align*}
\]
As a consequence, we can define a group structure on the set \( G/ \sim \) of equivalence classes modulo \( \sim \), by setting
\[
(g_1N)(g_2N) = (g_1g_2)N.
\]
This group is denoted \( G/N \). The equivalence class \( gN \) of an element \( g \in G \) is also denoted \( \bar{g} \). The map \( \pi : G \to G/N \), given by
\[
\pi(g) = \bar{g} = gN
\]
is clearly a group homomorphism called the canonical projection.

Given a homomorphism of groups \( \varphi : G \to G' \), we easily check that the groups \( G/\text{Ker } \varphi \) and \( \text{Im } \varphi = \varphi(G) \) are isomorphic.
5.2 Group Actions and Homogeneous Spaces, I

If $X$ is a set (usually, some kind of geometric space, for example, the sphere in $\mathbb{R}^3$, the upper half-plane, etc.), the “symmetries” of $X$ are often captured by the action of a group $G$ on $X$. In fact, if $G$ is a Lie group and the action satisfies some simple properties, the set $X$ can be given a manifold structure which makes it a projection (quotient) of $G$, a so-called “homogeneous space.”

**Definition 5.6.** Given a set $X$ and a group $G$, a **left action of $G$ on $X$** (for short, an **action of $G$ on $X$**) is a function $\varphi: G \times X \to X$, such that:

1. For all $g, h \in G$ and all $x \in X$,
$$\varphi(g, \varphi(h, x)) = \varphi(gh, x),$$
2. For all $x \in X$,
$$\varphi(1, x) = x,$$
where $1 \in G$ is the identity element of $G$.

To alleviate the notation, we usually write $g \cdot x$ or even $gx$ for $\varphi(g, x)$, in which case the above axioms read:

1. For all $g, h \in G$ and all $x \in X$,
$$g \cdot (h \cdot x) = gh \cdot x,$$
2. For all $x \in X$,
$$1 \cdot x = x.$$

The set $X$ is called a **(left) $G$-set**. The action $\varphi$ is **faithful** or **effective** iff for every $g$, if $g \cdot x = x$ for all $x \in X$, then $g = 1$. Faithful means that if the action behaves like the identity, it must be the identity. The action $\varphi$ is **transitive** iff for any two elements $x, y \in X$, there is some $g \in G$ so that $g \cdot x = y$.

Given an action $\varphi: G \times X \to X$, for every $g \in G$, we have a function $\varphi_g: X \to X$ defined by
$$\varphi_g(x) = g \cdot x,$$
for all $x \in X$.

Observe that $\varphi_g$ has $\varphi_{g^{-1}}$ as inverse, since
$$\varphi_{g^{-1}}(\varphi_g(x)) = \varphi_{g^{-1}}(g \cdot x) = g^{-1} \cdot (g \cdot x) = (g^{-1}g) \cdot x = 1 \cdot x = x,$$
and similarly, $\varphi_g \circ \varphi_{g^{-1}} = \text{id}$. Therefore, $\varphi_g$ is a bijection of $X$; that is, $\varphi_g$ is a permutation of $X$. Moreover, we check immediately that
$$\varphi_g \circ \varphi_h = \varphi_{gh},$$
so the map \( g \mapsto \varphi_g \) is a group homomorphism from \( G \) to \( \mathfrak{S}_X \), the group of permutations of \( X \). With a slight abuse of notation, this group homomorphism \( G \to \mathfrak{S}_X \) is also denoted \( \varphi \).

Conversely, it is easy to see that any group homomorphism \( \varphi: G \to \mathfrak{S}_X \) yields a group action \( \cdot: G \times X \to X \), by setting

\[ g \cdot x = \varphi(g)(x). \]

Observe that an action \( \varphi \) is faithful iff the group homomorphism \( \varphi: G \to \mathfrak{S}_X \) is injective, i.e. iff \( \varphi \) has a trivial kernel. Also, we have \( g \cdot x = y \) iff \( g^{-1} \cdot y = x \), since \( (gh) \cdot x = g \cdot (h \cdot x) \) and \( 1 \cdot x = x \), for all \( g, h \in G \) and all \( x \in X \).

**Definition 5.7.** Given two \( G \)-sets \( X \) and \( Y \), a function \( f: X \to Y \) is said to be equivariant, or a \( G \)-map, iff for all \( x \in X \) and all \( g \in G \), we have

\[ f(g \cdot x) = g \cdot f(x). \]

Equivalently, if the \( G \)-actions are denoted by \( \varphi: G \times X \to X \) and \( \psi: G \times Y \to Y \), we have the following commutative diagram for all \( g \in G \):

\[
\begin{array}{ccc}
X & \xrightarrow{\varphi_g} & X \\
\downarrow{f} & & \downarrow{f} \\
Y & \xrightarrow{\psi_g} & Y
\end{array}
\]

**Remark:** We can also define a right action \( \cdot: X \times G \to X \) of a group \( G \) on a set \( X \) as a map satisfying the conditions

1. For all \( g, h \in G \) and all \( x \in X \),
   \[(x \cdot g) \cdot h = x \cdot gh,
   \]
2. For all \( x \in X \),
   \[x \cdot 1 = x.
   \]

Every notion defined for left actions is also defined for right actions, in the obvious way.

However, one change is necessary. For every \( g \in G \), the map \( \varphi_g: X \to X \) must be defined as

\[ \varphi_g(x) = x \cdot g^{-1}, \]

in order for the map \( g \mapsto \varphi_g \) from \( G \) to \( \mathfrak{S}_X \) to be a homomorphism \( (\varphi_g \circ \varphi_h = \varphi_{gh}) \).

Conversely, given a homomorphism \( \varphi: G \to \mathfrak{S}_X \), we get a right action \( \cdot: X \times G \to X \) by setting

\[ x \cdot g = \varphi(g^{-1})(x). \]

Here are some examples of (left) group actions.
Example 5.2. The unit sphere $S^2$ (more generally, $S^{n-1}$).

Recall that for any $n \geq 1$, the (real) unit sphere $S^{n-1}$ is the set of points in $\mathbb{R}^n$ given by

$$S^{n-1} = \{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1 \}.$$

In particular, $S^2$ is the usual sphere in $\mathbb{R}^3$. Since the group $SO(3) = SO(3, \mathbb{R})$ consists of (orientation preserving) linear isometries, i.e., linear maps that are distance preserving (and of determinant $+1$), and every linear map leaves the origin fixed, we see that any rotation maps $S^2$ into itself.

Beware that this would be false if we considered the group of affine isometries $SE(3)$ of $\mathbb{E}^3$. For example, a screw motion does not map $S^2$ into itself, even though it is distance preserving, because the origin is translated.

Thus, for $X = S^2$ and $G = SO(3)$, we have an action $\cdot : SO(3) \times S^2 \to S^2$, given by the matrix multiplication

$$R \cdot x = Rx.$$

The verification that the above is indeed an action is trivial. This action is transitive. This is because, for any two points $x, y$ on the sphere $S^2$, there is a rotation whose axis is perpendicular to the plane containing $x, y$ and the center $O$ of the sphere (this plane is not unique when $x$ and $y$ are antipodal, i.e., on a diameter) mapping $x$ to $y$. See Figure 5.1.

![Figure 5.1: The rotation which maps $x$ to $y$](image)

Similarly, for any $n \geq 1$, let $X = S^{n-1}$ and $G = SO(n)$ and define the action $\cdot : SO(n) \times S^{n-1} \to S^{n-1}$ as $R \cdot x = Rx$. It is easy to show that this action is transitive.

Analogously, we can define the (complex) unit sphere $\Sigma^{n-1}$, as the set of points in $\mathbb{C}^n$ given by

$$\Sigma^{n-1} = \{ (z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1 \overline{z}_1 + \cdots + z_n \overline{z}_n = 1 \}.$$
If we write \( z_j = x_j + iy_j \), with \( x_j, y_j \in \mathbb{R} \), then
\[
\Sigma^{n-1} = \{(x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^{2n} \mid x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 = 1\}.
\]
Therefore, we can view the complex sphere \( \Sigma^{n-1} \) (in \( \mathbb{C}^n \)) as the real sphere \( S^{2n-1} \) (in \( \mathbb{R}^{2n} \)). By analogy with the real case, we can define for \( X = \Sigma^{n-1} \) and \( G = \text{SU}(n) \) an action \( \cdot : \text{SU}(n) \times \Sigma^{n-1} \to \Sigma^{n-1} \) of the group \( \text{SU}(n) \) of linear maps of \( \mathbb{C}^n \) preserving the hermitian inner product (and the origin, as all linear maps do), and this action is transitive.

One should not confuse the unit sphere \( \Sigma^{n-1} \) with the hypersurface \( S^{n-1}_{\mathbb{C}} \), given by
\[
S^{n-1}_{\mathbb{C}} = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1^2 + \cdots + z_n^2 = 1\}.
\]
For instance, one should check that a line \( L \) through the origin intersects \( \Sigma^{n-1} \) in a circle, whereas it intersects \( S^{n-1}_{\mathbb{C}} \) in exactly two points! Recall for a fixed \( u = (x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{C}^n \), that \( L = \{\gamma u | \gamma \in \mathbb{C}\} \). Since \( \gamma = \rho (\cos \theta + i \sin \theta) \), we deduce that \( L \) is actually the two dimensional subspace through the origin spanned by the orthogonal vectors \((x_1, \ldots, x_n, y_1, \ldots, y_n)\) and \((−y_1, \ldots, −y_n, x_1, \ldots, x_n)\).

**Example 5.3.** The upper half-plane.

The **upper half-plane** \( H \) is the open subset of \( \mathbb{R}^2 \) consisting of all points \((x, y) \in \mathbb{R}^2\), with \( y > 0 \). It is convenient to identify \( H \) with the set of complex numbers \( z \in \mathbb{C} \) such that \( \Im z > 0 \). Then, we can let \( X = H \) and \( G = \text{SL}(2, \mathbb{R}) \) and define an action \( \cdot : \text{SL}(2, \mathbb{R}) \times H \to H \) of the group \( \text{SL}(2, \mathbb{R}) \) on \( H \), as follows: For any \( z \in H \), for any \( A \in \text{SL}(2, \mathbb{R}) \),
\[
A \cdot z = \frac{az + b}{cz + d},
\]
where
\[
A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}
\]
with \( ad - bc = 1 \).

It is easily verified that \( A \cdot z \) is indeed always well defined and in \( H \) when \( z \in H \) (check this). To see why this action is transitive, let \( z \) and \( w \) be two arbitrary points of \( H \) where \( z = x + iy \) and \( w = u + iv \) with \( x, u \in \mathbb{R} \) and \( y, v \in \mathbb{R}^+ \) (i.e. \( y \) and \( v \) are positive real numbers). Define \( A = \left( \begin{smallmatrix} \sqrt{v} & ny-vz \\ \sqrt{y} & \sqrt{v} \end{smallmatrix} \right) \). Note that \( A \in \text{SL}(2, \mathbb{R}) \). A routine calculation shows that \( A \cdot z = w \).

Before introducing Example 5.4, we need to define the groups of Möbius transformations and the Riemann sphere. Maps of the form
\[
z \mapsto \frac{az + b}{cz + d},
\]
where $z \in \mathbb{C}$ and $ad - bc = 1$, are called Möbius transformations. Here, $a, b, c, d \in \mathbb{R}$, but in general, we allow $a, b, c, d \in \mathbb{C}$. Actually, these transformations are not necessarily defined everywhere on $\mathbb{C}$, for example, for $z = -d/c$ if $c \neq 0$. To fix this problem, we add a “point at infinity” $\infty$ to $\mathbb{C}$, and define Möbius transformations as functions $\mathbb{C} \cup \{\infty\} \longrightarrow \mathbb{C} \cup \{\infty\}$. If $c = 0$, the Möbius transformation sends $\infty$ to itself, otherwise, $-d/c \mapsto \infty$ and $\infty \mapsto a/c$.

The space $\mathbb{C} \cup \{\infty\}$ can be viewed as the plane $\mathbb{R}^2$ extended with a point at infinity. Using a stereographic projection from the sphere $S^2$ to the plane (say from the north pole to the equatorial plane), we see that there is a bijection between the sphere $S^2$ and $\mathbb{C} \cup \{\infty\}$. More precisely, the stereographic projection $\sigma_N$ of the sphere $S^2$ from the north pole $N = (0, 0, 1)$ to the plane $z = 0$ (extended with the point at infinity $\infty$) is given by

$$(x, y, z) \in S^2 - \{(0, 0, 1)\} \mapsto \left( \frac{x}{1 - z}, \frac{y}{1 - z} \right) = \frac{x + iy}{1 - z} \in \mathbb{C}, \quad \text{with} \quad (0, 0, 1) \mapsto \infty.$$ 

The inverse stereographic projection $\sigma_N^{-1}$ is given by

$$(x, y) \mapsto \left( \frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1} \right), \quad \text{with} \quad \infty \mapsto (0, 0, 1).$$

Intuitively, the inverse stereographic projection “wraps” the equatorial plane around the sphere. See Figure 4.3.

The space $\mathbb{C} \cup \{\infty\}$ is known as the Riemann sphere. We will see shortly that $\mathbb{C} \cup \{\infty\} \cong S^2$ is also the complex projective line $\mathbb{CP}^1$. In summary, Möbius transformations are bijections of the Riemann sphere. It is easy to check that these transformations form a group under composition for all $a, b, c, d \in \mathbb{C}$, with $ad - bc = 1$. This is the Möbius group, denoted $\text{Möb}^+$. The Möbius transformations corresponding to the case $a, b, c, d \in \mathbb{R}$, with $ad - bc = 1$ form a subgroup of $\text{Möb}^+$ denoted $\text{Möb}^+_R$.

The map from $\text{SL}(2, \mathbb{C})$ to $\text{Möb}^+$ that sends $A \in \text{SL}(2, \mathbb{C})$ to the corresponding Möbius transformation is a surjective group homomorphism, and one checks easily that its kernel is $\{-I, I\}$ (where $I$ is the $2 \times 2$ identity matrix). Therefore, the Möbius group $\text{Möb}^+$ is isomorphic to the quotient group $\text{SL}(2, \mathbb{C})/\{-I, I\}$, denoted $\text{PSL}(2, \mathbb{C})$. This latter group turns out to be the group of projective transformations of the projective space $\mathbb{CP}^1$. The same reasoning shows that the subgroup $\text{Möb}^+_R$ is isomorphic to $\text{SL}(2, \mathbb{R})/\{-I, I\}$, denoted $\text{PSL}(2, \mathbb{R})$.

**Example 5.4.** The Riemann sphere $\mathbb{C} \cup \{\infty\}$.

Let $X = \mathbb{C} \cup \{\infty\}$ and $G = \text{SL}(2, \mathbb{C})$. The group $\text{SL}(2, \mathbb{C})$ acts on $\mathbb{C} \cup \{\infty\} \cong S^2$ the same way that $\text{SL}(2, \mathbb{R})$ acts on $H$, namely: For any $A \in \text{SL}(2, \mathbb{C})$, for any $z \in \mathbb{C} \cup \{\infty\}$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad ad - bc = 1.$$
This action is transitive, an exercise we leave for the reader.

**Example 5.5.** The unit disk.

One may recall from complex analysis that the (complex) Möbius transformation

\[ z \mapsto \frac{z - i}{z + i} \]

is a biholomorphic or analytic isomorphism between the upper half plane \( H \) and the open unit disk

\[ D = \{ z \in \mathbb{C} \mid |z| < 1 \}. \]

As a consequence, it is possible to define a transitive action of \( \text{SL}(2, \mathbb{R}) \) on \( D \). This can be done in a more direct fashion, using a group isomorphic to \( \text{SL}(2, \mathbb{R}) \), namely, \( \text{SU}(1, 1) \) (a group of complex matrices), but we don’t want to do this right now.

**Example 5.6.** The unit Riemann sphere revisited.

Another interesting action is the action of \( \text{SU}(2) \) on the extended plane \( \mathbb{C} \cup \{ \infty \} \). Recall that the group \( \text{SU}(2) \) consists of all complex matrices of the form

\[ A = \begin{pmatrix} \alpha & \beta \\ -\bar{\beta} & \bar{\alpha} \end{pmatrix}, \quad \alpha, \beta \in \mathbb{C}, \quad \alpha \bar{\alpha} + \beta \bar{\beta} = 1, \]

Let \( X = \mathbb{C} \cup \{ \infty \} \) and \( G = \text{SU}(2) \). The action \( \cdot : \text{SU}(2) \times (\mathbb{C} \cup \{ \infty \}) \to \mathbb{C} \cup \{ \infty \} \) is given by

\[ A \cdot w = \frac{\alpha w + \beta}{-\bar{\beta} w + \bar{\alpha}}, \quad w \in \mathbb{C} \cup \{ \infty \}. \]

This action is transitive, but the proof of this fact relies on the surjectivity of the group homomorphism

\[ \rho : \text{SU}(2) \to \text{SO}(3) \]

defined below, and the stereographic projection \( \sigma_N \) from \( S^2 \) onto \( \mathbb{C} \cup \{ \infty \} \). In particular, take \( z, w \in \mathbb{C} \cup \{ \infty \} \), use the inverse stereographic projection to obtain two points on \( S^2 \), namely \( \sigma_N^{-1}(z) \) and \( \sigma_N^{-1}(w) \). Then apply the appropriate rotation \( R \in \text{SO}(3) \) to map \( \sigma_N^{-1}(z) \) onto \( \sigma_N^{-1}(w) \). Such a rotation exists by the argument presented in Example 5.2. Since \( \rho : \text{SU}(2) \to \text{SO}(3) \) is surjective (see below), we know there must exist \( A \in \text{SU}(2) \) such that \( \rho(A) = R \) and \( A \cdot z = w \).

Using the stereographic projection \( \sigma_N \) from \( S^2 \) onto \( \mathbb{C} \cup \{ \infty \} \) and its inverse \( \sigma_N^{-1} \), we can define an action of \( \text{SU}(2) \) on \( S^2 \) by

\[ A \cdot (x, y, z) = \sigma_N^{-1}(A \cdot \sigma_N(x, y, z)), \quad (x, y, z) \in S^2. \]

Although this is not immediately obvious, it turns out that \( \text{SU}(2) \) acts on \( S^2 \) by maps that are restrictions of linear maps to \( S^2 \), and since these linear maps preserve \( S^2 \), they
are orthogonal transformations. Thus, we obtain a continuous (in fact, smooth) group homomorphism

$$\rho: \text{SU}(2) \to \text{O}(3).$$

Since \( \text{SU}(2) \) is connected and \( \rho \) is continuous, the image of \( \text{SU}(2) \) is contained in the connected component of \( I \) in \( \text{O}(3) \), namely \( \text{SO}(3) \), so \( \rho \) is a homomorphism

$$\rho: \text{SU}(2) \to \text{SO}(3).$$

We will see that this homomorphism is surjective and that its kernel is \( \{I, -I\} \). The upshot is that we have an isomorphism

$$\text{SO}(3) \cong \text{SU}(2)/\{I, -I\}.$$

The homomorphism \( \rho \) is a way of describing how a unit quaternion (any element of \( \text{SU}(2) \)) induces a rotation, via the stereographic projection and its inverse. If we write \( \alpha = a + ib \) and \( \beta = c + id \), a rather tedious computation yields

$$\rho(A) = \begin{pmatrix} a^2 - b^2 - c^2 + d^2 & -2ab - 2cd & -2ac + 2bd \\ 2ab - 2cd & a^2 - b^2 + c^2 - d^2 & -2ad - 2bc \\ 2ac + 2bd & 2ad - 2bc & a^2 + b^2 - c^2 - d^2 \end{pmatrix}.$$ 

One can check that \( \rho(A) \) is indeed a rotation matrix which represents the rotation whose axis is the line determined by the vector \((d, -c, b)\) and whose angle \( \theta \in [-\pi, \pi] \) is determined by

$$\cos \frac{\theta}{2} = |a|.$$ 

We can also compute the derivative \( d\rho_I: \text{su}(2) \to \text{so}(3) \) of \( \rho \) at \( I \) as follows. Recall that \( \text{su}(2) \) consists of all complex matrices of the form

$$\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}, \quad b, c, d \in \mathbb{R},$$

so pick the following basis for \( \text{su}(2) \),

$$X_1 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and define the curves in \( \text{SU}(2) \) through \( I \) given by

$$c_1(t) = \begin{pmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{pmatrix}, \quad c_2(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}, \quad c_3(t) = \begin{pmatrix} \cos t & i \sin t \\ i \sin t & \cos t \end{pmatrix}.$$ 

It is easy to check that \( c_i'(0) = X_i \) for \( i = 1, 2, 3 \), and that

$$d\rho_I(X_1) = 2 \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_2) = 2 \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_3) = 2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$
Thus, we have
\[ d\rho_I(X_1) = 2E_3, \quad d\rho_I(X_2) = -2E_2, \quad d\rho_I(X_3) = 2E_1, \]
where \((E_1, E_2, E_3)\) is the basis of \(\mathfrak{so}(3)\) given in Section 4.1, which means that \(d\rho_I\) is an isomorphism between the Lie algebras \(\mathfrak{su}(2)\) and \(\mathfrak{so}(3)\).

Recall from Proposition 4.13 that we have the commutative diagram
\[
\begin{array}{ccc}
\mathbf{SU}(2) & \xrightarrow{\rho} & \mathbf{SO}(3) \\
\exp & & \exp
\end{array}
\]
\[
\begin{array}{ccc}
\mathfrak{su}(2) & \xrightarrow{d\rho_I} & \mathfrak{so}(3)
\end{array}
\]
Since \(d\rho_I\) is surjective and the exponential map \(\exp: \mathfrak{so}(3) \to \mathbf{SO}(3)\) is surjective, we conclude that \(\rho\) is surjective. (We also know from Section 4.1 that \(\exp: \mathfrak{su}(2) \to \mathbf{SU}(2)\) is surjective.) Observe that \(\rho(-A) = \rho(A)\), and it is easy to check that \(\text{Ker} \rho = \{I, -I\}\).

**Example 5.7.** The set of \(n \times n\) symmetric, positive, definite matrices, \(\text{SPD}(n)\).

Let \(X = \text{SPD}(n)\) and \(G = \mathbf{GL}(n)\). The group \(\mathbf{GL}(n) = \mathbf{GL}(n, \mathbb{R})\) acts on \(\text{SPD}(n)\) as follows: For all \(A \in \mathbf{GL}(n)\) and all \(S \in \text{SPD}(n)\),
\[ A \cdot S = ASA^\top. \]
It is easily checked that \(ASA^\top\) is in \(\text{SPD}(n)\) if \(S\) is in \(\text{SPD}(n)\). First observe that \(ASA^\top\) is symmetric since
\[ (ASA^\top)^\top = AS^\top A^\top = ASA^\top. \]
Next recall the following characterization of positive definite matrix, namely
\[ y^\top Sy > 0, \quad \text{whenever} \ y \neq 0. \]
We want to show \(x^\top (A^\top SA)x > 0\) for all \(x \neq 0\). Since \(A\) is invertible, we have \(x = A^{-1}y\) for some nonzero \(y\), and hence
\[
x^\top (A^\top SA)x = y^\top (A^{-1})^\top A^\top SAA^{-1}y
\[
= y^\top Sy > 0.
\]
Hence \(A^\top SA\) is positive definite. This action is transitive because every SPD matrix \(S\) can be written as \(S = AA^\top\), for some invertible matrix \(A\) (prove this as an exercise). Given any two SPD matrices \(S_1 = A_1A_1^\top\) and \(S_2 = A_2A_2^\top\) with \(A_1\) and \(A_2\) invertible, if \(A = A_2A_1^{-1}\), we have
\[
A \cdot S_1 = A_2A_1^{-1}S_1(A_2A_1^{-1})^\top = A_2A_1^{-1}S_1(A_1^\top)^{-1}A_2^\top
\]
\[
= A_2A_1^{-1}A_1A_1^\top (A_1^\top)^{-1}A_2^\top = A_2A_2^\top = S_2.
\]
Example 5.8. The projective spaces $\mathbb{RP}^n$ and $\mathbb{CP}^n$.

The (real) projective space $\mathbb{RP}^n$ is the set of all lines through the origin in $\mathbb{R}^{n+1}$; that is, the set of one-dimensional subspaces of $\mathbb{R}^{n+1}$ (where $n \geq 0$). Since a one-dimensional subspace $L \subseteq \mathbb{R}^{n+1}$ is spanned by any nonzero vector $u \in L$, we can view $\mathbb{RP}^n$ as the set of equivalence classes of nonzero vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \in \mathbb{R}, \; \lambda \neq 0.$$

In terms of this definition, there is a projection $pr: (\mathbb{R}^{n+1} - \{0\}) \rightarrow \mathbb{RP}^n$, given by $pr(u) = [u]_\sim$, the equivalence class of $u$ modulo $\sim$. Write $[u]$ for the line defined by the nonzero vector $u$. Since every line $L$ in $\mathbb{R}^{n+1}$ intersects the sphere $S^n$ in two antipodal points, we can view $\mathbb{RP}^n$ as the quotient of the sphere $S^n$ by identification of antipodal points. See Figures 5.2 and 5.3.

![Figure 5.2: Three constructions for $\mathbb{RP}^1 \cong S^1$. Illustration (i.) applies the equivalence relation. Since any line through the origin, excluding the $x$-axis, intersects the line $y = 1$, its equivalence class is represented by its point of intersection on $y = 1$. Hence, $\mathbb{RP}^n$ is the disjoint union of the line $y = 1$ and the point of infinity given by the $x$-axis. Illustration (ii.) represents $\mathbb{RP}^1$ as the quotient of the circle $S^1$ by identification of antipodal points. Illustration (iii.) is a variation which glues the equatorial points of the upper semicircle.](image)

Let $X = \mathbb{RP}^n$ and $G = \text{SO}(n + 1)$. We define an action of $\text{SO}(n + 1)$ on $\mathbb{RP}^n$ as follows:
For any line \( L = [u] \), for any \( R \in SO(n + 1) \),

\[
R \cdot L = [Ru].
\]

Since \( R \) is linear, the line \([Ru]\) is well defined; that is, does not depend on the choice of \( u \in L \). The reader can show that this action is transitive.

The \((\text{complex})\) projective space \( \mathbb{C}P^n \) is defined analogously as the set of all lines through the origin in \( \mathbb{C}^{n+1} \); that is, the set of one-dimensional subspaces of \( \mathbb{C}^{n+1} \) (where \( n \geq 0 \)). This time, we can view \( \mathbb{C}P^n \) as the set of equivalence classes of vectors in \( \mathbb{C}^{n+1} - \{0\} \) modulo the equivalence relation

\[
u \sim v \iff v = \lambda u, \text{ for some } \lambda \neq 0 \in \mathbb{C}.
\]

We have the projection \( pr : \mathbb{C}^{n+1} - \{0\} \to \mathbb{C}P^n \), given by \( pr(u) = [u]_\sim \), the equivalence class of \( u \) modulo \( \sim \). Again, write \([u]\) for the line defined by the nonzero vector \( u \). Let \( X = \mathbb{C}P^n \) and \( G = SU(n + 1) \). We define an action of \( SU(n + 1) \) on \( \mathbb{C}P^n \) as follows: For
any line \( L = [u] \), for any \( R \in \text{SU}(n+1) \),
\[
R \cdot L = [Ru].
\]
Again, this action is well defined and it is transitive. (Check this.)

Before progressing to our final example of group actions, we take a moment to construct \( \mathbb{CP}^n \) as a quotient space of \( S^{2n+1} \). Recall that \( \Sigma^n \subseteq \mathbb{C}^{n+1} \), the unit sphere in \( \mathbb{C}^{n+1} \), is defined by
\[
\Sigma^n = \{ (z_1, \ldots, z_{n+1}) \in \mathbb{C}^{n+1} \mid z_1 \bar{z}_1 + \cdots + z_{n+1} \bar{z}_{n+1} = 1 \}.
\]
For any line \( L = [u] \), where \( u \in \mathbb{C}^{n+1} \) is a nonzero vector, writing \( u = (u_1, \ldots, u_{n+1}) \), a point \( z \in \mathbb{C}^{n+1} \) belongs to \( L \) iff \( z = \lambda(u_1, \ldots, u_{n+1}) \), for some \( \lambda \in \mathbb{C} \). Therefore, the intersection \( L \cap \Sigma^n \) of the line \( L \) and the sphere \( \Sigma^n \) is given by
\[
L \cap \Sigma^n = \{ \lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \mid \lambda \in \mathbb{C}, \ |\lambda| = \frac{1}{\sqrt{|u_1|^2 + \cdots + |u_{n+1}|^2}} \}.
\]
Thus, we see that there is a bijection between \( L \cap \Sigma^n \) and the circle \( S^1 \); that is, geometrically \( L \cap \Sigma^n \) is a circle. Moreover, since any line \( L \) through the origin is determined by just one other point, we see that for any two lines \( L_1 \) and \( L_2 \) through the origin,
\[
L_1 \neq L_2 \iff (L_1 \cap \Sigma^n) \cap (L_2 \cap \Sigma^n) = \emptyset.
\]
However, \( \Sigma^n \) is the sphere \( S^{2n+1} \) in \( \mathbb{R}^{2n+2} \). It follows that \( \mathbb{CP}^n \) is the quotient of \( S^{2n+1} \) by the equivalence relation \( \sim \) defined such that
\[
y \sim z \iff y, z \in L \cap \Sigma^n, \quad \text{for some line, } L, \text{ through the origin}.
\]
Therefore, we can write
\[
S^{2n+1}/S^1 \cong \mathbb{CP}^n.
\]
The case \( n = 1 \) is particularly interesting, as it turns out that
\[
S^3/S^1 \cong S^2.
\]
This is the famous Hopf fibration. To show this, proceed as follows: As
\[
S^3 \cong \Sigma^1 = \{ (z, z') \in \mathbb{C}^2 \mid |z|^2 + |z'|^2 = 1 \},
\]
define a map, \( HF: S^3 \to S^2 \), by
\[
HF((z, z')) = (2z\bar{z}', |z|^2 - |z'|^2).
\]
We leave as a homework exercise to prove that this map has range $S^2$ and that

$$HF((z_1, z'_1)) = HF((z_2, z'_2)) \iff (z_1, z'_1) = \lambda(z_2, z'_2), \text{ for some } \lambda \text{ with } |\lambda| = 1.$$  

In other words, for any point, $p \in S^2$, the inverse image $HF^{-1}(p)$ (also called fibre over $p$) is a circle on $S^3$. Consequently, $S^3$ can be viewed as the union of a family of disjoint circles. This is the Hopf fibration. It is possible to visualize the Hopf fibration using the stereographic projection from $S^3$ onto $\mathbb{R}^3$. This is a beautiful and puzzling picture. For example, see Berger [18]. Therefore, HF induces a bijection from $\mathbb{C}P^1$ to $S^2$, and it is a homeomorphism.

**Example 5.9.** Affine spaces.

Let $X$ be a set and $E$ a real vector space. A transitive and faithful action $\cdot : E \times X \to X$ of the additive group of $E$ on $X$ makes $X$ into an affine space. The intuition is that the members of $E$ are translations.

Those familiar with affine spaces as in Gallier [72] (Chapter 2) or Berger [18] will point out that if $X$ is an affine space, then not only is the action of $E$ on $X$ transitive, but more is true: For any two points $a, b \in E$, there is a unique vector $u \in E$, such that $u \cdot a = b$. By the way, the action of $E$ on $X$ is usually considered to be a right action and is written additively, so $u \cdot a$ is written $a + u$ (the result of translating $a$ by $u$). Thus, it would seem that we have to require more of our action. However, this is not necessary because $E$ (under addition) is abelian. More precisely, we have the proposition

**Proposition 5.1.** If $G$ is an abelian group acting on a set $X$ and the action $\cdot : G \times X \to X$ is transitive and faithful, then for any two elements $x, y \in X$, there is a unique $g \in G$ so that $g \cdot x = y$ (the action is simply transitive).

**Proof.** Since our action is transitive, there is at least some $g \in G$ so that $g \cdot x = y$. Assume that we have $g_1, g_2 \in G$ with

$$g_1 \cdot x = g_2 \cdot x = y.$$  

We shall prove that, actually

$$g_1 \cdot z = g_2 \cdot z, \text{ for all } z \in X.$$  

This implies that

$$g_1 g_2^{-1} \cdot z = z, \text{ for all } z \in X.$$  

As our action is faithful, $g_1 g_2^{-1} = 1$, and we must have $g_1 = g_2$, which proves our proposition.

Pick any $z \in X$. As our action is transitive, there is some $h \in G$ so that $z = h \cdot x$. Then,
5.2. GROUP ACTIONS AND HOMOGENEOUS SPACES, I

we have

\[ g_1 \cdot z = g_1 \cdot (h \cdot x) \]

\[ = (g_1 h) \cdot x \]

\[ = (h g_1) \cdot x \quad \text{(since } G \text{ is abelian)} \]

\[ = h \cdot (g_1 \cdot x) \]

\[ = h \cdot (g_2 \cdot x) \quad \text{(since } g_1 \cdot x = g_2 \cdot x) \]

\[ = (h g_2) \cdot x \]

\[ = (g_2 h) \cdot x \quad \text{(since } G \text{ is abelian)} \]

\[ = g_2 \cdot (h \cdot x) \]

\[ = g_2 \cdot z. \]

Therefore, \( g_1 \cdot z = g_2 \cdot z \) for all \( z \in X \), as claimed. \( \square \)

The subset of group elements that leave some given element \( x \in X \) fixed plays an important role.

**Definition 5.8.** Given an action \( \cdot : G \times X \to X \) of a group \( G \) on a set \( X \), for any \( x \in X \), the group \( G_x \) (also denoted \( \text{Stab}_G(x) \)), called the stabilizer of \( x \) or isotropy group at \( x \), is given by

\[ G_x = \{ g \in G \mid g \cdot x = x \}. \]

We have to verify that \( G_x \) is indeed a subgroup of \( G \), but this is easy. Indeed, if \( g \cdot x = x \) and \( h \cdot x = x \), then we also have \( h^{-1} \cdot x = x \) and so, we get \( gh^{-1} \cdot x = x \), proving that \( G_x \) is a subgroup of \( G \). In general, \( G_x \) is **not** a normal subgroup.

Observe that

\[ G_{g \cdot x} = gG_x g^{-1}, \]

for all \( g \in G \) and all \( x \in X \). Indeed,

\[ G_{g \cdot x} = \{ h \in G \mid h \cdot (g \cdot x) = g \cdot x \} \]

\[ = \{ h \in G \mid h g \cdot x = g \cdot x \} \]

\[ = \{ h \in G \mid g^{-1} h g \cdot x = x \}, \]

which shows \( g^{-1} G_{g \cdot x} g \subseteq G_x \), or equivalently that \( G_{g \cdot x} \subseteq gG_x g^{-1} \). It remains to show that \( gG_x g^{-1} \subseteq G_{g \cdot x} \). Take an element of \( gG_x g^{-1} \), which has the form \( ghg^{-1} \) with \( h \cdot x = x \). Since \( h \cdot x = x \), we have \( (ghg^{-1}) \cdot g x = g x \), which shows that \( ghg^{-1} \in G_{g \cdot x} \).

Because \( G_{g \cdot x} = gG_x g^{-1} \), the stabilizers of \( x \) and \( g \cdot x \) are conjugate of each other.

When the action of \( G \) on \( X \) is transitive, for any fixed \( x \in G \), the set \( X \) is a quotient (as a set, not as group) of \( G \) by \( G_x \). Indeed, we can define the map, \( \pi_x : G \to X \), by

\[ \pi_x(g) = g \cdot x, \quad \text{for all } g \in G. \]
Observe that
\[ \pi_x(gG_x) = (gG_x) \cdot x = g \cdot (G_x \cdot x) = g \cdot x = \pi_x(g). \]
This shows that \( \pi_x : G \to X \) induces a quotient map \( \pi_x : G/G_x \to X \), from the set \( G/G_x \) of (left) cosets of \( G_x \) to \( X \), defined by
\[ \pi_x(gG_x) = g \cdot x. \]
Since
\[ \pi_x(g) = \pi_x(h) \iff g \cdot x = h \cdot x \iff g^{-1}h \cdot x = x \iff g^{-1}h \in G_x \iff gG_x = hG_x, \]
we deduce that \( \pi_x : G/G_x \to X \) is injective. However, since our action is transitive, for every \( y \in X \), there is some \( g \in G \) so that \( g \cdot x = y \), and so \( \pi_x(gG_x) = g \cdot x = y \); that is, the map \( \pi_x \) is also surjective. Therefore, the map \( \pi_x : G/G_x \to X \) is a bijection (of sets, not groups). The map \( \pi_x : G \to X \) is also surjective. Let us record this important fact as

**Proposition 5.2.** If \( \cdot : G \times X \to X \) is a transitive action of a group \( G \) on a set \( X \), for every fixed \( x \in X \), the surjection \( \pi : G \to X \) given by
\[ \pi(g) = g \cdot x \]
induces a bijection
\[ \pi_x : G/G_x \to X, \]
where \( G_x \) is the stabilizer of \( x \). See Figure 5.4.

![Figure 5.4](image_url)

Figure 5.4: A schematic representation of \( G/G_x \cong X \), where \( G \) is the gray solid, \( X \) is its purple circular base, and \( G_x \) is the pink vertical strand. The dotted strands are the fibres \( gG_x \).
The map \( \pi : G \to X \) (corresponding to a fixed \( x \in X \)) is sometimes called a projection of \( G \) onto \( X \). Proposition 5.2 shows that for every \( y \in X \), the subset \( \pi^{-1}(y) \) of \( G \) (called the fibre above \( y \)) is equal to some coset \( gG_x \) of \( G \), and thus is in bijection with the group \( G_x \) itself. We can think of \( G \) as a moving family of fibres \( G_x \) parametrized by \( X \). This point of view of viewing a space as a moving family of simpler spaces is typical in (algebraic) geometry, and underlies the notion of (principal) fibre bundle.

Note that if the action \( \cdot : G \times X \to X \) is transitive, then the stabilizers \( G_x \) and \( G_y \) of any two elements \( x, y \in X \) are isomorphic, as they as conjugates. Thus, in this case, it is enough to compute one of these stabilizers for a “convenient” \( x \).

As the situation of Proposition 5.2 is of particular interest, we make the following definition:

**Definition 5.9.** A set \( X \) is said to be a homogeneous space if there is a transitive action \( \cdot : G \times X \to X \) of some group \( G \) on \( X \).

We see that all the spaces of Examples 5.2–5.9, are homogeneous spaces. Another example that will play an important role when we deal with Lie groups is the situation where we have a group \( G \), a subgroup \( H \) of \( G \) (not necessarily normal), and where \( X = G/H \), the set of left cosets of \( G \) modulo \( H \). The group \( G \) acts on \( G/H \) by left multiplication:

\[
a \cdot (gH) = (ag)H,
\]

where \( a, g \in G \). This action is clearly transitive and one checks that the stabilizer of \( gH \) is \( gHg^{-1} \). If \( G \) is a topological group and \( H \) is a closed subgroup of \( G \) (see later for an explanation), it turns out that \( G/H \) is Hausdorff. If \( G \) is a Lie group, we obtain a manifold.

Even if \( G \) and \( X \) are topological spaces and the action \( \cdot : G \times X \to X \) is continuous, in general, the space \( G/G_x \) under the quotient topology is not homeomorphic to \( X \).

We will give later sufficient conditions that insure that \( X \) is indeed a topological space or even a manifold. In particular, \( X \) will be a manifold when \( G \) is a Lie group.

In general, an action \( \cdot : G \times X \to X \) is not transitive on \( X \), but for every \( x \in X \), it is transitive on the set

\[
O(x) = G \cdot x = \{ g \cdot x \mid g \in G \}.
\]

Such a set is called the orbit of \( x \). The orbits are the equivalence classes of the following equivalence relation:

**Definition 5.10.** Given an action \( \cdot : G \times X \to X \) of some group \( G \) on \( X \), the equivalence relation \( \sim \) on \( X \) is defined so that, for all \( x, y \in X \),

\[
x \sim y \quad \text{iff} \quad y = g \cdot x, \quad \text{for some} \ g \in G.
\]

For every \( x \in X \), the equivalence class of \( x \) is the orbit of \( x \), denoted \( O(x) \) or \( G \cdot x \), with

\[
G \cdot x = O(x) = \{ g \cdot x \mid g \in G \}.
\]

The set of orbits is denoted \( X/G \).
We warn the reader that some authors use the notation \( G \backslash X \) for the set of orbits \( G \cdot x \), because these orbits can be considered as right orbits, by analogy with right cosets \( Hg \) of a subgroup \( H \) of \( G \).

The orbit space \( X/G \) is obtained from \( X \) by an identification (or merging) process: For every orbit, all points in that orbit are merged into a single point. This akin to the process of forming the identification topology. For example, if \( X = S^2 \) and \( G \) is the group consisting of the restrictions of the two linear maps \( I \) and \( -I \) of \( \mathbb{R}^3 \) to \( S^2 \) (where \( (-I)(x) = -x \) for all \( x \in \mathbb{R}^3 \)), then

\[
X/G = S^2/\{I, -I\} \cong \mathbb{R}P^2.
\]

See Figure 5.3. More generally, if \( S^n \) is the \( n \)-sphere in \( \mathbb{R}^{n+1} \), then we have a bijection between the orbit space \( S^n/\{I, -I\} \) and \( \mathbb{R}P^n \):

\[
S^n/\{I, -I\} \cong \mathbb{R}P^n.
\]

Many manifolds can be obtained in this fashion, including the torus, the Klein bottle, the Möbius band, etc.

Since the action of \( G \) is transitive on \( O(x) \), by Proposition 5.2, we see that for every \( x \in X \), we have a bijection

\[
O(x) \cong G/G_x.
\]

As a corollary, if both \( X \) and \( G \) are finite, for any set \( A \subseteq X \) of representatives from every orbit, we have the orbit formula:

\[
|X| = \sum_{a \in A} |G: G_a| = \sum_{a \in A} |G|/|G_a|.
\]

Even if a group action \( \cdot : G \times X \to X \) is not transitive, when \( X \) is a manifold, we can consider the set of orbits \( X/G \), and if the action of \( G \) on \( X \) satisfies certain conditions, \( X/G \) is actually a manifold. Manifolds arising in this fashion are often called orbifolds. In summary, we see that manifolds arise in at least two ways from a group action:

1. As homogeneous spaces \( G/G_x \), if the action is transitive.
2. As orbifolds \( X/G \) (under certain conditions on the action).

Of course, in both cases, the action must satisfy some additional properties.

For the rest of this section, we reconsider Examples 5.2–5.9 in the context of homogeneous space by determining some stabilizers for those actions.

(a) Consider the action \( \cdot : \text{SO}(n) \times S^{n-1} \to S^{n-1} \) of \( \text{SO}(n) \) on the sphere \( S^{n-1} \) \((n \geq 1)\) defined in Example 5.2. Since this action is transitive, we can determine the stabilizer of
any convenient element of $S^{n-1}$, say $e_1 = (1, 0, \ldots, 0)$. In order for any $R \in SO(n)$ to leave $e_1$ fixed, the first column of $R$ must be $e_1$, so $R$ is an orthogonal matrix of the form

$$R = \begin{pmatrix} 1 & U \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1,$$

where $U$ is a $1 \times (n-1)$ row vector. As the rows of $R$ must be unit vectors, we see that $U = 0$ and $S \in SO(n-1)$. Therefore, the stabilizer of $e_1$ is isomorphic to $SO(n-1)$, and we deduce the bijection

$$SO(n)/SO(n-1) \cong S^{n-1}.$$  

Strictly speaking, $SO(n-1)$ is not a subgroup of $SO(n)$, and in all rigor, we should consider the subgroup $\widetilde{SO}(n-1)$ of $SO(n)$ consisting of all matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1,$$

and write

$$SO(n)/\widetilde{SO}(n-1) \cong S^{n-1}.$$  

However, it is common practice to identify $SO(n-1)$ with $\widetilde{SO}(n-1)$.

When $n = 2$, as $SO(1) = \{1\}$, we find that $SO(2) \cong S^1$, a circle, a fact that we already knew. When $n = 3$, we find that $SO(3)/SO(2) \cong S^2$. This says that $SO(3)$ is somehow the result of gluing circles to the surface of a sphere (in $\mathbb{R}^3$), in such a way that these circles do not intersect. This is hard to visualize!

A similar argument for the complex unit sphere $\Sigma^{n-1}$ shows that

$$SU(n)/SU(n-1) \cong \Sigma^{n-1} \cong S^{2n-1}.$$  

Again, we identify $SU(n-1)$ with a subgroup of $SU(n)$, as in the real case. In particular, when $n = 2$, as $SU(1) = \{1\}$, we find that

$$SU(2) \cong S^3;$$

that is, the group $SU(2)$ is topologically the sphere $S^3$! Actually, this is not surprising if we remember that $SU(2)$ is in fact the group of unit quaternions.

(b) We saw in Example 5.3 that the action $\cdot : SL(2, \mathbb{R}) \times H \rightarrow H$ of the group $SL(2, \mathbb{R})$ on the upper half plane is transitive. Let us find out what the stabilizer of $z = i$ is. We should have

$$\frac{ai + b}{ci + d} = i,$$

that is, $ai + b = -c + di$, i.e.,

$$(d - a)i = b + c.$$
Since \(a, b, c, d\) are real, we must have \(d = a\) and \(b = -c\). Moreover, \(ad - bc = 1\), so we get \(a^2 + b^2 = 1\). We conclude that a matrix in \(\text{SL}(2, \mathbb{R})\) fixes \(i\) iff it is of the form

\[
\begin{pmatrix}
a & -b \\
b & a
\end{pmatrix}, \quad \text{with} \quad a^2 + b^2 = 1.
\]

Clearly, these are the rotation matrices in \(\text{SO}(2)\), and so the stabilizer of \(i\) is \(\text{SO}(2)\). We conclude that

\[
\text{SL}(2, \mathbb{R})/\text{SO}(2) \cong H.
\]

This time, we can view \(\text{SL}(2, \mathbb{R})\) as the result of glueing circles to the upper half plane. This is not so easy to visualize. There is a better way to visualize the topology of \(\text{SL}(2, \mathbb{R})\) by making it act on the open disk \(D\). We will return to this action in a little while.

(c) Now, consider the action of \(\text{SL}(2, \mathbb{C})\) on \(\mathbb{C} \cup \{\infty\} \cong \mathbb{S}^2\) given in Example 5.4. As it is transitive, let us find the stabilizer of \(z = 0\). We must have

\[
\begin{pmatrix}b \\ d \end{pmatrix} = 0,
\]

and as \(ad - bc = 1\), we must have \(b = 0\) and \(ad = 1\). Thus, the stabilizer of 0 is the subgroup \(\text{SL}(2, \mathbb{C})_0\) of \(\text{SL}(2, \mathbb{C})\) consisting of all matrices of the form

\[
\begin{pmatrix}a & 0 \\ c & a^{-1}\end{pmatrix}, \quad \text{where} \quad a \in \mathbb{C} - \{0\} \quad \text{and} \quad c \in \mathbb{C}.
\]

We get

\[
\text{SL}(2, \mathbb{C})/\text{SL}(2, \mathbb{C})_0 \cong \mathbb{C} \cup \{\infty\} \cong \mathbb{S}^2,
\]

but this is not very illuminating.

(d) In Example 5.7 we considered the action \(\cdot: \text{GL}(n) \times \text{SPD}(n) \to \text{SPD}(n)\) of \(\text{GL}(n)\) on \(\text{SPD}(n)\), the set of symmetric positive definite matrices. As this action is transitive, let us find the stabilizer of \(I\). For any \(A \in \text{GL}(n)\), the matrix \(A\) stabilizes \(I\) iff

\[
AA^\top = AA^\top = I.
\]

Therefore, the stabilizer of \(I\) is \(\text{O}(n)\), and we find that

\[
\text{GL}(n)/\text{O}(n) = \text{SPD}(n).
\]

Observe that if \(\text{GL}^+(n)\) denotes the subgroup of \(\text{GL}(n)\) consisting of all matrices with a strictly positive determinant, then we have an action \(\cdot: \text{GL}^+(n) \times \text{SPD}(n) \to \text{SPD}(n)\) of \(\text{GL}^+(n)\) on \(\text{SPD}(n)\). This action is transitive and we find that the stabilizer of \(I\) is \(\text{SO}(n)\); consequently, we get

\[
\text{GL}^+(n)/\text{SO}(n) = \text{SPD}(n).
\]
(e) In Example 5.8 we considered the action \( \cdot : \text{SO}(n+1) \times \mathbb{R}P^n \to \mathbb{R}P^n \) of \( \text{SO}(n+1) \) on the (real) projective space \( \mathbb{R}P^n \). As this action is transitive, let us find the stabilizer of the line \( L = [e_1] \), where \( e_1 = (1, 0, \ldots, 0) \). For any \( R \in \text{SO}(n+1) \), the line \( L \) is fixed iff either \( R(e_1) = e_1 \) or \( R(e_1) = -e_1 \), since \( e_1 \) and \( -e_1 \) define the same line. As \( R \) is orthogonal with \( \det(R) = 1 \), this means that \( R \) is of the form

\[
R = \begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with } \alpha = \pm 1 \text{ and } \det(S) = \alpha.
\]

But, \( S \) must be orthogonal, so we conclude \( S \in \text{O}(n) \). Therefore, the stabilizer of \( L = [e_1] \) is isomorphic to the group \( \text{O}(n) \), and we find that

\[
\text{SO}(n+1)/\text{O}(n) \cong \mathbb{R}P^n.
\]

Strictly speaking, \( \text{O}(n) \) is not a subgroup of \( \text{SO}(n+1) \), so the above equation does not make sense. We should write

\[
\text{SO}(n+1)/\tilde{\text{O}}(n) \cong \mathbb{R}P^n,
\]

where \( \tilde{\text{O}}(n) \) is the subgroup of \( \text{SO}(n+1) \) consisting of all matrices of the form

\[
\begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with } S \in \text{O}(n), \alpha = \pm 1 \text{ and } \det(S) = \alpha.
\]

This groups is also denoted \( S(\text{O}(1) \times \text{O}(n)) \). However, the common practice is to write \( \text{O}(n) \) instead of \( S(\text{O}(1) \times \text{O}(n)) \).

We should mention that \( \mathbb{R}P^3 \) and \( \text{SO}(3) \) are homeomorphic spaces. This is shown using the quaternions; for example, see Gallier [72], Chapter 8.

A similar argument applies to the action \( \cdot : \text{SU}(n+1) \times \mathbb{C}P^n \to \mathbb{C}P^n \) of \( \text{SU}(n+1) \) on the (complex) projective space \( \mathbb{C}P^n \). We find that

\[
\text{SU}(n+1)/\text{U}(n) \cong \mathbb{C}P^n.
\]

Again, the above is a bit sloppy as \( \text{U}(n) \) is not a subgroup of \( \text{SU}(n+1) \). To be rigorous, we should use the subgroup \( \tilde{\text{U}}(n) \) consisting of all matrices of the form

\[
\begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with } S \in \text{U}(n), |\alpha| = 1 \text{ and } \det(S) = \overline{\alpha}.
\]

This groups is also denoted \( S(\text{U}(1) \times \text{U}(n)) \). The common practice is to write \( \text{U}(n) \) instead of \( S(\text{U}(1) \times \text{U}(n)) \). In particular, when \( n = 1 \), we find that

\[
\text{SU}(2)/\text{U}(1) \cong \mathbb{C}P^1.
\]
But, we know that $\text{SU}(2) \cong S^3$, and clearly $\text{U}(1) \cong S^1$. So, again, we find that $S^3/S^1 \cong \mathbb{C}P^1$ (we know more, namely, $S^3/S^1 \cong S^2 \cong \mathbb{C}P^1$.)

Observe that $\mathbb{C}P^n$ can also be viewed as the orbit space of the action $\cdot : S^1 \times S^{2n+1} \to S^{2n+1}$ given by

$$\lambda \cdot (z_1, \ldots, z_{n+1}) = (\lambda z_1, \ldots, \lambda z_{n+1}),$$

where $S^1 = \text{U}(1)$ (the group of complex numbers of modulus 1) and $S^{2n+1}$ is identified with $\Sigma^n$.

We now return to Case (b) to give a better picture of $\text{SL}(2, \mathbb{R})$. Instead of having $\text{SL}(2, \mathbb{R})$ act on the upper half plane, we define an action of $\text{SL}(2, \mathbb{R})$ on the open unit disk $D$ as we did in Example 4. Technically, it is easier to consider the group $\text{SU}(1, 1)$, which is isomorphic to $\text{SL}(2, \mathbb{R})$, and to make $\text{SU}(1, 1)$ act on $D$. The group $\text{SU}(1, 1)$ is the group of $2 \times 2$ complex matrices of the form

$$\begin{pmatrix} a & b \\ b & \bar{a} \end{pmatrix}, \quad \text{with} \quad a\bar{a} - b\bar{b} = 1.$$  

The reader should check that if we let $$g = \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix},$$
then the map from $\text{SL}(2, \mathbb{R})$ to $\text{SU}(1, 1)$ given by

$$A \mapsto gAg^{-1}$$

is an isomorphism. Observe that the Möbius transformation associated with $g$ is

$$z \mapsto \frac{z - i}{z + i},$$

which is the holomorphic isomorphism mapping $H$ to $D$ mentioned earlier! Now, we can define a bijection between $\text{SU}(1, 1)$ and $S^1 \times D$ given by

$$\begin{pmatrix} a & b \\ b & \bar{a} \end{pmatrix} \mapsto (a/|a|, b/a).$$

We conclude that $\text{SL}(2, \mathbb{R}) \cong \text{SU}(1, 1)$ is topologically an open solid torus (i.e., with the surface of the torus removed). It is possible to further classify the elements of $\text{SL}(2, \mathbb{R})$ into three categories and to have geometric interpretations of these as certain regions of the torus. For details, the reader should consult Carter, Segal and Macdonald [38] or Duistermatt and Kolk [64] (Chapter 1, Section 1.2).

The group $\text{SU}(1, 1)$ acts on $D$ by interpreting any matrix in $\text{SU}(1, 1)$ as a Möbius transformation; that is,

$$\begin{pmatrix} a & b \\ b & \bar{a} \end{pmatrix} \mapsto \left( z \mapsto \frac{az + b}{bz + \bar{a}} \right).$$
The reader should check that these transformations preserve $D$.

Both the upper half-plane and the open disk are models of Lobachevsky’s non-Euclidean geometry (where the parallel postulate fails). They are also models of hyperbolic spaces (Riemannian manifolds with constant negative curvature, see Gallot, Hulin and Lafontaine [73], Chapter III). According to Dubrovin, Fomenko, and Novikov [62] (Chapter 2, Section 13.2), the open disk model is due to Poincaré and the upper half-plane model to Klein, although Poincaré was the first to realize that the upper half-plane is a hyperbolic space.

5.3 The Grassmann and Stiefel Manifolds

In this section we introduce two very important homogeneous manifolds, the Grassmann manifolds and the Stiefel manifolds. The Grassmann manifolds are generalizations of projective spaces (real and complex), while the Stiefel manifold are generalizations of $O(n)$. Both of these manifolds are examples of reductive homogeneous spaces; See Chapter 19. We begin by defining the Grassmann manifolds $G(k, n)$.

First, consider the real case. Given any $n \geq 1$, for any $k$, with $0 \leq k \leq n$, let $G(k, n)$ be the set of all linear $k$-dimensional subspaces of $\mathbb{R}^n$ (also called $k$-planes). Any $k$-dimensional subspace $U$ of $\mathbb{R}^n$ is spanned by $k$ linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$; write $U = \text{span}(u_1, \ldots, u_k)$. We can define an action $\cdot: O(n) \times G(k, n) \rightarrow G(k, n)$ as follows: For any $R \in O(n)$, for any $U = \text{span}(u_1, \ldots, u_k)$, let

$$R \cdot U = \text{span}(Ru_1, \ldots, Ru_k).$$

We have to check that the above is well defined. If $U = \text{span}(v_1, \ldots, v_k)$ for any other $k$ linearly independent vectors $v_1, \ldots, v_k$, we have

$$v_i = \sum_{j=1}^{k} a_{ij} u_j, \quad 1 \leq i \leq k,$$

for some $a_{ij} \in \mathbb{R}$, and so

$$Rv_i = \sum_{j=1}^{k} a_{ij} Ru_j, \quad 1 \leq i \leq k,$$

which shows that

$$\text{span}(Ru_1, \ldots, Ru_k) = \text{span}(Rv_1, \ldots, Rv_k);$$

that is, the above action is well defined.

We claim this action is transitive. This is because if $U$ and $V$ are any two $k$-planes, we may assume that $U = \text{span}(u_1, \ldots, u_k)$ and $V = \text{span}(v_1, \ldots, v_k)$, where the $u_i$’s form an orthonormal family and similarly for the $v_i$’s. Then, we can extend these families to orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ on $\mathbb{R}^n$, and w.r.t. the orthonormal basis
(u_1, \ldots, u_n), the matrix of the linear map sending u_i to v_i is orthogonal. Hence G(k, n) is a homogeneous space.

In order to represent G(k, n) as a quotient space, Proposition 5.2 implies it is enough to find the stabilizer of any k-plane. Pick \( U = \text{span}(e_1, \ldots, e_k) \), where \((e_1, \ldots, e_n)\) is the canonical basis of \( \mathbb{R}^n \) (i.e., \( e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \), with the 1 in the \( i \)th position). Now, any \( R \in O(n) \) stabilizes \( U \) iff \( R \) maps \( e_1, \ldots, e_k \) to \( k \) linearly independent vectors in the subspace \( U = \text{span}(e_1, \ldots, e_k) \), i.e., \( R \) is of the form

\[
R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},
\]

where \( S \) is \( k \times k \) and \( T \) is \((n - k) \times (n - k)\). Moreover, as \( R \) is orthogonal, \( S \) and \( T \) must be orthogonal, that is \( S \in O(k) \) and \( T \in O(n - k) \). We deduce that the stabilizer of \( U \) is isomorphic to \( O(k) \times O(n - k) \) and we find that

\[
O(n)/(O(k) \times O(n - k)) \cong G(k, n).
\]

It turns out that this makes \( G(k, n) \) into a smooth manifold of dimension \( k(n - k) \) called a Grassmannian.

The restriction of the action of \( O(n) \) on \( G(k, n) \) to \( SO(n) \) yields an action \( \cdot : SO(n) \times G(k, n) \to G(k, n) \) of \( SO(n) \) on \( G(k, n) \). Then, it is easy to see that this action is transitive and that the stabilizer of the subspace \( U \) is isomorphic to the subgroup \( S(O(k) \times O(n - k)) \) of \( SO(n) \) consisting of the rotations of the form

\[
R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},
\]

with \( S \in O(k), T \in O(n - k) \) and \( \det(S) \det(T) = 1 \). Thus, we also have

\[
SO(n)/S(O(k) \times O(n - k)) \cong G(k, n).
\]

If we recall the projection map of Example 5.8 in Section 5.2, namely \( pr : \mathbb{R}^{n+1} - \{0\} \to \mathbb{R}^n \), by definition, a \( k \)-plane in \( \mathbb{R}^n \) is the image under \( pr \) of any \((k + 1)\)-plane in \( \mathbb{R}^{n+1} \). So, for example, a line in \( \mathbb{R}^n \) is the image of a 2-plane in \( \mathbb{R}^{n+1} \), and a hyperplane in \( \mathbb{R}^n \) is the image of a hyperplane in \( \mathbb{R}^{n+1} \). The advantage of this point of view is that the \( k \)-planes in \( \mathbb{R}^n \) are arbitrary; that is, they do not have to go through “the origin” (which does not make sense, anyway!). Then, we see that we can interpret the Grassmannian, \( G(k + 1, n + 1) \), as a space of “parameters” for the \( k \)-planes in \( \mathbb{R}^n \). For example, \( G(2, n + 1) \) parametrizes the lines in \( \mathbb{R}^n \). In this viewpoint, \( G(k + 1, n + 1) \) is usually denoted \( G(k, n) \).

It can be proved (using some exterior algebra) that \( G(k, n) \) can be embedded in \( \mathbb{R}P^{(k)} \) (see Section 22.7). Much more is true. For example, \( G(k, n) \) is a projective variety, which means that it can be defined as a subset of \( \mathbb{R}P^{(k)} \) equal to the zero locus of a set of
5.3. THE GRASSMANN AND STIEFEL MANIFOLDS

homogeneous equations. There is even a set of quadratic equations known as the *Plücker equations* defining $G(k, n)$; for details, see Section 22.7. In particular, when $n = 4$ and $k = 2$, we have $G(2, 4) \subseteq \mathbb{R}P^5$, and $G(2, 4)$ is defined by a single equation of degree 2. The Grassmannian $G(2, 4) = \mathbb{G}(1, 3)$ is known as the *Klein quadric*. This hypersurface in $\mathbb{R}P^5$ parametrizes the lines in $\mathbb{R}P^3$.

Complex Grassmannians are defined in a similar way, by replacing $\mathbb{R}$ by $\mathbb{C}$ and $O(n)$ by $U(n)$ throughout. The complex Grassmannian $G_C(k, n)$ is a complex manifold as well as a real manifold, and we have

$$U(n)/(U(k) \times U(n-k)) \cong G_C(k, n).$$

As in the case of the real Grassmannians, the action of $U(n)$ on $G_C(k, n)$ yields an action of $SU(n)$ on $G_C(k, n)$, and we get

$$SU(n)/S(U(k) \times U(n-k)) \cong G_C(k, n),$$

where $S(U(k) \times U(n-k))$ is the subgroup of $SU(n)$ consisting of all matrices $R \in SU(n)$ of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

with $S \in U(k), T \in U(n-k)$ and $\det(S) \det(T) = 1$.

Closely related to Grassmannians are the *Stiefel manifolds* $S(k, n)$. Again, let us begin with the real case. For any $n \geq 1$ and any $k$ with $1 \leq k \leq n$, let $S(k, n)$ be the set of all *orthonormal* $k$-frames; that is, of $k$-tuples of orthonormal vectors $(u_1, \ldots, u_k)$ with $u_i \in \mathbb{R}^n$. Obviously, $S(1, n) = S^{n-1}$ and $S(n, n) = O(n)$, so assume $k \leq n - 1$. There is a natural action $\cdot : SO(n) \times S(k, n) \rightarrow S(k, n)$ of $SO(n)$ on $S(k, n)$ given by

$$R \cdot (u_1, \ldots, u_k) = (Ru_1, \ldots, Ru_k).$$

This action is transitive, because if $(u_1, \ldots, u_k)$ and $(v_1, \ldots, v_k)$ are any two orthonormal $k$-frames, then they can be extended to orthonormal bases (for example, by Gram-Schmidt) $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ with the same orientation (since we can pick $u_n$ and $v_n$ so that our bases have the same orientation), and there is a unique orthogonal transformation $R \in SO(n)$ such that $Ru_i = v_i$ for $i = 1, \ldots, n$.

In order to apply Proposition 5.2, we need to find the stabilizer of the orthonormal $k$-frame $(e_1, \ldots, e_k)$ consisting of the first canonical basis vectors of $\mathbb{R}^n$. A matrix $R \in SO(n)$ stabilizes $(e_1, \ldots, e_k)$ iff it is of the form

$$R = \begin{pmatrix} I_k & 0 \\ 0 & S \end{pmatrix},$$

where $S \in SO(n-k)$. Therefore, for $1 \leq k \leq n - 1$, we have

$$SO(n)/SO(n-k) \cong S(k, n).$$
This makes $S(k, n)$ a smooth manifold of dimension
\[ nk - \frac{k(k + 1)}{2} = k(n - k) + \frac{k(k - 1)}{2}. \]

**Remark:** It should be noted that we can define another type of Stiefel manifolds, denoted by $V(k, n)$, using linearly independent $k$-tuples $(u_1, \ldots, u_k)$ that do not necessarily form an orthonormal system. In this case, there is an action $\cdot : \text{GL}(n, \mathbb{R}) \times V(k, n) \to V(k, n)$, and the stabilizer $H$ of the first $k$ canonical basis vectors $(e_1, \ldots, e_k)$ is a closed subgroup of $\text{GL}(n, \mathbb{R})$, but it doesn’t have a simple description (see Warner [175], Chapter 3). We get an isomorphism
\[ V(k, n) \cong \text{GL}(n, \mathbb{R})/H. \]

The version of the Stiefel manifold $S(k, n)$ using orthonormal frames is sometimes denoted by $V^0(k, n)$ (Milnor and Stasheff [129] use the notation $V^0_k(\mathbb{R}^n)$). Beware that the notation is not standardized. Certain authors use $V(k, n)$ for what we denote by $S(k, n)$!

Complex Stiefel manifolds are defined in a similar way by replacing $\mathbb{R}$ by $\mathbb{C}$ and $\text{SO}(n)$ by $\text{SU}(n)$. For $1 \leq k \leq n - 1$, the complex Stiefel manifold $S_\mathbb{C}(k, n)$ is isomorphic to the quotient
\[ \text{SU}(n)/\text{SU}(n - k) \cong S_\mathbb{C}(k, n). \]
If $k = 1$, we have $S_\mathbb{C}(1, n) = S^{2n-1}$, and if $k = n$, we have $S_\mathbb{C}(n, n) = \text{U}(n)$.

The Grassmannians can also be viewed as quotient spaces of the Stiefel manifolds. Every orthonormal $k$-frame $(u_1, \ldots, u_k)$ can be represented by an $n \times k$ matrix $Y$ over the canonical basis of $\mathbb{R}^n$, and such a matrix $Y$ satisfies the equation
\[ Y^\top Y = I. \]

We have a right action $\cdot : S(k, n) \times \text{O}(k) \to S(k, n)$ given by
\[ Y \cdot R = YR, \]
for any $R \in \text{O}(k)$. This action is well defined since
\[ (YR)^\top YR = R^\top Y^\top YR = I. \]
However, this action is not transitive (unless $k = 1$), but the orbit space $S(k, n)/\text{O}(k)$ is isomorphic to the Grassmannian $G(k, n)$, so we can write
\[ G(k, n) \cong S(k, n)/\text{O}(k). \]

Similarly, the complex Grassmannian is isomorphic to the orbit space $S_\mathbb{C}(k, n)/\text{U}(k)$:
\[ G_\mathbb{C}(k, n) \cong S_\mathbb{C}(k, n)/\text{U}(k). \]
5.4 Topological Groups

Since Lie groups are topological groups (and manifolds), it is useful to gather a few basic facts about topological groups.

**Definition 5.11.** A set $G$ is a topological group iff

(a) $G$ is a Hausdorff topological space;

(b) $G$ is a group (with identity 1);

(c) Multiplication $\cdot : G \times G \to G$, and the inverse operation $G \to G: g \mapsto g^{-1}$, are continuous, where $G \times G$ has the product topology.

It is easy to see that the two requirements of condition (c) are equivalent to

(c') The map $G \times G \to G: (g, h) \mapsto gh^{-1}$ is continuous.

**Proposition 5.3.** If $G$ is a topological group and $H$ is any subgroup of $G$, then the closure $H$ of $H$ is a subgroup of $G$.

**Proof.** This follows easily from the continuity of multiplication and of the inverse operation, the details are left as an exercise to the reader.

Given a topological group $G$, for every $a \in G$ we define the left translation $L_a$ as the map $L_a: G \to G$ such that $L_a(b) = ab$, for all $b \in G$, and the right translation $R_a$ as the map $R_a: G \to G$ such that $R_a(b) = ba$, for all $b \in G$. Observe that $L_a^{-1}$ is the inverse of $L_a$ and similarly, $R_a^{-1}$ is the inverse of $R_a$. As multiplication is continuous, we see that $L_a$ and $R_a$ are continuous. Moreover, since they have a continuous inverse, they are homeomorphisms. As a consequence, if $U$ is an open subset of $G$, then so is $gU = L_g(U)$ (resp. $Ug = R_gU$), for all $g \in G$. Therefore, the topology of a topological group is determined by the knowledge of the open subsets containing the identity 1.

Given any subset $S \subseteq G$, let $S^{-1} = \{s^{-1} \mid s \in S\}$; let $S^0 = \{1\}$, and $S^{n+1} = S^nS$, for all $n \geq 0$. Property (c) of Definition 5.11 has the following useful consequences, which shows there exists an open set containing 1 which has a special symmetrical structure.

**Proposition 5.4.** If $G$ is a topological group and $U$ is any open subset containing 1, then there is some open subset $V \subseteq U$, with $1 \in V$, so that $V = V^{-1}$ and $V^2 \subseteq U$. Furthermore, $V \subseteq U$.

**Proof.** Since multiplication $G \times G \to G$ is continuous and $G \times G$ is given the product topology, there are open subsets $U_1$ and $U_2$, with $1 \in U_1$ and $1 \in U_2$, so that $U_1U_2 \subseteq U$. Let $W = U_1 \cap U_2$ and $V = W \cap W^{-1}$. Then, $V$ is an open set containing 1, and clearly $V = V^{-1}$ and $V^2 \subseteq U$. Let $g \in V$, then $gV$ is an open set containing $g$ (since $1 \in V$) and thus, $gV \cap V \neq \emptyset$. This means that there are some $h_1, h_2 \in V$ so that $gh_1 = h_2$, but then, $g = h_2h_1^{-1} \in VV^{-1} = VV \subseteq U$. 

Definition 5.12. A subset $U$ containing 1 and such that $U = U^{-1}$ is called symmetric.

Proposition 5.4 is used in the proofs of many the propositions and theorems on the structure of topological groups. For example, it is key in verifying the following proposition regarding discrete topological subgroups.

Definition 5.13. A subgroup $H$ of a topological group $G$ is discrete iff the induced topology on $H$ is discrete; that is, for every $h \in H$, there is some open subset $U$ of $G$ so that $U \cap H = \{h\}$.

Proposition 5.5. If $G$ is a topological group and $H$ is a discrete subgroup of $G$, then $H$ is closed.

Proof. As $H$ is discrete, there is an open subset $U$ of $G$ so that $U \cap H = \{1\}$, and by Proposition 5.4, we may assume that $U = U^{-1}$. If $g \in H$, as $gU$ is an open set containing $g$, we have $gU \cap H \neq \emptyset$. Consequently, there is some $y \in gU \cap H = gU^{-1} \cap H$, so $g \in gU$ with $y \in H$. We claim that $yU \cap H = \{y\}$. Note that $x \in yU \cap H$ means $x = yu$ with $yu \in H$ and $u \in U$. Since $H$ is a subgroup of $G$ and $y \in H$, $y^{-1}yu = u \in H$. Thus $u \in U \cap H$, which implies $u = 1$ and $yu = y$, and we have

$$g \in yU \cap H \subseteq yU \cap H = \{y\} = \{y\}.$$ 

since $G$ is Hausdorff. Therefore, $g = y \in H$. \hfill \Box

Using Proposition 5.4, we can give a very convenient characterization of the Hausdorff separation property in a topological group.

Proposition 5.6. If $G$ is a topological group, then the following properties are equivalent:

1. $G$ is Hausdorff;
2. The set $\{1\}$ is closed;
3. The set $\{g\}$ is closed, for every $g \in G$.

Proof. The implication (1) $\rightarrow$ (2) is true in any Hausdorff topological space. We just have to prove that $G - \{1\}$ is open, which goes as follows: For any $g \neq 1$, since $G$ is Hausdorff, there exists disjoint open subsets $U_g$ and $V_g$, with $g \in U_g$ and $1 \in V_g$. Thus, $\bigcup U_g = G - \{1\}$, showing that $G - \{1\}$ is open. Since $L_g$ is a homeomorphism, (2) and (3) are equivalent. Let us prove that (3) $\rightarrow$ (1). Let $g_1, g_2 \in G$ with $g_1 \neq g_2$. Then, $g_1^{-1}g_2 \neq 1$ and if $U$ and $V$ are distinct open subsets such that $1 \in U$ and $g_1^{-1}g_2 \in V$, then $g_1 \in g_1U$ and $g_2 \in g_1V$, where $g_1U$ and $g_1V$ are still open and disjoint. Thus, it is enough to separate 1 and $g \neq 1$. Pick any $g \neq 1$. If every open subset containing 1 also contained $g$, then 1 would be in the closure of $\{g\}$, which is absurd since $\{g\}$ is closed and $g \neq 1$. Therefore, there is some open subset $U$ such that $1 \in U$ and $g \notin U$. By Proposition 5.4, we can find an open subset $V$
containing 1, so that $VV \subseteq U$ and $V = V^{-1}$. We claim that $V$ and $gV$ are disjoint open sets with $1 \in V$ and $g \in gV$.

Since $1 \in V$, it is clear that $g \in gV$. If we had $V \cap gV \neq \emptyset$, then by the last sentence in the proof of Proposition 5.4 we would have $g \in VV^{-1} = VV \subseteq U$, a contradiction.

If $H$ is a subgroup of $G$ (not necessarily normal), we can form the set of left coset $G/H$, and we have the projection $p: G \to G/H$, where $p(g) = gH = \bar{g}$. If $G$ is a topological group, then $G/H$ can be given the quotient topology, where a subset $U \subseteq G/H$ is open iff $p^{-1}(U)$ is open in $G$. With this topology, $p$ is continuous. The trouble is that $G/H$ is not necessarily Hausdorff. However, we can neatly characterize when this happens.

**Proposition 5.7.** If $G$ is a topological group and $H$ is a subgroup of $G$, then the following properties hold:

1. The map $p: G \to G/H$ is an open map, which means that $p(V)$ is open in $G/H$ whenever $V$ is open in $G$.

2. The space $G/H$ is Hausdorff iff $H$ is closed in $G$.

3. If $H$ is open, then $H$ is closed and $G/H$ has the discrete topology (every subset is open).

4. The subgroup $H$ is open iff $1 \in H$ (i.e., there is some open subset $U$ so that $1 \in U \subseteq H$).

**Proof.** (1) Observe that if $V$ is open in $G$, then $VH = \bigcup_{h \in H} Vh$ is open, since each $Vh$ is open (as right translation is a homeomorphism). However, it is clear that

$$p^{-1}(p(V)) = VH,$$

i.e., $p^{-1}(p(V))$ is open which, by definition of the quotient topology, means that $p(V)$ is open.

(2) If $G/H$ is Hausdorff, then by Proposition 5.6, every point of $G/H$ is closed, i.e., each coset $gH$ is closed, so $H$ is closed. Conversely, assume $H$ is closed. Let $\bar{x}$ and $\bar{y}$ be two distinct point in $G/H$ and let $x, y \in G$ be some elements with $p(x) = \bar{x}$ and $p(y) = \bar{y}$. As $\bar{x} \neq \bar{y}$, the elements $x$ and $y$ are not in the same coset, so $x \notin yH$. As $H$ is closed, so is $yH$, and since $x \notin yH$, there is some open containing $x$ which is disjoint from $yH$, and we may assume (by translation) that it is of the form $Ux$, where $U$ is an open containing 1. By Proposition 5.4, there is some open $V$ containing 1 so that $VV \subseteq U$ and $V = V^{-1}$. Thus, we have

$$V^2 x \cap yH = \emptyset$$

and in fact,

$$V^2 xH \cap yH = \emptyset.$$
since $H$ is a group and $xH \cap yH = \emptyset$. (Recall that the cosets of $H$ partition $G$ into equivalence classes.) Since $V = V^{-1}$, we get

$$VxH \cap VyH = \emptyset,$$

and then, since $V$ is open, both $VxH$ and $VyH$ are disjoint, open, so $p(VxH)$ and $p(VyH)$ are open sets (by (1)) containing $\pi$ and $\gamma$ respectively and $p(VxH)$ and $p(VyH)$ are disjoint (because $p^{-1}(p(VxH)) = VxHH = VxH$, $p^{-1}(p(VyH)) = VyHH = VyH$, and $VxH \cap VyH = \emptyset$).

(3) If $H$ is open, then every coset $gH$ is open, so every point of $G/H$ is open and $G/H$ is discrete. Also, $\bigcup_{g \notin H} gH$ is open, i.e., $H$ is closed.

(4) Say $U$ is an open subset such that $1 \in U \subseteq H$. Then, for every $h \in H$, the set $hU$ is an open subset of $H$ with $h \in hU$, which shows that $H$ is open. The converse is trivial.

We next provide a criterion relating the connectivity of $G$ with that of $G/H$.

**Proposition 5.8.** Let $G$ be a topological group and $H$ be any subgroup of $G$. If $H$ and $G/H$ are connected, then $G$ is connected.

**Proof.** It is a standard fact of topology that a space $G$ is connected iff every continuous function $f$ from $G$ to the discrete space $\{0, 1\}$ is constant. See Proposition 3.15. Pick any continuous function $f$ from $G$ to $\{0, 1\}$. As $H$ is connected and left translations are homeomorphisms, all cosets $gH$ are connected. Thus, $f$ is constant on every coset $gH$. It follows that the function $f: G \to \{0, 1\}$ induces a continuous function $\overline{f}: G/H \to \{0, 1\}$ such that $f = \overline{f} \circ p$ (where $p: G \to G/H$; the continuity of $\overline{f}$ follows immediately from the definition of the quotient topology on $G/H$). As $G/H$ is connected, $\overline{f}$ is constant, and so $f = \overline{f} \circ p$ is constant.

The next three propositions describe how to generate a topological group from its symmetric neighborhood of $1$.

**Proposition 5.9.** If $G$ is a connected topological group, then $G$ is generated by any symmetric neighborhood $V$ of $1$. In fact,

$$G = \bigcup_{n \geq 1} V^n.$$

**Proof.** Since $V = V^{-1}$, it is immediately checked that $H = \bigcup_{n \geq 1} V^n$ is the group generated by $V$. As $V$ is a neighborhood of $1$, there is some open subset $U \subseteq V$, with $1 \in U$, and so $1 \in H$. From Proposition 5.7 (3), the subgroup $H$ is open and closed, and since $G$ is connected, $H = G$. 

Proposition 5.10. Let $G$ be a topological group and let $V$ be any connected symmetric open subset containing 1. Then, if $G_0$ is the connected component of the identity, we have

$$G_0 = \bigcup_{n \geq 1} V^n,$$

and $G_0$ is a normal subgroup of $G$. Moreover, the group $G/G_0$ is discrete.

Proof. First, as $V$ is open, every $V^n$ is open, so the group $\bigcup_{n \geq 1} V^n$ is open, and thus closed, by Proposition 5.7 (3). For every $n \geq 1$, we have the continuous map

$$V \times \cdots \times V \longrightarrow V^n : (g_1, \ldots, g_n) \mapsto g_1 \cdots g_n.$$

As $V$ is connected, $V \times \cdots \times V$ is connected, and so $V^n$ is connected. See Theorem 3.18 and Proposition 3.11. Since $1 \in V^n$ for all $n \geq 1$ and every $V^n$ is connected, we use Lemma 3.12 to conclude that $\bigcup_{n \geq 1} V^n$ is connected. Now, $\bigcup_{n \geq 1} V^n$ is connected, open and closed, so it is the connected component of 1. Finally, for every $g \in G$, the group $gG_0g^{-1}$ is connected and contains 1, so it is contained in $G_0$, which proves that $G_0$ is normal. Since $G_0$ is open, the group $G/G_0$ is discrete. \qed

Recall that a topological space $X$ is locally compact iff for every point $p \in X$, there is a compact neighborhood $C$ of $p$; that is, there is a compact $C$ and an open $U$, with $p \in U \subseteq C$. For example, manifolds are locally compact.

Proposition 5.11. Let $G$ be a topological group and assume that $G$ is connected and locally compact. Then, $G$ is countable at infinity, which means that $G$ is the union of a countable family of compact subsets. In fact, if $V$ is any symmetric compact neighborhood of 1, then

$$G = \bigcup_{n \geq 1} V^n.$$

Proof. Since $G$ is locally compact, there is some compact neighborhood $K$ of 1. Then, $V = K \cap K^{-1}$ is also compact and a symmetric neighborhood of 1. By Proposition 5.9, we have

$$G = \bigcup_{n \geq 1} V^n.$$

An argument similar to the one used in the proof of Proposition 5.10 to show that $V^n$ is connected if $V$ is connected proves that each $V^n$ compact if $V$ is compact. \qed

We end this section by combining the various properties of a topological group $G$ to characterize when $G/G_x$ is homeomorphic to $X$. In order to do so, we need two definitions.

Definition 5.14. Let $G$ be a topological group and let $X$ be a topological space. An action $\varphi: G \times X \rightarrow X$ is continuous (and $G$ acts continuously on $X$) if the map $\varphi$ is continuous.
If an action \( \varphi : G \times X \to X \) is continuous, then each map \( \varphi_g : X \to X \) is a homeomorphism of \( X \) (recall that \( \varphi_g(x) = g \cdot x \), for all \( x \in X \)).

Under some mild assumptions on \( G \) and \( X \), the quotient space \( G/G_x \) is homeomorphic to \( X \). For example, this happens if \( X \) is a Baire space.

**Definition 5.15.** A Baire space \( X \) is a topological space with the property that if \( \{F_i\}_{i \geq 1} \) is any countable family of closed sets \( F_i \) such that each \( F_i \) has empty interior, then \( \bigcup_{i \geq 1} F_i \) also has empty interior. By complementation, this is equivalent to the fact that for every countable family of open sets \( U_i \) such that each \( U_i \) is dense in \( X \) (i.e., \( \overline{U}_i = X \)), then \( \bigcap_{i \geq 1} U_i \) is also dense in \( X \).

**Remark:** A subset \( A \subseteq X \) is rare if its closure \( \overline{A} \) has empty interior. A subset \( Y \subseteq X \) is meager if it is a countable union of rare sets. Then, it is immediately verified that a space \( X \) is a Baire space iff every nonempty open subset of \( X \) is not meager.

The following theorem shows that there are plenty of Baire spaces:

**Theorem 5.12.** (Baire) (1) Every locally compact topological space is a Baire space.

(2) Every complete metric space is a Baire space.

A proof of Theorem 5.12 can be found in Bourbaki [30], Chapter IX, Section 5, Theorem 1.

We can now greatly improve Proposition 5.2 when \( G \) and \( X \) are topological spaces having some “nice” properties.

**Theorem 5.13.** Let \( G \) be a topological group which is locally compact and countable at infinity, \( X \) a Hausdorff topological space which is a Baire space, and assume that \( G \) acts transitively and continuously on \( X \). Then, for any \( x \in X \), the map \( \varphi : G/G_x \to X \) is a homeomorphism.

**Proof.** We follow the proof given in Bourbaki [30], Chapter IX, Section 5, Proposition 6 (Essentially the same proof can be found in Mneimné and Testard [130], Chapter 2). First, observe that if a topological group acts continuously and transitively on a Hausdorff topological space, then for every \( x \in X \), the stabilizer \( G_x \) is a closed subgroup of \( G \). This is because, as the action is continuous, the projection \( \pi : G \to X : g \mapsto g \cdot x \) is continuous, and \( G_x = \pi^{-1}\{x\} \), with \( \{x\} \) closed. Therefore, by Proposition 5.7, the quotient space \( G/G_x \) is Hausdorff. As the map \( \pi : G \to X \) is continuous, the induced map \( \varphi : G/G_x \to X \) is continuous, and by Proposition 5.2, it is a bijection. Therefore, to prove that \( \varphi \) is a homeomorphism, it is enough to prove that \( \varphi \) is an open map. For this, it suffices to show that \( \pi \) is an open map. Given any open \( U \) in \( G \), we will prove that for any \( g \in U \), the element \( \pi(g) = g \cdot x \) is contained in the interior of \( U \cdot x \). However, observe that this is equivalent to proving that \( x \) belongs to the interior of \( (g^{-1} \cdot U) \cdot x \). Therefore, we are reduced to the case: If \( U \) is any open subset of \( G \) containing \( 1 \), then \( x \) belongs to the interior of \( U \cdot x \).
Since $G$ is locally compact, using Proposition 5.4, we can find a compact neighborhood of the form $W = \overline{V}$, such that $1 \in W$, $W = W^{-1}$ and $W^2 \subseteq U$, where $V$ is open with $1 \in V \subseteq U$. As $G$ is countable at infinity, $G = \bigcup_{i \geq 1} K_i$, where each $K_i$ is compact. Since $V$ is open, all the cosets $gV$ are open, and as each $K_i$ is covered by the $gV$’s, by compactness of $K_i$, finitely many cosets $gV$ cover each $K_i$, and so

$$G = \bigcup_{i \geq 1} g_i V = \bigcup_{i \geq 1} g_i W,$$

for countably many $g_i \in G$, where each $g_i W$ is compact. As our action is transitive, we deduce that

$$X = \bigcup_{i \geq 1} g_i W \cdot x,$$

where each $g_i W \cdot x$ is compact, since our action is continuous and the $g_i W$ are compact. As $X$ is Hausdorff, each $g_i W \cdot x$ is closed, and as $X$ is a Baire space expressed as a union of closed sets, one of the $g_i W \cdot x$ must have nonempty interior; that is, there is some $w \in W$, with $g_i w \cdot x$ in the interior of $g_i W \cdot x$, for some $i$. But then, as the map $y \mapsto g \cdot y$ is a homeomorphism for any given $g \in G$ (where $y \in X$), we see that $x$ is in the interior of

$$w^{-1} g_i^{-1} \cdot (g_i W \cdot x) = w^{-1} W \cdot x \subseteq W^{-1} W \cdot x = W^2 \cdot x \subseteq U \cdot x,$$

as desired.

By Theorem 5.12, we get the following important corollary:

**Theorem 5.14.** Let $G$ be a topological group which is locally compact and countable at infinity, $X$ a Hausdorff locally compact topological space, and assume that $G$ acts transitively and continuously on $X$. Then, for any $x \in X$, the map $\varphi: G/G_x \to X$ is a homeomorphism.

Readers who wish to learn more about topological groups may consult Sagle and Walde [149] and Chevalley [41] for an introductory account, and Bourbaki [29], Weil [177] and Pontryagin [141, 142], for a more comprehensive account (especially the last two references).
Chapter 6

The Lorentz Groups ⊗

6.1 The Lorentz Groups $O(n, 1)$, $SO(n, 1)$ and $SO_0(n, 1)$

In chapter we study a class of linear Lie groups known as the Lorentz groups. As we will see, the Lorentz groups provide interesting examples of homogeneous spaces. Moreover, the Lorentz group $SO(3, 1)$ shows up in an interesting way in computer vision.

Denote the $p \times p$-identity matrix by $I_p$, for $p, q, \geq 1$, and define

$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}.$$  

If $n = p + q$, the matrix $I_{p,q}$ is associated with the nondegenerate symmetric bilinear form

$$\varphi_{p,q}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{n} x_j y_j$$  

with associated quadratic form

$$\Phi_{p,q}((x_1, \ldots, x_n)) = \sum_{i=1}^{p} x_i^2 - \sum_{j=p+1}^{n} x_j^2.$$  

In particular, when $p = 1$ and $q = 3$, we have the Lorentz metric

$$x_1^2 - x_2^2 - x_3^2 - x_4^2.$$  

In physics, $x_1$ is interpreted as time and written $t$, and $x_2, x_3, x_4$ as coordinates in $\mathbb{R}^3$ and written $x, y, z$. Thus, the Lorentz metric is usually written a

$$t^2 - x^2 - y^2 - z^2,$$

although it also appears as

$$x^2 + y^2 + z^2 - t^2,$$
which is equivalent but slightly less convenient for certain purposes, as we will see later. The space \( \mathbb{R}^4 \) with the Lorentz metric is called \textit{Minkowski space}. It plays an important role in Einstein’s theory of special relativity.

The group \( O(p,q) \) is the set of all \( n \times n \)-matrices

\[
O(p,q) = \{ A \in GL(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q} \}.
\]

This is the group of all invertible linear maps of \( \mathbb{R}^n \) that preserve the quadratic form \( \Phi_{p,q} \), i.e., the group of isometries of \( \Phi_{p,q} \). Let us check that \( O(p,q) \) is indeed a group.

If \( A, B \in O(p,q) \), then \( A^\top I_{p,q} A = I_{p,q} \) and \( B^\top I_{p,q} B = I_{p,q} \), so we get

\[
(AB^\top)I_{p,q}AB = B^\top A^\top I_{p,q} AB = B^\top I_{p,q} B = I_{p,q},
\]

which shows that \( AB \in O(p,q) \). Clearly, \( I \in O(p,q) \). Since \( I_{p,q}^2 = I \) and \( I_{p,q}^\top = I_{p,q} \), we have

\[
I_{p,q}^\top I_{p,q} = I_{p,q} I_{p,q} = I_{p,q},
\]

so \( I_{p,q} \in O(p,q) \). Since \( I_{p,q}^2 = I \), the condition \( A^\top I_{p,q} A = I_{p,q} \)

is equivalent to \( I_{p,q} A^\top I_{p,q} A = I \), which means that

\[
A^{-1} = I_{p,q} A^\top I_{p,q}.
\]

Consequently \( I = AA^{-1} = AI_{p,q} A^\top I_{p,q} \), so \( AI_{p,q} A^\top = I_{p,q} \) also holds, which shows that \( O(p,q) \) is closed under transposition (i.e., if \( A \in O(p,q) \), then \( A^\top \in O(p,q) \)). Then, if \( A \in O(p,q) \), since \( A^\top \in O(p,q) \) and \( I_{p,q} \in O(p,q) \), we have \( A^{-1} = I_{p,q} A^\top I_{p,q} \in O(p,q) \). So \( O(p,q) \) is indeed a subgroup of \( GL(n, \mathbb{R}) \) with inverse given by \( A^{-1} = I_{p,q} A^\top I_{p,q} \).

We have the subgroup

\[
SO(p,q) = \{ A \in O(p,q) \mid \det(A) = 1 \}
\]

consisting of the isometries of \( (\mathbb{R}^n, \Phi_{p,q}) \) with determinant +1. It is clear that \( SO(p,q) \) is also closed under transposition. The condition \( A^\top I_{p,q} A = I_{p,q} \) has an interpretation in terms of the inner product \( \varphi_{p,q} \) and the columns (and rows) of \( A \). Indeed, if we denote the \( j \)th column of \( A \) by \( A_j \), then

\[
A^\top I_{p,q} A = (\varphi_{p,q}(A_i, A_j)),
\]

so \( A \in O(p,q) \) iff the columns of \( A \) form an “orthonormal basis” w.r.t. \( \varphi_{p,q} \), i.e.,

\[
\varphi_{p,q}(A_i, A_j) = \begin{cases} 
\delta_{ij} & \text{if } 1 \leq i, j \leq p; \\
-\delta_{ij} & \text{if } p + 1 \leq i, j \leq p + q.
\end{cases}
\]

The difference with the usual orthogonal matrices is that \( \varphi_{p,q}(A_i, A_i) = -1 \), if \( p + 1 \leq i \leq p + q \). As \( O(p,q) \) is closed under transposition, the rows of \( A \) also form an orthonormal basis w.r.t. \( \varphi_{p,q} \).

It turns out that \( SO(p,q) \) has two connected components, and the component containing the identity is a subgroup of \( SO(p,q) \) denoted \( SO_0(p,q) \). The group \( SO_0(p,q) \) is actually homeomorphic to \( SO(p) \times SO(q) \times \mathbb{R}^{pq} \). This is not immediately obvious. A way to prove
this fact is to work out the polar decomposition for matrices in O(p, q). This is nicely done in Dragon [61] (see Section 6.2). A close examination of the factorization obtained in Section 6.3 also shows that there is bijection between O(p, q) and O(p) \times O(q) \times \mathbb{R}^{pq}. Another way to prove these results (in a stronger form, namely that there is a homeomorphism) is to use results on pseudo-algebraic subgroups of GL(n, \mathbb{C}); see Sections 6.4 and 6.5.

We will now determine the polar decomposition and the SVD decomposition of matrices in the Lorentz groups O(n, 1) and SO(n, 1). Write

\[ J = I_{n,1} \]

and given any \( A \in O(n, 1) \), write

\[ A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix}, \]

where \( B \) is an \( n \times n \) matrix, \( u, v \) are (column) vectors in \( \mathbb{R}^n \) and \( c \in \mathbb{R} \). We begin with the polar decomposition of matrices in the Lorentz groups O(n, 1).

**Proposition 6.1.** Every matrix \( A \in O(n, 1) \) has a polar decomposition of the form

\[ A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \quad \text{or} \quad A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}, \]

where \( Q \in O(n) \) and \( c = \sqrt{\|v\|^2 + 1} \).

**Proof.** Write \( A \) in block form as above. As the condition for \( A \) to be in \( O(n, 1) \) is \( A^\top JA = J \), we get

\[ \begin{pmatrix} B^\top & v \\ u^\top & c \end{pmatrix} \begin{pmatrix} B & u \\ -v^\top & -c \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix}, \]

i.e.,

\[ B^\top B = I_n + vv^\top \]
\[ u^\top u = c^2 - 1 \]
\[ B^\top u = cv. \]

If we remember that we also have \( AJA^\top = J \), then

\[ \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix} \begin{pmatrix} B^\top & v \\ -u^\top & -c \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix}, \]

and

\[ BB^\top = I_n + uu^\top \]
\[ v^\top v = c^2 - 1 \]
\[ Bv = cu. \]

From \( u^\top u = \|u\|^2 = c^2 - 1 \), we deduce that \( |c| \geq 1 \). From \( B^\top B = I_n + vv^\top \), we deduce that \( B^\top B \) is clearly symmetric; we also deduce that \( B^\top B \) positive definite since

\[ x^\top (I_n + vv^\top) x = \|x\|^2 + x^\top vv^\top x = \|x\|^2 + \|v^\top x\|^2, \]
and \( \|x\|^2 + \|v^TX\|^2 \) whenever \( x \neq 0 \). Now, geometrically, it is well known that \( vv^T/v^Tv \) is the orthogonal projection onto the line determined by \( v \). Consequently, the kernel of \( vv^T \) is the orthogonal complement of \( v \), and \( vv^T \) has the eigenvalue 0 with multiplicity \( n-1 \) and the eigenvalue \( c^2 - 1 = \|v\|^2 = v^Tv \) with multiplicity 1. The eigenvectors associated with 0 are orthogonal to \( v \) and the eigenvectors associated with \( c^2 - 1 \) are proportional with \( v \) since \( vv^T/\|v\|^2 = (c^2 - 1)v\|v\| \). It follows that \( I_n + vv^T \) has the eigenvalue 1 with multiplicity \( n-1 \) and the eigenvalue \( c^2 \) with multiplicity 1, the eigenvectors being as before. Now, \( B \) has polar form \( B = QS_1 \), where \( Q \) is orthogonal and \( S_1 \) is symmetric positive definite and \( S_1^2 = B^TB = I_n + vv^T \). Therefore, if \( c > 0 \), then \( S_1 = \sqrt{I_n + vv^T} \) is a symmetric positive definite matrix with eigenvalue 1 with multiplicity \( n-1 \) and eigenvalue \( c \) with multiplicity 1, the eigenvectors being as before. If \( c < 0 \), then change \( c \) to \( -c \).

**Case 1:** \( c > 0 \). Then \( v \) is an eigenvector of \( S_1 \) for \( c \) and we must also have \( Bv = cu \), which implies

\[
Bv = QS_1v = Q(cv) = cQv = cu,
\]

so

\[
Qv = u.
\]

It follows that

\[
A = \begin{pmatrix} B & u \\ v^T & c \end{pmatrix} = \begin{pmatrix} QS_1 & Qv \\ v^T & c \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^T} & v \\ v^T & c \end{pmatrix},
\]

where \( Q \in O(n) \) and \( c = \sqrt{\|v\|^2 + 1} \).

**Case 2:** \( c < 0 \). Then \( v \) is an eigenvector of \( S_1 \) for \( -c \) and we must also have \( Bv = cu \), which implies

\[
Bv = QS_1v = Q(-cv) = cQ(-v) = cu,
\]

so

\[
Q(-v) = u.
\]

It follows that

\[
A = \begin{pmatrix} B & u \\ v^T & c \end{pmatrix} = \begin{pmatrix} QS_1 & Q(-v) \\ v^T & c \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^T} & -v \\ -v^T & -c \end{pmatrix},
\]

where \( Q \in O(n) \) and \( c = -\sqrt{\|v\|^2 + 1} \).

We conclude that any \( A \in O(n,1) \) has a factorization of the form

\[
A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^T} & v \\ v^T & c \end{pmatrix} \quad \text{or} \quad A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^T} & v \\ v^T & c \end{pmatrix},
\]

where \( Q \in O(n) \) and \( c = \sqrt{\|v\|^2 + 1} \). Note that the matrix \( \begin{pmatrix} Q & 0 \\ 0 & \pm 1 \end{pmatrix} \) is orthogonal and \( \begin{pmatrix} \sqrt{I_n + vv^T} & v \\ v^T & c \end{pmatrix} \) is symmetric. Proposition 6.2 will show that \( \sqrt{I_n + vv^T} \) is positive definite. Hence the above factorizations are polar decompositions.
6.1. THE LORENTZ GROUPS $O(N,1)$, $SO(N,1)$ AND $SO_0(N,1)$

In order to show that $S = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$ is positive definite, we show that the eigenvalues are strictly positive. Such a matrix is called a Lorentz boost. Observe that if $v = 0$, then $c = 1$ and $S = I_{n+1}$.

**Proposition 6.2.** Assume $v \neq 0$. The eigenvalues of the symmetric positive definite matrix

$$S = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $c = \sqrt{\|v\|^2 + 1}$, are 1 with multiplicity $n - 1$, and $e^\alpha$ and $e^{-\alpha}$ each with multiplicity 1 (for some $\alpha \geq 0$). An orthonormal basis of eigenvectors of $S$ consists of vectors of the form

$$\begin{pmatrix} u_1 \\ 0 \\ \vdots \\ u_{n-1} \\ 0 \\ \sqrt{\frac{2\|v\|}{\sqrt{1 + \|v\|^2}}} \\ \frac{\sqrt{2\|v\|}}{\sqrt{1 + \|v\|^2}} \end{pmatrix},$$

where the $u_i \in \mathbb{R}^n$ are all orthogonal to $v$ and pairwise orthogonal.

**Proof.** Let us solve the linear system

$$\begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \begin{pmatrix} v \\ d \end{pmatrix} = \lambda \begin{pmatrix} v \\ d \end{pmatrix}.$$

We get

$$\sqrt{I_n + vv^\top}(v) + dv = \lambda v,$$

$$v^\top v + cd = \lambda d.$$

Since the proof of Proposition 6.1 implies that $c = \sqrt{\|v\|^2 + 1}$ and $\sqrt{I_n + vv^\top}(v) = cv$, the previous two equations are equivalent to

$$(c + d)v = \lambda v,$$

$$c^2 - 1 + cd = \lambda d.$$

Because $v \neq 0$, we get $\lambda = c + d$. Substituting in the second equation, we get

$$c^2 - 1 + cd = (c + d)d,$$

that is,

$$d^2 = c^2 - 1.$$

Thus, either $\lambda_1 = c + \sqrt{c^2 - 1}$ and $d = \sqrt{c^2 - 1}$, or $\lambda_2 = c - \sqrt{c^2 - 1}$ and $d = -\sqrt{c^2 - 1}$. Since $c \geq 1$ and $\lambda_1 \lambda_2 = 1$, set $\alpha = \log(c + \sqrt{c^2 - 1}) \geq 0$, so that $-\alpha = \log(c - \sqrt{c^2 - 1})$, and then $\lambda_1 = e^\alpha$ and $\lambda_2 = e^{-\alpha}$. On the other hand, if $u$ is orthogonal to $v$, observe that

$$\begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} u \\ 0 \end{pmatrix},$$

since the kernel of $vv^\top$ is the orthogonal complement of $v$. The rest is clear. \qed
Corollary 6.3. The singular values of any matrix $A \in O(n, 1)$ are 1 with multiplicity $n - 1$, $e^{\alpha}$, and $e^{-\alpha}$, for some $\alpha \geq 0$.

Note that the case $\alpha = 0$ is possible, in which case $A$ is an orthogonal matrix of the form

$$
\begin{pmatrix}
Q & 0 \\
0 & 1
\end{pmatrix}
\text{ or }
\begin{pmatrix}
Q & 0 \\
0 & -1
\end{pmatrix},
$$

with $Q \in O(n)$. The two singular values $e^{\alpha}$ and $e^{-\alpha}$ tell us how much $A$ deviates from being orthogonal.

By using Proposition 6.1 we see that $O(n, 1)$ has four components corresponding to the cases:

1. $Q \in O(n)$; $\det(Q) < 0$; +1 as the lower right entry of the orthogonal matrix;
2. $Q \in SO(n)$; −1 as the lower right entry of the orthogonal matrix;
3. $Q \in O(n)$; $\det(Q) < 0$; −1 as the lower right entry of the orthogonal matrix;
4. $Q \in SO(n)$; +1 as the lower right entry of the orthogonal matrix.

Observe that $\det(A) = -1$ in cases (1) and (2) and that $\det(A) = +1$ in cases (3) and (4). Thus, cases (3) and (4) correspond to the group $SO(n, 1)$, in which case the polar decomposition is of the form

$$
A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \left( \sqrt{I_n + vv^\top} \right) \begin{pmatrix} v \\ c \end{pmatrix},
$$

where $Q \in O(n)$, with $\det(Q) = -1$ and $c = \sqrt{\|v\|^2 + 1}$, or

$$
A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \left( \sqrt{I_n + vv^\top} \right) \begin{pmatrix} v \\ c \end{pmatrix},
$$

where $Q \in SO(n)$ and $c = \sqrt{\|v\|^2 + 1}$. The components in cases (1), (2) and (3) are not groups. We will show later that all four components are connected and that case (4) corresponds to a group (Proposition 6.6). This group is the connected component of the identity and it is denoted $SO_0(n, 1)$ (see Corollary 6.10). For the time being, note that $A \in SO_0(n, 1)$ iff $A \in SO(n, 1)$ and $a_{n+1,n+1} = c > 0$ (here, $A = (a_{ij})$.) In fact, we proved above that if $a_{n+1,n+1} > 0$, then $a_{n+1,n+1} \geq 1$.

Remark: If we let

$$
\Lambda_p = \begin{pmatrix} I_{n-1,1} & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Lambda_T = I_{n,1},
$$

where $I_{n,1} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix}$,
then we have the disjoint union

\[ \mathbf{O}(n,1) = \mathbf{SO}_0(n,1) \cup \Lambda_P \mathbf{SO}_0(n,1) \cup \Lambda_T \mathbf{SO}_0(n,1) \cup \Lambda_P \Lambda_T \mathbf{SO}_0(n,1). \]

We can now determine a convenient form for the SVD of matrices in \( \mathbf{O}(n,1) \).

**Theorem 6.4.** Every matrix \( A \in \mathbf{O}(n,1) \) can be written as

\[
A = \begin{pmatrix} P & \epsilon \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh \alpha & \sinh \alpha \\ 0 & \cdots & 0 & \sinh \alpha & \cosh \alpha \end{pmatrix} \begin{pmatrix} Q^T & 0 \\ 0 & 1 \end{pmatrix}
\]

with \( \epsilon = \pm 1 \), \( P \in \mathbf{O}(n) \) and \( Q \in \mathbf{SO}(n) \). When \( A \in \mathbf{SO}(n,1) \), we have \( \det(P)\epsilon = +1 \), and when \( A \in \mathbf{SO}_0(n,1) \), we have \( \epsilon = +1 \) and \( P \in \mathbf{SO}(n) \); that is,

\[
A = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \\ 0 & \cdots & 0 & \cosh \alpha \\ 0 & \cdots & 0 & \sinh \alpha \end{pmatrix} \begin{pmatrix} Q^T & 0 \\ 0 & 1 \end{pmatrix}
\]

with \( P \in \mathbf{SO}(n) \) and \( Q \in \mathbf{SO}(n) \).

**Proof.** By Proposition 6.1, any matrix \( A \in \mathbf{O}(n) \) can be written as

\[
A = \begin{pmatrix} R & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^T} & v \\ v^T & c \end{pmatrix}
\]

where \( \epsilon = \pm 1 \), \( R \in \mathbf{O}(n) \) and \( c = \sqrt{\|v\|^2 + 1} \). The case where \( c = 1 \) is trivial, so assume \( c > 1 \), which means that \( \alpha \) from Proposition 6.2 is such that \( \alpha > 0 \). The key fact is that the eigenvalues of the matrix

\[
\begin{pmatrix} \cosh \alpha & \sinh \alpha \\ \sinh \alpha & \cosh \alpha \end{pmatrix}
\]

are \( e^\alpha \) and \( e^{-\alpha} \). To verify this fact, observe that

\[
\det \begin{pmatrix} \cosh \alpha - \lambda & \sinh \alpha \\ \sinh \alpha & \cosh \alpha - \lambda \end{pmatrix} = (\cosh \alpha - \lambda)^2 - \sinh^2 \alpha = \lambda^2 - 2\lambda \cosh \alpha + 1 = 0,
\]

which in turn implies

\[
\lambda = \cosh \alpha \pm \sinh \alpha.
\]
and the conclusion follows from the definitions of \( \cosh \alpha = \frac{e^\alpha + e^{-\alpha}}{2} \) and \( \sinh \alpha = \frac{e^\alpha - e^{-\alpha}}{2} \).

Also observe that the definitions of \( \cosh \alpha \) and \( \sinh \alpha \) imply that

\[
\begin{pmatrix} e^\alpha & 0 \\ 0 & e^{-\alpha} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \cosh \alpha & \sinh \alpha \\ \sinh \alpha & \cosh \alpha \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix},
\]

which is equivalent to the observation that \( \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \) is the eigenvector associated with \( e^\alpha \), while \( \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \) is the eigenvector associated with \( e^{-\alpha} \).

From these two facts we see that the diagonal matrix

\[
D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & e^\alpha & 0 \\ 0 & \cdots & 0 & 0 & e^{-\alpha} \end{pmatrix}
\]

of eigenvalues of \( S = \begin{pmatrix} \sqrt{I_n + vv^\top} \\ v \end{pmatrix} \) is given by

\[
D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.
\]

By Proposition 6.2, an orthonormal basis of eigenvectors of \( S \) consists of vectors of the form

\[
\begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} u_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} v \sqrt{\|v\|} \\ 0 \end{pmatrix}, \begin{pmatrix} v \sqrt{\|v\|} \\ 0 \end{pmatrix}, \begin{pmatrix} v \sqrt{\|v\|} \\ 0 \end{pmatrix},
\]

where the \( u_i \in \mathbb{R}^n \) are all orthogonal to \( v \) and pairwise orthogonal. Now, if we multiply the matrices

\[
\begin{pmatrix} u_1 & \cdots & u_{n-1} & v \sqrt{\|v\|} & v \sqrt{\|v\|} \\ 0 & \cdots & 0 & \frac{v \sqrt{\|v\|}}{\sqrt{2}} & -\frac{v \sqrt{\|v\|}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},
\]

the conclusion follows from the definitions of \( \cosh \alpha \) and \( \sinh \alpha \).
we get an orthogonal matrix of the form

\[
\begin{pmatrix}
Q & 0 \\
0 & 1
\end{pmatrix}
\]

where the columns of \(Q\) are the vectors

\[
u_1, \dotsc, u_{n-1}, \frac{v}{\|v\|}.
\]

By flipping \(u_1\) to \(-u_1\) if necessary, we can make sure that this matrix has determinant +1. Consequently,

\[
S = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}
\begin{pmatrix}
1 & \cdots & 0 & 0 & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & 1 & 0 & 0 \\
0 & \cdots & 0 & \cosh \alpha & \sinh \alpha \\
0 & \cdots & 0 & \sinh \alpha & \cosh \alpha
\end{pmatrix}
\begin{pmatrix}
Q^T & 0 \\
0 & 1
\end{pmatrix},
\]

so

\[
A = \begin{pmatrix} R & 0 \\ 0 & \epsilon \end{pmatrix}
\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}
\begin{pmatrix}
1 & \cdots & 0 & 0 & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & 1 & 0 & 0 \\
0 & \cdots & 0 & \cosh \alpha & \sinh \alpha \\
0 & \cdots & 0 & \sinh \alpha & \cosh \alpha
\end{pmatrix}
\begin{pmatrix}
Q^T & 0 \\
0 & 1
\end{pmatrix},
\]

and if we let \(P = RQ\), we get the desired decomposition. \(\square\)

**Remark:** We warn our readers about Chapter 6 of Baker’s book [16]. Indeed, this chapter is seriously flawed. The main two Theorems (Theorem 6.9 and Theorem 6.10) are false, and as consequence, the proof of Theorem 6.11 is wrong too. Theorem 6.11 states that the exponential map \(\text{exp}: \mathfrak{so}(n, 1) \rightarrow \text{SO}_0(n, 1)\) is surjective, which is correct, but known proofs are nontrivial and quite lengthy (see Section 6.2). The proof of Theorem 6.12 is also false, although the theorem itself is correct (this is our Theorem 6.17, see Section 6.2). The main problem with Theorem 6.9 (in Baker) is that the existence of the normal form for matrices in \(\text{SO}_0(n, 1)\) claimed by this theorem is unfortunately false on several accounts. Firstly, it would imply that every matrix in \(\text{SO}_0(n, 1)\) can be diagonalized, but this is false for \(n \geq 2\). Secondly, even if a matrix \(A \in \text{SO}_0(n, 1)\) is diagonalizable as \(A = PDP^{-1}\), Theorem 6.9 (and Theorem 6.10) miss some possible eigenvalues and the matrix \(P\) is not necessarily in \(\text{SO}_0(n, 1)\) (as the case \(n = 1\) already shows). For a thorough analysis of the eigenvalues of Lorentz isometries (and much more), one should consult Riesz [145] (Chapter III).

Clearly, a result similar to Theorem 6.4 also holds for the matrices in the groups \(\text{O}(1, n),\)
SO(1, n) and SO_0(1, n). For example, every matrix \(A \in SO_0(1, n)\) can be written as

\[
A = \begin{pmatrix}
1 & 0 \\
0 & P
\end{pmatrix}
\begin{pmatrix}
\cosh \alpha & \sinh \alpha & 0 & \cdots & 0 \\
\sinh \alpha & \cosh \alpha & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 \\
0 & Q^T
\end{pmatrix},
\]

where \(P, Q \in SO(n)\).

In the case \(n = 3\), we obtain the proper orthochronous Lorentz group \(SO_0(1, 3)\), also denoted \(Lor(1, 3)\). By the way, \(O(1, 3)\) is called the (full) Lorentz group and \(SO(1, 3)\) is the special Lorentz group.

Theorem 6.4 (really, the version for \(SO_0(1, n)\)) shows that the Lorentz group \(SO_0(1, 3)\) is generated by the matrices of the form

\[
\begin{pmatrix}
1 & 0 \\
0 & P
\end{pmatrix}
\]

with \(P \in SO(3)\) and the matrices of the form

\[
\begin{pmatrix}
\cosh \alpha & \sinh \alpha & 0 & 0 \\
\sinh \alpha & \cosh \alpha & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}.
\]

This fact will be useful when we prove that the homomorphism \(\varphi : SL(2, \mathbb{C}) \to SO_0(1, 3)\) is surjective.

Remark: Unfortunately, unlike orthogonal matrices which can always be diagonalized over \(\mathbb{C}\), not every matrix in \(SO(1, n)\) can be diagonalized for \(n \geq 2\). This has to do with the fact that the Lie algebra \(so(1, n)\) has non-zero idempotents (see Section 6.2).

It turns out that the group \(SO_0(1, 3)\) admits another interesting characterization involving the hypersurface

\[
\mathcal{H} = \{ (t, x, y, z) \in \mathbb{R}^4 \mid t^2 - x^2 - y^2 - z^2 = 1 \}.
\]

This surface has two sheets, and it is not hard to show that \(SO_0(1, 3)\) is the subgroup of \(SO(1, 3)\) that preserves these two sheets (does not swap them). Actually, we will prove this fact for any \(n\). In preparation for this, we need some definitions and a few propositions.

Let us switch back to \(SO(n, 1)\). First, as a matter of notation, we write every \(u \in \mathbb{R}^{n+1}\) as \(u = (u, t)\), where \(u \in \mathbb{R}^n\) and \(t \in \mathbb{R}\), so that the Lorentz inner product can be expressed as

\[
\langle u, v \rangle = \langle (u, t), (v, s) \rangle = u \cdot v - ts,
\]

where \(u \cdot v\) is the standard Euclidean inner product (the Euclidean norm of \(x\) is denoted \(\|x\|\)). Then, we can classify the vectors in \(\mathbb{R}^{n+1}\) as follows:
Definition 6.1. A nonzero vector \( u = (u, t) \in \mathbb{R}^{n+1} \) is called

(a) spacelike iff \( \langle u, u \rangle > 0 \), i.e., iff \( \|u\|^2 > t^2 \);

(b) timelike iff \( \langle u, u \rangle < 0 \), i.e., iff \( \|u\|^2 < t^2 \);

(c) lightlike or isotropic iff \( \langle u, u \rangle = 0 \), i.e., iff \( \|u\|^2 = t^2 \).

A spacelike (resp. timelike, resp. lightlike) vector is said to be positive iff \( t > 0 \) and negative iff \( t < 0 \). The set of all isotropic vectors

\[
\mathcal{H}_n(0) = \{ u = (u, t) \in \mathbb{R}^{n+1} \mid \|u\|^2 = t^2 \}
\]

is called the light cone. For every \( r > 0 \), let

\[
\mathcal{H}_n(r) = \{ u = (u, t) \in \mathbb{R}^{n+1} \mid \|u\|^2 - t^2 = -r \},
\]

a hyperboloid of two sheets.

It is easy to check that \( \mathcal{H}_n(r) \) has two connected components as follows: First, since \( r > 0 \) and

\[
\|u\|^2 + r = t^2,
\]

we have \( |t| \geq \sqrt{r} \). Now, for any \( x = (x_1, \ldots, x_n, t) \in \mathcal{H}_n(r) \) with \( t \geq \sqrt{r} \), we have the continuous path from \((0, \ldots, 0, \sqrt{r})\) to \( x \) given by

\[
\lambda \mapsto (\lambda x_1, \ldots, \lambda x_n, \sqrt{r + \lambda^2(t^2 - r)}),
\]

where \( \lambda \in [0, 1] \), proving that the component of \( (0, \ldots, 0, \sqrt{r}) \) is connected. Similarly, when \( t \leq -\sqrt{r} \), we have the continuous path from \((0, \ldots, 0, -\sqrt{r})\) to \( x \) given by

\[
\lambda \mapsto (\lambda x_1, \ldots, \lambda x_n, -\sqrt{r + \lambda^2(t^2 - r)}),
\]

where \( \lambda \in [0, 1] \), proving that the component of \( (0, \ldots, 0, -\sqrt{r}) \) is connected. We denote the sheet containing \((0, \ldots, 0, \sqrt{r})\) by \( \mathcal{H}_n^+(r) \) and sheet containing \((0, \ldots, 0, -\sqrt{r})\) by \( \mathcal{H}_n^-(r) \).

Since every Lorentz isometry \( A \in \text{SO}(n, 1) \) preserves the Lorentz inner product, we conclude that \( A \) globally preserves every hyperboloid \( \mathcal{H}_n(r) \), for \( r > 0 \). We claim that every \( A \in \text{SO}_0(n, 1) \) preserves both \( \mathcal{H}_n^+(r) \) and \( \mathcal{H}_n^-(r) \). This follows immediately from

Proposition 6.5. If \( a_{n+1,n+1} > 0 \), then every isometry \( A \in \text{O}(n, 1) \) preserves all positive (resp. negative) timelike vectors and all positive (resp. negative) lightlike vectors. Moreover, if \( A \in \text{O}(n, 1) \) preserves all positive timelike vectors, then \( a_{n+1,n+1} > 0 \).
Proof. Let \( u = (u, t) \) be a nonzero timelike or lightlike vector. This means that
\[
\|u\|^2 \leq t^2 \quad \text{and} \quad t \neq 0.
\]
Since \( A \in O(n, 1) \), the matrix \( A \) preserves the inner product; if \( \langle u, u \rangle = \|u\|^2 - t^2 < 0 \), we get \( \langle Au, Au \rangle < 0 \), which shows that \( Au \) is also timelike. Similarly, if \( \langle u, u \rangle = 0 \), then \( \langle Au, Au \rangle = 0 \). Define \( A_{n+1} = (A, a_{n+1,n+1}) \) is the \((n+1)\)th row of the matrix \( A \). As \( A \in O(n, 1) \), we know that
\[
\langle A_{n+1}, A_{n+1} \rangle = -1,
\]
that is,
\[
\|A_{n+1}\|^2 - a_{n+1,n+1}^2 = -1.
\]
The \((n+1)\)th component of the vector \( Au \) is
\[
u \cdot A_{n+1} + a_{n+1,n+1}t.
\]
By Cauchy-Schwarz,
\[
(u \cdot A_{n+1})^2 \leq \|u\|^2 \|A_{n+1}\|^2,
\]
so we get,
\[
(u \cdot A_{n+1})^2 \leq \|u\|^2 \|A_{n+1}\|^2 \leq t^2 (a_{n+1,n+1}^2 - 1) = t^2 a_{n+1,n+1}^2 - t^2 < t^2 a_{n+1,n+1}^2,
\]
since \( t \neq 0 \). These calculations imply that
\[
(u \cdot A_{n+1})^2 - t^2 a_{n+1,n+1}^2 = (u \cdot A_{n+1} - ta_{n+1,n+1})(u \cdot A_{n+1} + ta_{n+1,n+1}) < 0,
\]
and that
\[
|u \cdot A_{n+1}| < |t|a_{n+1,n+1}.
\]
Note that either \((u \cdot A_{n+1} - ta_{n+1,n+1}) < 0 \) or \((u \cdot A_{n+1} + ta_{n+1,n+1}) < 0 \), but not both. If \( t < 0 \), since \( |u \cdot A_{n+1}| < |t|a_{n+1,n+1} \) and \( a_{n+1,n+1} > 0 \), then \((u \cdot A_{n+1} - ta_{n+1,n+1}) > 0 \) and \((u \cdot A_{n+1} + ta_{n+1,n+1}) < 0 \). On the other hand, if \( t > 0 \), the fact that \( |u \cdot A_{n+1}| < |t|a_{n+1,n+1} \) and \( a_{n+1,n+1} > 0 \) implies \((u \cdot A_{n+1} - ta_{n+1,n+1}) < 0 \) and \((u \cdot A_{n+1} + ta_{n+1,n+1}) > 0 \). From this it follows that \( u \cdot A_{n+1} + a_{n+1,n+1}t \) has the same sign as \( t \), since \( a_{n+1,n+1} > 0 \). Consequently, if \( a_{n+1,n+1} > 0 \), we see that \( A \) maps positive timelike (resp. lightlike) vectors to positive timelike (resp. lightlike) vectors and similarly with negative timelike (resp. lightlike) vectors.

Conversely, as \( e_{n+1} = (0, \ldots, 0, 1) \) is timelike and positive, if \( A \) preserves all positive timelike vectors, then \( Ae_{n+1} \) is timelike positive, which implies \( a_{n+1,n+1} > 0 \). \hfill \( \square \)
Let $O^+(n, 1)$ denote the subset of $O(n, 1)$ consisting of all matrices $A = (a_{ij})$ such that $a_{n+1,n+1} > 0$. Using Proposition 6.5, we can now show that $O^+(n, 1)$ is a subgroup of $O(n, 1)$ and that $SO_0(n, 1)$ is a subgroup of $SO(n, 1)$. Recall that

$$SO_0(n, 1) = \{ A \in SO(n, 1) \mid a_{n+1,n+1} > 0 \}.$$ 

Note that $SO_0(n, 1) = O^+(n, 1) \cap SO(n, 1)$.

**Proposition 6.6.** The set $O^+(n, 1)$ is a subgroup of $O(n, 1)$ and the set $SO_0(n, 1)$ is a subgroup of $SO(n, 1)$.

**Proof.** Let $A \in O^+(n, 1) \subseteq O(n, 1)$, so that $a_{n+1,n+1} > 0$. The inverse of $A$ in $O(n, 1)$ is $J A^\top J$, where

$$J = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

which implies that $a_{n+1,n+1}^{-1} = a_{n+1,n+1} > 0$, and so $A^{-1} \in O^+(n, 1)$. If $A, B \in O^+(n, 1)$, then by Proposition 6.5, both $A$ and $B$ preserve all positive timelike vectors, so $AB$ preserve all positive timelike vectors. By Proposition 6.5 again, $AB \in O^+(n, 1)$. Therefore, $O^+(n, 1)$ is a group. But then, $SO_0(n, 1) = O^+(n, 1) \cap SO(n, 1)$ is also a group.

Since any matrix $A \in SO_0(n, 1)$ preserves the Lorentz inner product and all positive timelike vectors and since $H^+_n(1)$ consists of timelike vectors, we see that every $A \in SO_0(n, 1)$ maps $H^+_n(1)$ into itself. Similarly, every $A \in SO_0(n, 1)$ maps $H^-_n(1)$ into itself. Thus, we can define an action $\cdot : SO_0(n, 1) \times H^+_n(1) \rightarrow H^+_n(1)$ by

$$A \cdot u = Au$$

and similarly, we have an action $\cdot : SO_0(n, 1) \times H^-_n(1) \rightarrow H^-_n(1)$.

**Proposition 6.7.** The group $SO_0(n, 1)$ is the subgroup of $SO(n, 1)$ that preserves $H^+_n(1)$ (and $H^-_n(1)$); that is,

$$SO_0(n, 1) = \{ A \in SO(n, 1) \mid A(H^+_n(1)) = H^+_n(1) \text{ and } A(H^-_n(1)) = H^-_n(1) \}.$$ 

**Proof.** We already observed that $A(H^+_n(1)) = H^+_n(1)$ if $A \in SO_0(n, 1)$ (and similarly, $A(H^-_n(1)) = H^-_n(1)$). Conversely, for any $A \in SO(n, 1)$ such that $A(H^+_n(1)) = H^+_n(1)$, as $e_{n+1} = (0, \ldots, 0, 1) \in H^+_n(1)$, the vector $Ae_{n+1}$ must be positive timelike, but this says that $a_{n+1,n+1} > 0$, i.e., $A \in SO_0(n, 1)$.

Next, we wish to prove that the action $SO_0(n, 1) \times H^+_n(1) \rightarrow H^+_n(1)$ is transitive. For this, we need the next two propositions.

**Proposition 6.8.** Let $u = (u, t)$ and $v = (v, s)$ be nonzero vectors in $\mathbb{R}^{n+1}$ with $\langle u, v \rangle = 0$. If $u$ is timelike, then $v$ is spacelike (i.e., $\langle v, v \rangle > 0$).
Proof. Since $u$ is timelike, we have $\|u\|^2 < t^2$, so $t \neq 0$. The condition $\langle u, v \rangle = 0$ is equivalent to $u \cdot v - ts = 0$. If $u = 0$, then $ts = 0$, and since $t \neq 0$, then $s = 0$. Then $\langle v, v \rangle = \|v\|^2 - s^2 = \|v\|^2 > 0$ since $v$ is a nonzero vector in $\mathbb{R}^{n+1}$. We now assume $u \neq 0$. In this case $u \cdot v - ts = 0$, and we get

$$\langle v, v \rangle = \|v\|^2 - s^2 = \|v\|^2 - \frac{(u \cdot v)^2}{t^2}.$$ 

But Cauchy-Schwarz implies that $(u \cdot v)^2 \leq \|u\|^2 \|v\|^2$, so we get when $u \neq 0$

$$\langle v, v \rangle = \|v\|^2 - \frac{(u \cdot v)^2}{t^2} > \|v\|^2 - \frac{(u \cdot v)^2}{\|u\|^2} \geq 0,$$

as $\|u\|^2 < t^2$. \hfill \Box

Lemma 6.8 also holds if $u = (u, t)$ is a nonzero isotropic vector and $v = (v, s)$ is a nonzero vector that is not collinear with $u$: If $\langle u, v \rangle = 0$, then $v$ is spacelike (i.e., $\langle v, v \rangle > 0$). The proof is left as an exercise to the reader.

**Proposition 6.9.** The action $\text{SO}_0(n, 1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ is transitive.

**Proof.** Let $e_{n+1} = (0, \ldots, 0, 1) \in \mathcal{H}_n^+(1)$. It is enough to prove that for every $u = (u, t) \in \mathcal{H}_n^+(1)$, there is some $A \in \text{SO}_0(n, 1)$ such that $A e_{n+1} = u$. By hypothesis,

$$\langle u, u \rangle = \|u\|^2 - t^2 = -1.$$ 

We show that we can construct an orthonormal basis, $e_1, \ldots, e_n, u$, with respect to the Lorentz inner product. Consider the hyperplane

$$H = \{v \in \mathbb{R}^{n+1} \mid \langle u, v \rangle = 0\}.$$ 

Since $u$ is timelike, by Proposition 6.8, every nonzero vector $v \in H$ is spacelike, that is $\langle v, v \rangle > 0$. Let $v_1, \ldots, v_n$ be a basis of $H$. Since all (nonzero) vectors in $H$ are spacelike, we can apply the Gram-Schmidt orthonormalization procedure and we get a basis $e_1, \ldots, e_n$ of $H$, such that

$$\langle e_i, e_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq n.$$ 

Now, by construction, we also have

$$\langle e_i, u \rangle = 0, \quad 1 \leq i \leq n, \quad \text{and} \quad \langle u, u \rangle = -1.$$ 

Therefore, $e_1, \ldots, e_n, u$ are the column vectors of a Lorentz matrix $A$ such that $A e_{n+1} = u$, proving our assertion. \hfill \Box
6.2 THE LIE ALGEBRA OF THE LORENTZ GROUP $\text{SO}_0(N, 1)$

Let us find the stabilizer of $e_{n+1} = (0, \ldots, 0, 1)$. We must have $Ae_{n+1} = e_{n+1}$, and the polar form implies that

$$A = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{with} \quad P \in \text{SO}(n).$$

Therefore, the stabilizer of $e_{n+1}$ is isomorphic to $\text{SO}(n)$, and we conclude that $H_n^+(1)$, as a homogeneous space, is

$$H_n^+(1) \cong \text{SO}_0(n, 1)/\text{SO}(n).$$

We will return to this homogeneous space in Chapter 18, and see that it is actually a symmetric space.

We end this section by showing that the Lorentz group $\text{SO}_0(n, 1)$ is connected. Firstly, it is easy to check that $\text{SO}_0(n, 1)$ and $H_n^+(1)$ satisfy the assumptions of Theorem 5.14 because they are both manifolds, although this notion has not been discussed yet (but will be in Chapter 7). Since the action $\cdot : \text{SO}_0(n, 1) \times H_n^+(1) \rightarrow H_n^+(1)$ of $\text{SO}_0(n, 1)$ on $H_n^+(1)$ is transitive, Theorem 5.14 implies that as topological spaces,

$$\text{SO}_0(n, 1)/\text{SO}(n) \cong H_n^+(1).$$

Now, we already showed that $H_n^+(1)$ is connected, so by Proposition 5.8, the connectivity of $\text{SO}_0(n, 1)$ follows from the connectivity of $\text{SO}(n)$ for $n \geq 1$. The connectivity of $\text{SO}(n)$ is a consequence of the surjectivity of the exponential map (for instance, see Gallier [72], Chapter 14) but we can also give a quick proof using Proposition 5.8. Indeed, $\text{SO}(n + 1)$ and $S^n$ are both manifolds and we saw in Section 5.2 that

$$\text{SO}(n + 1)/\text{SO}(n) \cong S^n.$$

Now, $S^n$ is connected for $n \geq 1$ and $\text{SO}(1) \cong S^1$ is connected. We finish the proof by induction on $n$.

**Corollary 6.10.** The Lorentz group $\text{SO}_0(n, 1)$ is connected; it is the component of the identity in $\text{O}(n, 1)$.

6.2 More on the Lorentz Group $\text{SO}_0(n, 1)$

In this section we take a closer look at the Lorentz group $\text{SO}_0(n, 1)$, and in particular, at the relationship between $\text{SO}_0(n, 1)$ and its Lie algebra $\mathfrak{so}(n, 1)$. The Lie algebra of $\text{SO}_0(n, 1)$ is easily determined by computing the tangent vectors to curves $t \mapsto A(t)$ on $\text{SO}_0(n, 1)$ through the identity $I$. Since $A(t)$ satisfies

$$A^\top JA = J, \quad J = I_{n, 1} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

differentiating and using the fact that $A(0) = I$, we get

$$A^\top J + JA' = 0.$$
Therefore,
\[ \mathfrak{so}(n, 1) = \{ A \in \text{Mat}_{n+1,n+1}(\mathbb{R}) \mid A^\top J + JA = 0 \}. \]

Since \( J = J^\top \), this means that \( JA \) is skew-symmetric, and so
\[ \mathfrak{so}(n, 1) = \left\{ \begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} \in \text{Mat}_{n+1,n+1}(\mathbb{R}) \mid u \in \mathbb{R}^n, \quad B^\top = -B \right\}. \]

Since \( J^2 = I \), the condition \( A^\top J + JA = 0 \) is equivalent to
\[ A^\top = -JAJ. \]

Observe that every matrix \( A \in \mathfrak{so}(n, 1) \) can be written uniquely as
\[ \begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}, \]
where the first matrix is skew-symmetric, the second one is symmetric, and both belong to \( \mathfrak{so}(n, 1) \). Thus, it is natural to define
\[ \mathfrak{e} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \mid B^\top = -B \right\}, \]
and
\[ \mathfrak{p} = \left\{ \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \mid u \in \mathbb{R}^n \right\}. \]

It is immediately verified that both \( \mathfrak{e} \) and \( \mathfrak{p} \) are subspaces of \( \mathfrak{so}(n, 1) \) (as vector spaces) and that \( \mathfrak{e} \) is a Lie subalgebra isomorphic to \( \mathfrak{so}(n) \), but \( \mathfrak{p} \) is not a Lie subalgebra of \( \mathfrak{so}(n, 1) \) because it is not closed under the Lie bracket. Still, we have
\[ [\mathfrak{e}, \mathfrak{e}] \subseteq \mathfrak{e}, \quad [\mathfrak{e}, \mathfrak{p}] \subseteq \mathfrak{p}, \quad [\mathfrak{p}, \mathfrak{p}] \subseteq \mathfrak{e}. \]

Clearly, we have the direct sum decomposition
\[ \mathfrak{so}(n, 1) = \mathfrak{e} \oplus \mathfrak{p}, \]
known as \textit{Cartan decomposition}.

There is also an automorphism of \( \mathfrak{so}(n, 1) \) known as the \textit{Cartan involution}, namely
\[ \theta(A) = -A^\top = JAJ, \]
and we see that
\[ \mathfrak{e} = \{ A \in \mathfrak{so}(n, 1) \mid \theta(A) = A \} \quad \text{and} \quad \mathfrak{p} = \{ A \in \mathfrak{so}(n, 1) \mid \theta(A) = -A \}. \]
The involution $\theta$ defined on $\mathfrak{so}(n, 1)$ is the derivative at $I$ of the involutive isomorphism $\sigma$ of the group $\mathrm{SO}_0(n, 1)$ also defined by

$$\sigma(A) = JAJ, \quad A \in \mathrm{SO}_0(n, 1).$$

To justify this claim, let $\gamma(t)$ be a curve in $\mathrm{SO}_0(n, 1)$ through $I$. Define $h(t) = \sigma \circ \gamma(t) = J\gamma(t)J$. The product rule implies $h'(0) = J\gamma'(0)J$. On the other hand, the chain rule implies $h'(0) = D\sigma_I \circ \gamma'(0)$. Combining the two equivalent forms of $h'(0)$ implies $D\sigma_I(X) = JXJ$, whenever $X \in \mathrm{SO}_0(n, 1)$.

Since the inverse of an element $A \in \mathrm{SO}_0(n, 1)$ is given by $A^{-1} = JA^\top J$, we see that $\sigma$ is also given by

$$\sigma(A) = (A^{-1})^\top.$$

Unfortunately, there does not appear to be any simple way of obtaining a formula for $\exp(A)$, where $A \in \mathfrak{so}(n, 1)$ (except for small $n$—there is such a formula for $n = 3$ due to Chris Geyer). However, it is possible to obtain an explicit formula for the matrices in $\mathfrak{p}$. This is because for such matrices $A$, if we let $\omega = \|u\| = \sqrt{u^\top u}$, we have

$$A^3 = \omega^2 A.$$

Thus, we get

**Proposition 6.11.** For every matrix $A \in \mathfrak{p}$ of the form

$$A = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix},$$

we have

$$e^A = \begin{pmatrix} I & \frac{(\cosh \omega - 1)}{\omega^2} uu^\top \\ \frac{\sinh \omega}{\omega^2} u^\top u & \cosh \omega \end{pmatrix} = \begin{pmatrix} I + \frac{\sinh^2 \omega}{\omega^2} uu^\top & \frac{\sinh \omega}{\omega^2} uu^\top \\ \frac{\sinh \omega}{\omega^2} u^\top u & \cosh \omega \end{pmatrix}.$$  

**Proof.** Using the fact that $A^3 = \omega^2 A$, we easily prove (by adjusting the calculations of Section 1.1) that

$$e^A = I + \frac{\sinh \omega}{\omega} A + \frac{\cosh \omega - 1}{\omega^2} A^2,$$

which is the first equation of the proposition, since

$$A^2 = \begin{pmatrix} uu^\top & 0 \\ 0 & u^\top u \end{pmatrix} = \begin{pmatrix} uu^\top & 0 \\ 0 & \omega^2 \end{pmatrix}.$$  

We leave as an exercise the fact that

$$\left( I + \frac{(\cosh \omega - 1)}{\omega^2} uu^\top \right)^2 = I + \frac{\sinh^2 \omega}{\omega^2} uu^\top.$$  

$\square$
Now, it clear from the above formula that each $e^B$ with $B \in \mathfrak{p}$ is a Lorentz boost. Conversely, every Lorentz boost is the exponential of some $B \in \mathfrak{p}$, as shown below.

**Proposition 6.12.** Every Lorentz boost

$$A = \begin{pmatrix} \sqrt{I + vv^T} & v \\ v^T & c \end{pmatrix},$$

with $c = \sqrt{||v||^2 + 1}$, is of the form $A = e^B$ for some $B \in \mathfrak{p}$; that is, for some $B \in \mathfrak{so}(n,1)$ of the form

$$B = \begin{pmatrix} 0 & u \\ u^T & 0 \end{pmatrix}.$$

Proof. Given

$$A = \begin{pmatrix} \sqrt{I + vv^T} & v \\ v^T & c \end{pmatrix},$$

we need to find some

$$B = \begin{pmatrix} 0 & u \\ u^T & 0 \end{pmatrix}$$

by solving the equation

$$\begin{pmatrix} \sqrt{I + \sinh^2 \omega u u^T} & \sinh \omega u \\ \sinh \omega u^T & \cosh \omega \end{pmatrix} = \begin{pmatrix} \sqrt{I + vv^T} & v \\ v^T & c \end{pmatrix},$$

with $\omega = ||u||$ and $c = \sqrt{||v||^2 + 1}$. When $v = 0$, we have $A = I$, and the matrix $B = 0$ corresponding to $u = 0$ works. So, assume $v \neq 0$. In this case, $c > 1$. We have to solve the equation $\cosh \omega = c$, that is,

$$e^{2\omega} - 2ce^{\omega} + 1 = 0.$$

The roots of the corresponding algebraic equation $X^2 - 2cX + 1 = 0$ are

$$X = c \pm \sqrt{c^2 - 1}.$$

As $c > 1$, both roots are strictly positive, so we can solve for $\omega$, say $\omega = \log(c + \sqrt{c^2 - 1}) \neq 0$. Then, $\sinh \omega \neq 0$, so we can solve the equation

$$\frac{\sinh \omega}{\omega} u = v$$

for $u$, which yields a $B \in \mathfrak{so}(n,1)$ of the right form with $A = e^B$. □

Combining Proposition 6.1 and Proposition 6.12, we have the corollary:
Corollary 6.13. Every matrix \( A \in O(n, 1) \) can be written as

\[
A = \begin{pmatrix} Q & 0 \\ 0 & \epsilon \end{pmatrix} e^{\begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}},
\]

where \( Q \in O(n) \), \( \epsilon = \pm 1 \), and \( u \in \mathbb{R}^n \).

Remarks:

(1) It is easy to show that the eigenvalues of matrices

\[
B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}
\]

are 0, with multiplicity \( n - 1 \), \( \|u\| \), and \( -\|u\| \). In particular, the eigenvalue relation

\[
\begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \lambda \begin{pmatrix} c \\ d \end{pmatrix}, \quad c \in \mathbb{R}^n, \ d, \lambda \in \mathbb{R}
\]

implies

\[
du = \lambda c, \quad u^\top c = \lambda d.
\]

If \( \lambda \neq 0 \), \( c = \frac{du}{\lambda} \), which in turn implies \( u^\top ud = \lambda^2 d \), i.e. \( \lambda^2 = u^\top u = \|u\|^2 \). If \( \lambda = 0 \), \( u^\top c = 0 \), which implies that \( c \) is in the \( n - 1 \)-dimensional hyperplane perpendicular to \( u \). Eigenvectors are then easily determined.

(2) The matrices \( B \in so(n, 1) \) of the form

\[
B = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha \\ 0 & \cdots & \alpha & 0 \end{pmatrix}
\]

are easily seen to form an abelian Lie subalgebra \( a \) of \( so(n, 1) \) (which means that for all \( B, C \in a \), \([B, C] = 0\), i.e., \( BC = CB \)). Proposition 6.11 implies that any \( B \in a \) as above, we get

\[
e^B = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh \alpha & \sinh \alpha \\ 0 & \cdots & 0 & \sinh \alpha & \cosh \alpha \end{pmatrix}.
\]
The matrices of the form $e^B$ with $B \in \mathfrak{a}$ form an abelian subgroup $A$ of $\text{SO}_0(n,1)$ isomorphic to $\text{SO}_0(1,1)$. As we already know, the matrices $B \in \mathfrak{so}(n,1)$ of the form

$$
\begin{pmatrix}
B & 0 \\
0 & 0
\end{pmatrix},
$$

where $B$ is skew-symmetric, form a Lie subalgebra $\mathfrak{k}$ of $\mathfrak{so}(n,1)$. Clearly, $\mathfrak{k}$ is isomorphic to $\mathfrak{so}(n)$. And using the exponential, we get a subgroup $K$ of $\text{SO}_0(n,1)$ isomorphic to $\text{SO}(n)$. It is also clear that $\mathfrak{k} \cap \mathfrak{a} = (0)$, but $\mathfrak{k} \oplus \mathfrak{a}$ is not equal to $\mathfrak{so}(n,1)$. What is the missing piece?

Consider the matrices $N \in \mathfrak{so}(n,1)$ of the form

$$
N = \begin{pmatrix}
0 & -u & u \\
u^\top & 0 & 0 \\
u^\top & 0 & 0
\end{pmatrix},
$$

where $u \in \mathbb{R}^{n-1}$. The reader should check that these matrices form an abelian Lie subalgebra $\mathfrak{n}$ of $\mathfrak{so}(n,1)$. Furthermore, since

$$
\mathfrak{so}(n,1) = \begin{pmatrix}
B_1 & u_1 & u \\
-u_1^\top & 0 & \alpha \\
u^\top & \alpha & 0
\end{pmatrix}
$$

$$
= \begin{pmatrix}
B_1 & u_1 + u & 0 \\
-u_1 + u^\top & 0 & 0 \\
0 & 0 & 0
\end{pmatrix} + \begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & \alpha \\
u^\top & 0 & 0
\end{pmatrix} + \begin{pmatrix}
0 & -u & u \\
u^\top & 0 & 0 \\
u^\top & \alpha & 0
\end{pmatrix},
$$

where $B_1 \in \mathfrak{so}(n-1)$, $u, u_1 \in \mathbb{R}^{n-1}$, and $\alpha \in \mathbb{R}$, we conclude that

$$
\mathfrak{so}(n,1) = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}.
$$

This is the Iwasawa decomposition of the Lie algebra $\mathfrak{so}(n,1)$. Furthermore, the reader should check that every $N \in \mathfrak{n}$ is nilpotent; in fact, $N^3 = 0$. (It turns out that $\mathfrak{n}$ is a nilpotent Lie algebra, see Knapp [106]).

The connected Lie subgroup of $\text{SO}_0(n,1)$ associated with $\mathfrak{n}$ is denoted $N$ and it can be shown that we have the Iwasawa decomposition of the Lie group $\text{SO}_0(n,1)$:

$$
\text{SO}_0(n,1) = KAN.
$$

It is easy to check that $[\mathfrak{a}, \mathfrak{n}] \subset \mathfrak{n}$, so $\mathfrak{a} \oplus \mathfrak{n}$ is a Lie subalgebra of $\mathfrak{so}(n,1)$ and $\mathfrak{n}$ is an ideal of $\mathfrak{a} \oplus \mathfrak{n}$. This implies that $N$ is normal in the group corresponding to $\mathfrak{a} \oplus \mathfrak{n}$, so $AN$ is a subgroup (in fact, solvable) of $\text{SO}_0(n,1)$. For more on the Iwasawa decomposition, see Knapp [106].
Observe that the image $\tilde{n}$ of $n$ under the Cartan involution $\theta$ is the Lie subalgebra

$$\tilde{n} = \left\{ \begin{pmatrix} 0 & u & u \\ -u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix} \mid u \in \mathbb{R}^{n-1} \right\}.$$ 

By using the Iwasawa decomposition, we can show that centralizer of $a$, namely $\{m \in so(n, 1) \mid ma = am \text{ whenever } a \in a\}$, is the Lie subalgebra

$$m = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \in \text{Mat}_{n+1,n+1}(\mathbb{R}) \mid B \in so(n - 1) \right\},$$

and hence

$$so(n, 1) = m \oplus a \oplus n \oplus \tilde{n}.$$ 

We also have

$$[m, n] \subseteq n,$$

so $m \oplus a \oplus n$ is a subalgebra of $so(n, 1)$.

The group $M$ associated with $m$ is isomorphic to $SO(n - 1)$, and it can be shown that $B = MAN$ is a subgroup of $SO_0(n, 1)$. In fact,

$$SO_0(n, 1) / (MAN) = KAN/MAN = K/M = SO(n) / SO(n - 1) = S^{n-1}.$$ 

It is customary to denote the subalgebra $m \oplus a$ by $g_0$, the algebra $n$ by $g_1$, and $\tilde{n}$ by $g_{-1}$, so that $so(n, 1) = m \oplus a \oplus n \oplus \tilde{n}$ is also written

$$so(n, 1) = g_0 \oplus g_{-1} \oplus g_1.$$ 

By the way, if $N \in n$, then

$$e^N = I + N + \frac{1}{2} N^2,$$

and since $N + \frac{1}{2} N^2$ is also nilpotent, $e^N$ can’t be diagonalized when $N \neq 0$. This provides a simple example of matrices in $SO_0(n, 1)$ that can’t be diagonalized.

Observe that Corollary 6.13 proves that every matrix $A \in SO_0(n, 1)$ can be written as

$$A = Pe^S, \quad \text{with } P \in K \cong SO(n) \text{ and } S \in p,$$

i.e.,

$$SO_0(n, 1) = K \exp(p),$$

a version of the polar decomposition for $SO_0(n, 1)$. 
Now, it is known that the exponential map \( \exp: \mathfrak{so}(n) \to \text{SO}(n) \) is surjective. So, when \( A \in \text{SO}_0(n, 1) \), since then \( Q \in \text{SO}(n) \) and \( \epsilon = +1 \), the matrix

\[
\begin{pmatrix}
Q & 0 \\
0 & 1
\end{pmatrix}
\]

is the exponential of some skew symmetric matrix

\[
C = \begin{pmatrix}
B & 0 \\
0 & 0
\end{pmatrix} \in \mathfrak{so}(n, 1),
\]

and we can write \( A = e^C e^Z \), with \( C \in \mathfrak{k} \) and \( Z \in \mathfrak{p} \). Unfortunately, \( C \) and \( Z \) generally don’t commute, so it is generally not true that \( A = e^{C+Z} \). Thus, we don’t get an “easy” proof of the surjectivity of the exponential, \( \exp: \mathfrak{so}(n, 1) \to \text{SO}_0(n, 1) \).

This is not too surprising because to the best of our knowledge, proving surjectivity for all \( n \) is not a simple matter. One proof is due to Nishikawa [137] (1983). Nishikawa’s paper is rather short, but this is misleading. Indeed, Nishikawa relies on a classic paper by Djokovic [58], which itself relies heavily on another fundamental paper by Burgoyne and Cushman [33], published in 1977. Burgoyne and Cushman determine the conjugacy classes for some linear Lie groups and their Lie algebras, where the linear groups arise from an inner product space (real or complex). This inner product is nondegenerate, symmetric, or Hermitian or skew-symmetric or skew-Hermitian. Altogether, one has to read over 40 pages to fully understand the proof of surjectivity.

In his introduction, Nishikawa states that he is not aware of any other proof of the surjectivity of the exponential for \( \text{SO}_0(n, 1) \). However, such a proof was also given by Marcel Riesz as early as 1957, in some lectures notes that he gave while visiting the University of Maryland in 1957-1958. These notes were probably not easily available until 1993, when they were published in book form, with commentaries, by Bolinder and Lounesto [145].

Interestingly, these two proofs use very different methods. The Nishikawa–Djokovic–Burgoyne and Cushman proof makes heavy use of methods in Lie groups and Lie algebra, although not far beyond linear algebra. Riesz’s proof begins with a deep study of the structure of the minimal polynomial of a Lorentz isometry (Chapter III). This is a beautiful argument that takes about 10 pages. The story is not over, as it takes most of Chapter IV (some 40 pages) to prove the surjectivity of the exponential (actually, Riesz proves other things along the way). In any case, the reader can see that both proofs are quite involved.

It is worth noting that Milnor (1969) also uses techniques very similar to those used by Riesz (in dealing with minimal polynomials of isometries) in his paper on isometries of inner product spaces [126].

What we will do to close this section is to give a relatively simple proof that the exponential map \( \exp: \mathfrak{so}(1, 3) \to \text{SO}_0(1, 3) \) is surjective.

In the case of \( \text{SO}_0(1, 3) \), we can use the fact that \( \text{SL}(2, \mathbb{C}) \) is a two-sheeted covering space of \( \text{SO}_0(1, 3) \), which means that there is a homomorphism \( \phi: \text{SL}(2, \mathbb{C}) \to \text{SO}_0(1, 3) \)
which is surjective and that $\text{Ker } \phi = \{-I, I\}$. Then, the small miracle is that, although the exponential $\exp: \mathfrak{sl}(2, \mathbb{C}) \to \text{SL}(2, \mathbb{C})$ is not surjective, for every $A \in \text{SL}(2, \mathbb{C})$, either $A$ or $-A$ is in the image of the exponential!

**Proposition 6.14.** Given any matrix

\[
B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2, \mathbb{C}),
\]

let $\omega$ be any of the two complex roots of $a^2 + bc$. If $\omega \neq 0$, then

\[
\begin{align*}
\exp(B) &= \cosh \omega I + \frac{\sinh \omega}{\omega} B, \\
\omega^2 &= (a^2 + bc) I.
\end{align*}
\]

and $\exp(B) = I + B$ if $a^2 + bc = 0$. Furthermore, every matrix $A \in \text{SL}(2, \mathbb{C})$ is in the image of the exponential map, unless $A = -I + N$, where $N$ is a nonzero nilpotent (i.e., $N^2 = 0$ with $N \neq 0$). Consequently, for any $A \in \text{SL}(2, \mathbb{C})$, either $A$ or $-A$ is of the form $\exp(B)$, for some $B \in \mathfrak{sl}(2, \mathbb{C})$.

**Proof.** Observe that

\[
B^2 = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \begin{pmatrix} a & b \\ c & -a \end{pmatrix} = (a^2 + bc) I.
\]

Then, it is straightforward to prove that

\[
\exp(B) = \cosh \omega I + \frac{\sinh \omega}{\omega} B,
\]

where $\omega$ is a square root of $a^2 + bc$ is $\omega \neq 0$, otherwise, $\exp(B) = I + B$.

Let

\[
A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \quad \alpha \delta - \gamma \beta = 1
\]

be any matrix in $\text{SL}(2, \mathbb{C})$. We would like to find a matrix $B \in \mathfrak{sl}(2, \mathbb{C})$ so that $A = \exp(B)$. In view of the above, we need to solve the system

\[
\begin{align*}
cosh \omega + \frac{\sinh \omega}{\omega} a &= \alpha \\
cosh \omega - \frac{\sinh \omega}{\omega} a &= \delta \\
\frac{\sinh \omega}{\omega} b &= \beta \\
\frac{\sinh \omega}{\omega} c &= \gamma
\end{align*}
\]

for $a, b, c, \text{ and } \omega$. From the first two equations, we get

\[
\begin{align*}
cosh \omega &= \frac{\alpha + \delta}{2} \\
\sinh \omega \frac{a}{\omega} &= \frac{\alpha - \delta}{2}.
\end{align*}
\]
Thus, we see that we need to know whether complex cosh is surjective and when complex sinh is zero. We claim:

1. cosh is surjective.

2. sinh $z = 0$ iff $z = n\pi i$, where $n \in \mathbb{Z}$.

Given any $c \in \mathbb{C}$, we have $\cosh \omega = c$ iff

$$e^{2\omega} - 2e^{\omega}c + 1 = 0.$$  

The corresponding algebraic equation

$$Z^2 - 2cZ + 1 = 0$$

has discriminant $4(c^2 - 1)$ and it has two complex roots

$$Z = c \pm \sqrt{c^2 - 1}$$

where $\sqrt{c^2 - 1}$ is some square root of $c^2 - 1$. Observe that these roots are never zero. Therefore, we can find a complex log of $c + \sqrt{c^2 - 1}$, say $\omega$, so that $e^{\omega} = c + \sqrt{c^2 - 1}$ is a solution of $e^{2\omega} - 2e^{\omega}c + 1 = 0$. This proves the surjectivity of cosh.

We have sinh $\omega = 0$ iff $e^{2\omega} = 1$; this holds iff $2\omega = n2\pi i$, i.e., $\omega = n\pi i$.

Observe that

$$\frac{\sinh n\pi i}{n\pi i} = 0 \quad \text{if } n \neq 0, \text{ but } \frac{\sinh n\pi i}{n\pi i} = 1 \quad \text{when } n = 0.$$  

We know that

$$\cosh \omega = \frac{\alpha + \delta}{2}$$

can always be solved.

Case 1. If $\omega \neq n\pi i$, with $n \neq 0$, then

$$\frac{\sinh \omega}{\omega} \neq 0$$

and the other equations can also be solved (this includes the case $\omega = 0$). We still have to check that

$$a^2 + bc = \omega^2.$$
This is because, using the fact that \( \cosh \omega = \frac{\alpha + \delta}{2} \), \( \alpha \delta - \beta \gamma = 1 \), and \( \cosh^2 \omega - \sinh^2 \omega = 1 \), we have

\[
\begin{align*}
a^2 + bc &= \frac{(\alpha - \delta)^2 \omega^2}{4 \sinh^2 \omega} + \frac{\beta \gamma \omega^2}{\sinh^2 \omega} \\
&= \frac{\omega^2(\alpha^2 + \delta^2 - 2\alpha \delta + 4\beta \gamma)}{4 \sinh^2 \omega} \\
&= \frac{\omega^2((\alpha + \delta)^2 - 4(\alpha \delta - \beta \gamma))}{4 \sinh^2 \omega} \\
&= \frac{4\omega^2(\cosh^2 \omega - 1)}{4 \sinh^2 \omega} \\
&= \omega^2.
\end{align*}
\]

Therefore, in this case, the exponential is surjective. It remains to examine the other case.

Case 2. Assume \( \omega = n \pi i \), with \( n \neq 0 \). If \( n \) is even, then \( e^\omega = 1 \), which implies

\[ \alpha + \delta = 2. \]

However, \( \alpha \delta - \beta \gamma = 1 \) (since \( A \in \text{SL}(2, \mathbb{C}) \)), so from the facts that \( \det(A) \) is the product of the eigenvalues and \( \text{tr}(A) \) is the sum of the eigenvalues, we deduce that \( A \) has the double eigenvalue 1. Thus, \( N = A - I \) is nilpotent (i.e., \( N^2 = 0 \)) and has zero trace; but then, \( N \in \text{sl}(2, \mathbb{C}) \) and

\[ e^N = I + N = I + A - I = A. \]

If \( n \) is odd, then \( e^\omega = -1 \), which implies

\[ \alpha + \delta = -2. \]

In this case, \( A \) has the double eigenvalue \(-1\) and \( A + I = N \) is nilpotent. So \( A = -I + N \), where \( N \) is nilpotent. If \( N \neq 0 \), then \( A \) cannot be diagonalized. We claim that there is no \( B \in \text{sl}(2, \mathbb{C}) \) so that \( e^B = A \).

Indeed, any matrix \( B \in \text{sl}(2, \mathbb{C}) \) has zero trace, which means that if \( \lambda_1 \) and \( \lambda_2 \) are the eigenvalues of \( B \), then \( \lambda_1 = -\lambda_2 \). If \( \lambda_1 \neq 0 \), then \( \lambda_1 \neq \lambda_2 \) so \( B \) can be diagonalized, but then \( e^B \) can also be diagonalized, contradicting the fact that \( A \) can’t be diagonalized. If \( \lambda_1 = \lambda_2 = 0 \), then \( e^B \) has the double eigenvalue +1, but \( A \) has eigenvalues \(-1\). Therefore, the only matrices \( A \in \text{SL}(2, \mathbb{C}) \) that are not in the image of the exponential are those of the form \( A = -I + N \), where \( N \) is a nonzero nilpotent. However, note that \(-A = I - N \) \textit{is} in the image of the exponential.

\[ \square \]

**Remark:** If we restrict our attention to \( \text{SL}(2, \mathbb{R}) \), then we have the following proposition that can be used to prove that the exponential map \( \exp : \text{so}(1, 2) \to \text{SO}_0(1, 2) \) is surjective:
Proposition 6.15. Given any matrix
\[
B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2, \mathbb{R}),
\]
if \(a^2 + bc > 0\), then let \(\omega = \sqrt{a^2 + bc} > 0\), and if \(a^2 + bc < 0\), then let \(\omega = \sqrt{-(a^2 + bc)} > 0\) (i.e., \(\omega^2 = -(a^2 + bc)\)). In the first case \((a^2 + bc > 0)\), we have
\[
e^B = \cosh \omega I + \frac{\sinh \omega}{\omega} B,
\]
and in the second case \((a^2 + bc < 0)\), we have
\[
e^B = \cos \omega I + \frac{\sin \omega}{\omega} B.
\]
If \(a^2 + bc = 0\), then \(e^B = I + B\). Furthermore, every matrix \(A \in \text{SL}(2, \mathbb{R})\) whose trace satisfies \(\text{tr}(A) \geq -2\) is in the image of the exponential map, unless \(A = -I + N\) with \(N \neq 0\) nilpotent. Consequently, for any \(A \in \text{SL}(2, \mathbb{R})\), either \(A\) or \(-A\) is of the form \(e^B\), for some \(B \in \mathfrak{sl}(2, \mathbb{R})\).

Proof. For any matrix
\[
B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2, \mathbb{R}),
\]
some simple calculations show that if \(a^2 + bc > 0\), then
\[
e^B = \cosh \omega I + \frac{\sinh \omega}{\omega} B
\]
with \(\omega = \sqrt{a^2 + bc} > 0\), and if \(a^2 + bc < 0\), then
\[
e^B = \cos \omega I + \frac{\sin \omega}{\omega} B
\]
with \(\omega = \sqrt{-(a^2 + bc)} > 0\) (and \(e^B = I + B\) when \(a^2 + bc = 0\)). Let
\[
A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \quad \alpha \delta - \beta \gamma = 1
\]
be any matrix in \(\text{SL}(2, \mathbb{R})\).

First, assume that \(\text{tr}(A) = \alpha + \delta > 2\). We would like to find a matrix \(B \in \mathfrak{sl}(2, \mathbb{R})\) so that \(A = e^B\). In view of the above, we need to solve the system
\[
cosh \omega + \frac{\sinh \omega}{\omega} a = \alpha \\
cosh \omega - \frac{\sinh \omega}{\omega} a = \delta \\
\frac{\sinh \omega}{\omega} b = \beta \\
\frac{\sinh \omega}{\omega} c = \gamma
\]
for $a, b, c,$ and $\omega$. From the first two equations we get
\[
\cosh \omega = \frac{\alpha + \delta}{2}, \\
\sinh \omega a = \frac{\alpha - \delta}{2}.
\]
As in the proof of Proposition 6.14, $\cosh \omega = c$ iff $e^c$ is a root of the quadratic equation
\[
Z^2 - 2cZ + 1 = 0.
\]
This equation has a real roots iff $c^2 \geq 1$, and since $c = \frac{\alpha + \delta}{2}$ and $\alpha + \delta > 2$, our equation has real roots. Furthermore, the root $c + \sqrt{c^2 - 1}$ is greater than 1, so $\log c$ is a positive real number. Then, as in the proof of Proposition 6.14, we find solutions of our system above. Moreover, these solutions are real and satisfy $a^2 + bc = \omega^2$.

Let us now consider the case where $-2 \leq \alpha + \delta \leq 2$. This time we try to solve the system
\[
\cos \omega + \frac{\sin \omega}{\omega} a = \alpha, \\
\cos \omega - \frac{\sin \omega}{\omega} a = \delta, \\
\frac{\sin \omega}{\omega} b = \beta, \\
\frac{\sin \omega}{\omega} c = \gamma.
\]
We get
\[
\cos \omega = \frac{\alpha + \delta}{2}, \\
\sinh \omega a = \frac{\alpha - \delta}{2}.
\]
Because $-2 \leq \alpha + \delta \leq 2$, the first equation has (real) solutions, and we may assume that $0 \leq \omega \leq \pi$.

If $\omega = 0$ is a solution, then $\alpha + \beta = 2$ and we already know via the arguments of Proposition 6.14 that $N = A - I$ is nilpotent and that $e^N = I + N = A$. If $\omega = \pi$, then $\alpha + \beta = -2$ and we know that $N = A + I$ is nilpotent. If $N = 0$, then $A = -I$, and otherwise we already know that $A = -I + N$ is not in the image of the exponential.

If $0 < \omega < \pi$, then $\sin \omega \neq 0$ and the other equations have a solution. We still need to check that
\[
a^2 + bc = -\omega^2.
\]
Because $\cos \omega = \frac{\alpha + \delta}{2}$, $\alpha \delta - \beta \gamma = 1$ and $\cos^2 \omega + \sin^2 \omega = 1$, we have

$$a^2 + bc = \frac{(\alpha - \delta)^2 \omega^2}{4 \sin^2 \omega} + \frac{\beta \gamma \omega^2}{\sin^2 \omega}$$

$$= \frac{\omega^2 (\alpha^2 + \delta^2 - 2 \alpha \delta + 4 \beta \gamma)}{4 \sinh^2 \omega}$$

$$= \frac{\omega^2 ((\alpha + \delta)^2 - 4(\alpha \delta - \beta \gamma))}{4 \sin^2 \omega}$$

$$= \frac{4 \omega^2 (\cos^2 \omega - 1)}{4 \sin^2 \omega}$$

$$= -\omega^2.$$

This proves that every matrix $A \in \text{SL}(2, \mathbb{R})$ whose trace satisfies $\text{tr}(A) \geq -2$ is in the image of the exponential map, unless $A = -I + N$ with $N \neq 0$ nilpotent.

We now return to the relationship between $\text{SL}(2, \mathbb{C})$ and $\text{SO}_0(1, 3)$. In order to define a homomorphism $\phi: \text{SL}(2, \mathbb{C}) \to \text{SO}_0(1, 3)$, we begin by defining a linear bijection $h$ between $\mathbb{R}^4$ and $\text{H}(2)$, the set of complex $2 \times 2$ Hermitian matrices, by

$$(t, x, y, z) \mapsto \begin{pmatrix} t + x & y - i z \\ y + i z & t - x \end{pmatrix}.$$  

Those familiar with quantum physics will recognize a linear combination of the Pauli matrices! The inverse map is easily defined For instance, given a Hermitian matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a, d \in \mathbb{R}, \quad c = \overline{b} \in \mathbb{C}$$

by setting

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} t + x & y - i z \\ y + i z & t - x \end{pmatrix},$$

we find that have

$$t = \frac{a + d}{2}, \quad x = \frac{a - d}{2}, \quad y = \frac{b + \overline{b}}{2}, \quad z = \frac{b - \overline{b}}{2i}.$$  

For any $A \in \text{SL}(2, \mathbb{C})$, we define a map $l_A: \text{H}(2) \to \text{H}(2)$, via

$$S \mapsto ASA^*.$$  

(Here, $A^* = \overline{A}^T$.) Using the linear bijection $h: \mathbb{R}^4 \to \text{H}(2)$ and its inverse, we obtain a map $\text{lor}_A: \mathbb{R}^4 \to \mathbb{R}^4$, where

$$\text{lor}_A = h^{-1} \circ l_A \circ h.$$
As $A^*A$ is Hermitian, we see that $l_A$ is well defined. It is obviously linear and since $\det(A) = 1$ (recall, $A \in \text{SL}(2, \mathbb{C})$) and
\[
\det\begin{pmatrix} t + x & y - iz \\ y + iz & t - x \end{pmatrix} = t^2 - x^2 - y^2 - z^2,
\]
we see that lor$_A$ preserves the Lorentz metric! Furthermore, it is not hard to prove that $\text{SL}(2, \mathbb{C})$ is connected (use the polar form or analyze the eigenvalues of a matrix in $\text{SL}(2, \mathbb{C})$, for example, as in Duistermatt and Kolk [64] (Chapter 1, Section 1.2)) and that the map $\phi: \text{SL}(2, \mathbb{C}) \to \text{GL}(4, \mathbb{R})$ with
\[
\phi: A \mapsto \text{lor}_A
\]
is a continuous group homomorphism. Thus, the range of $\phi$ is a connected subgroup of $\text{SO}_0(1,3)$. This shows that $\phi: \text{SL}(2, \mathbb{C}) \to \text{SO}_0(1,3)$ is indeed a homomorphism. It remains to prove that it is surjective and that its kernel is $\{I, -I\}$.

**Proposition 6.16.** The homomorphism $\phi: \text{SL}(2, \mathbb{C}) \to \text{SO}_0(1,3)$ is surjective and its kernel is $\{I, -I\}$.

**Proof.** Recall that from Theorem 6.4, the Lorentz group $\text{SO}_0(1,3)$ is generated by the matrices of the form
\[
\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \text{SO}(3)
\]
and the matrices of the form
\[
\begin{pmatrix} \cosh \alpha & \sinh \alpha & 0 & 0 \\ \sinh \alpha & \cosh \alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
\]
Thus, to prove the surjectivity of $\phi$, it is enough to check that the above matrices are in the range of $\phi$. For matrices of the second kind
\[
A = \begin{pmatrix} e^{\frac{\alpha}{2}} & 0 \\ 0 & e^{-\frac{\alpha}{2}} \end{pmatrix}
\]
does the job. Let $e_1, e_2, e_3$, and $e_4$ be the standard basis for $\mathbb{R}^4$. Then
\[
\text{lor}_A(e_1) = h^{-1} \circ l_A \circ h(e_1) = h^{-1} \circ l_A \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = h^{-1} \begin{pmatrix} e^\alpha \\ 0 \\ 0 \\ e^{-\alpha} \end{pmatrix} = \frac{e^\alpha + e^{-\alpha}}{2}, \frac{e^\alpha - e^{-\alpha}}{2}, 0, 0,
\]

\[
= (\cosh \alpha, \sinh \alpha, 0, 0).
\]
Similar calculations show that
\[
\text{lor}_A(e_2) = (\sinh \alpha, \cosh \alpha, 0, 0) \\
\text{lor}_A(e_3) = (0, 0, 1, 0) \\
\text{lor}_A(e_4) = (0, 0, 0, 1).
\]
For matrices of the first kind, we recall that the group of unit quaternions \( q = a1 + bi + cj + dk \) can be viewed as \( \text{SU}(2) \), via the correspondence
\[
a1 + bi + cj + dk \mapsto \begin{pmatrix} a + ib & c + id \\ -c + id & a - ib \end{pmatrix},
\]
where \( a, b, c, d \in \mathbb{R} \) and \( a^2 + b^2 + c^2 + d^2 = 1 \). Moreover, the algebra of quaternions \( \mathbb{H} \) is the real algebra of matrices as above, without the restriction \( a^2 + b^2 + c^2 + d^2 = 1 \), and \( \mathbb{R}^3 \) is embedded in \( \mathbb{H} \) as the pure quaternions, i.e., those for which \( a = 0 \). Observe that when \( a = 0 \),
\[
\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix} = i \begin{pmatrix} b & d - ic \\ d + ic & -b \end{pmatrix} = ih(0, b, d, c).
\]
Therefore, we have a bijection between the pure quaternions and the subspace of the Hermitian matrices
\[
\begin{pmatrix} b & d - ic \\ d + ic & -b \end{pmatrix}
\]
for which \( a = 0 \), the inverse being division by \( i \), i.e., multiplication by \( -i \). Also, when \( q \) is a unit quaternion, let \( \overline{q} = a1 - bi - cj - dk \), and observe that \( \overline{q} = q^{-1} \). Using the embedding \( \mathbb{R}^3 \hookrightarrow \mathbb{H} \), for every unit quaternion \( q \in \text{SU}(2) \), define the map \( \rho_q : \mathbb{R}^3 \to \mathbb{R}^3 \) by
\[
\rho_q(X) = qX\overline{q} = qXq^{-1},
\]
for all \( X \in \mathbb{R}^3 \hookrightarrow \mathbb{H} \). Then, it is well known that \( \rho_q \) is a rotation (i.e., \( \rho_q \in \text{SO}(3) \)), and moreover the map \( q \mapsto \rho_q \) is a surjective homomorphism \( \rho : \text{SU}(2) \to \text{SO}(3) \), and \( \text{Ker } \phi = \{ I, -I \} \) (For example, see Gallier [72], Chapter 8).

Now, consider a matrix \( A \) of the form
\[
\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix}
\]
with \( P \in \text{SO}(3) \).
We claim that we can find a matrix \( B \in \text{SL}(2, \mathbb{C}) \), such that \( \phi(B) = \text{lor}_B = A \). We claim that we can pick \( B \in \text{SU}(2) \subseteq \text{SL}(2, \mathbb{C}) \). Indeed, if \( B \in \text{SU}(2) \), then \( B^* = B^{-1} \), so
\[
B \begin{pmatrix} t + x & y - iz \\ y + iz & t - x \end{pmatrix} B^* = t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - iB \begin{pmatrix} ix & z + iy \\ -z + iy & -ix \end{pmatrix} B^{-1}.
\]
The above shows that \( \text{lor}_B \) leaves the coordinate \( t \) invariant. The term
\[
B \begin{pmatrix} ix & z + iy \\ -z + iy & -ix \end{pmatrix} B^{-1}
\]
is a pure quaternion corresponding to the application of the rotation $\rho_B$ induced by the quaternion $B$ to the pure quaternion associated with $(x, y, z)$ and multiplication by $-i$ is just the corresponding Hermitian matrix, as explained above. But, we know that for any $P \in \text{SO}(3)$, there is a quaternion $B$ so that $\rho_B = P$, so we can find our $B \in \text{SU}(2)$ so that
\[
\text{lor}_B = \begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} = A.
\]

Finally, assume that $\phi(A) = I$. This means that
\[
ASA^* = S,
\]
for all Hermitian matrices $S$ defined above. In particular, for $S = I$, we get $AA^* = I$, i.e., $A \in \text{SU}(2)$. Thus
\[
AS = SA
\]
for all Hermitian matrices $S$ defined above, so in particular, this holds for diagonal matrices of the form
\[
\begin{pmatrix} t + x & 0 \\ 0 & t - x \end{pmatrix},
\]
with $t + x \neq t - x$. We deduce that $A$ is a diagonal matrix, and since it is unitary, we must have $A = \pm I$. Therefore, $\text{Ker} \phi = \{I, -I\}$.

**Remark:** The group $\text{SL}(2, \mathbb{C})$ is isomorphic to the group $\text{Spin}(1, 3)$, which is a (simply-connected) double-cover of $\text{SO}_0(1, 3)$. This is a standard result of Clifford algebra theory; see Bröcker and tom Dieck [31] or Fulton and Harris [70]. What we just did is to provide a direct proof of this fact.

We just proved that there is an isomorphism
\[
\text{SL}(2, \mathbb{C})/\{I, -I\} \cong \text{SO}_0(1, 3).
\]
However, the reader may recall that $\text{SL}(2, \mathbb{C})/\{I, -I\} = \text{PSL}(2, \mathbb{C}) \cong \text{Möb}^+$. Therefore, the Lorentz group is isomorphic to the Möbius group.

We now have all the tools to prove that the exponential map $\exp: \mathfrak{so}(1, 3) \to \text{SO}_0(1, 3)$ is surjective.

**Theorem 6.17.** The exponential map $\exp: \mathfrak{so}(1, 3) \to \text{SO}_0(1, 3)$ is surjective.

**Proof.** First, recall from Proposition 4.13 that the following diagram commutes:
\[
\begin{array}{ccc}
\text{SL}(2, \mathbb{C}) & \xrightarrow{\phi} & \text{SO}_0(1, 3) \\
\exp \downarrow & & \exp \\
\mathfrak{sl}(2, \mathbb{C}) & \xrightarrow{d\phi_1} & \mathfrak{so}(1, 3)
\end{array}
\]
Pick any \( A \in \text{SO}_0(1, 3) \). By Proposition 6.16, the homomorphism \( \phi \) is surjective and as \( \text{Ker} \phi = \{ I, -I \} \), there exists some \( B \in \text{SL}(2, \mathbb{C}) \) so that
\[
\phi(B) = \phi(-B) = A.
\]
Now, by Proposition 6.14, for any \( B \in \text{SL}(2, \mathbb{C}) \), either \( B \) or \( -B \) is of the form \( e^C \), for some \( C \in \mathfrak{sl}(2, \mathbb{C}) \). By the commutativity of the diagram, if we let \( D = d\phi_1(C) \in \mathfrak{so}(1, 3) \), we get
\[
A = \phi(\pm e^C) = e^{d\phi_1(C)} = e^D,
\]
with \( D \in \mathfrak{so}(1, 3) \), as required.

**Remark:** We can restrict the bijection \( h : \mathbb{R}^4 \to \text{H}(2) \) defined earlier to a bijection between \( \mathbb{R}^3 \) and the space of real symmetric matrices of the form
\[
\begin{pmatrix}
t + x & y \\
y & t - x
\end{pmatrix}.
\]
Then, if we also restrict ourselves to \( \text{SL}(2, \mathbb{R}) \), for any \( A \in \text{SL}(2, \mathbb{R}) \) and any symmetric matrix \( S \) as above, we get a map
\[
S \mapsto ASA^\top.
\]
The reader should check that these transformations correspond to isometries in \( \text{SO}_0(1, 2) \) and we get a homomorphism \( \phi : \text{SL}(2, \mathbb{R}) \to \text{SO}_0(1, 2) \). Just as \( \text{SL}(2, \mathbb{C}) \) is connected, the group \( \text{SL}(2, \mathbb{R}) \) is also connected (but not simply connected, unlike \( \text{SL}(2, \mathbb{C}) \)). Then, we have a version of Proposition 6.16 for \( \text{SL}(2, \mathbb{R}) \) and \( \text{SO}_0(1, 2) \):

**Proposition 6.18.** The homomorphism \( \phi : \text{SL}(2, \mathbb{R}) \to \text{SO}_0(1, 2) \) is surjective and its kernel is \( \{ I, -I \} \).

Using Proposition 6.18, Proposition 6.15, and the commutative diagram
\[
\begin{array}{ccc}
\text{SL}(2, \mathbb{R}) & \xrightarrow{\phi} & \text{SO}_0(1, 2) \\
\exp & & \exp \\
\text{sl}(2, \mathbb{R}) & \xrightarrow{d\phi_1} & \mathfrak{so}(1, 2)
\end{array}
\]
we get a version of Theorem 6.17 for \( \text{SO}_0(1, 2) \):

**Theorem 6.19.** The exponential map \( \exp : \mathfrak{so}(1, 2) \to \text{SO}_0(1, 2) \) is surjective.

Also observe that \( \text{SO}_0(1, 1) \) consists of the matrices of the form
\[
A = \begin{pmatrix}
\cosh \alpha & \sinh \alpha \\
\sinh \alpha & \cosh \alpha
\end{pmatrix},
\]
and a direct computation shows that
\[
e^{\begin{pmatrix} 0 & \alpha \\ \alpha & 0 \end{pmatrix}} = \begin{pmatrix} \cosh \alpha & \sinh \alpha \\ \sinh \alpha & \cosh \alpha \end{pmatrix}.
\]

Thus, we see that the map \( \exp : \mathfrak{so}(1, 1) \to \mathrm{SO}_0(1, 1) \) is also surjective. Therefore, we have proved that \( \exp : \mathfrak{so}(1, n) \to \mathrm{SO}_0(1, n) \) is surjective for \( n = 1, 2, 3 \). This actually holds for all \( n \geq 1 \), but the proof is much more involved, as we already discussed earlier.

### 6.3 Polar Forms for Matrices in \( \mathbf{O}(p, q) \)

Recall from Section 6.1 that the group \( \mathbf{O}(p, q) \) is the set of all \( n \times n \)-matrices
\[
\mathbf{O}(p, q) = \{ A \in \mathrm{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q} \}.
\]
We deduce immediately that \( |\det(A)| = 1 \), and we also know that \( AI_{p,q} A^\top = I_{p,q} \) holds. Unfortunately, when \( p \neq 0, 1 \) and \( q \neq 0, 1 \), it does not seem possible to obtain a formula as nice as that given in Proposition 6.1. Nevertheless, we can obtain a formula for a polar form factorization of matrices in \( \mathbf{O}(p, q) \).

First, recall (for example, see Gallier [72], Chapter 12) that if \( S \) is a symmetric positive definite matrix, then there is a unique symmetric positive definite matrix, \( T \), so that
\[
S = T^2.
\]

We denote \( T \) by \( S^{\frac{1}{2}} \) or \( \sqrt{S} \). By \( S^{-\frac{1}{2}} \), we mean the inverse of \( S^{\frac{1}{2}} \). In order to obtain the polar form of a matrix in \( \mathbf{O}(p, q) \), we begin with the following proposition:

**Proposition 6.20.** Every matrix \( X \in \mathbf{O}(p, q) \) can be written as
\[
X = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix},
\]
where \( \alpha = (I_p - Z^\top Z)^{-1} \) and \( \delta = (I_q - ZZ^\top)^{-1} \), for some orthogonal matrices \( U \in \mathbf{O}(p) \), \( V \in \mathbf{O}(q) \) and for some \( q \times p \) matrix, \( Z \), such that \( I_p - Z^\top Z \) and \( I_q - ZZ^\top \) are symmetric positive definite matrices. Moreover, \( U, V, Z \) are uniquely determined by \( X \).

**Proof.** If we write
\[
X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},
\]

with $A$ a $p \times p$ matrix, $D$ a $q \times q$ matrix, $B$ a $p \times q$ matrix and $C$ a $q \times p$ matrix, then the equations $A^\top I_{p,q}A = I_{p,q}$ and $AI_{p,q}A^\top = I_{p,q}$ yield the (not independent) conditions

\[
\begin{align*}
A^\top A &= I_p + C^\top C \\
D^\top D &= I_q + B^\top B \\
A^\top B &= C^\top D \\
AA^\top &= I_p + BB^\top \\
DD^\top &= I_q + CC^\top \\
AC^\top &= BD^\top.
\end{align*}
\]

Since $C^\top C$ is symmetric and since

\[x^\top C^\top Cx = \|Cx\|^2 \geq 0,
\]

we see that $C^\top C$ is a positive semi-definite matrix with nonnegative eigenvalues. We then deduce that $A^\top A$ is symmetric positive definite and similarly for $D^\top D$. If we assume that the above decomposition of $X$ holds, we deduce that

\[
\begin{align*}
A &= U(I_p - Z^\top Z)^{-\frac{1}{2}} \\
B &= U(I_p - Z^\top Z)^{-\frac{1}{2}}Z^\top \\
C &= V(I_q - ZZ^\top)^{-\frac{1}{2}}Z \\
D &= V(I_q - ZZ^\top)^{-\frac{1}{2}},
\end{align*}
\]

which implies

\[
Z = D^{-1}C \quad \text{and} \quad Z^\top = A^{-1}B.
\]

Thus, we must check that

\[(D^{-1}C)^\top = A^{-1}B\]

i.e.,

\[C^\top (D^\top)^{-1} = A^{-1}B,
\]

namely,

\[AC^\top = BD^\top,
\]

which is indeed the last of our identities. Thus, we must have $Z = D^{-1}C = (A^{-1}B)^\top$. The above expressions for $A$ and $D$ also imply that

\[
A^\top A = (I_p - Z^\top Z)^{-1} \quad \text{and} \quad D^\top D = (I_q - ZZ^\top)^{-1},
\]

so we must check that the choice $Z = D^{-1}C = (A^{-1}B)^\top$ yields the above equations.
Since \( Z^\top = A^{-1}B \), we have
\[
Z^\top Z = A^{-1}BB^\top (A^\top)^{-1} = A^{-1}(AA^\top - I_p)(A^\top)^{-1}, \quad \text{since } AA^\top = I_p + BB^\top \\
= I_p - A^{-1}(A^\top)^{-1} = I_p - (A^\top A)^{-1}.
\]
Therefore,
\[
(A^\top A)^{-1} = I_p - Z^\top Z,
\]
i.e.,
\[
A^\top A = (I_p - Z^\top Z)^{-1},
\]
as desired. We also have, this time, with \( Z = D^{-1}C \),
\[
ZZ^\top = D^{-1}CC^\top (D^\top)^{-1} = D^{-1}(DD^\top - I_q)(D^\top)^{-1}, \quad \text{since } DD^\top = I_p + CC^\top \\
= I_q - D^{-1}(D^\top)^{-1} = I_q - (D^\top D)^{-1}.
\]
Therefore,
\[
(D^\top D)^{-1} = I_q - ZZ^\top,
\]
i.e.,
\[
D^\top D = (I_q - ZZ^\top)^{-1},
\]
as desired. Now, since \( A^\top A \) and \( D^\top D \) are positive definite, the polar form implies that
\[
A = U(A^\top A)^{\frac{1}{2}} = U(I_p - Z^\top Z)^{-\frac{1}{2}}
\]
and
\[
D = V(D^\top D)^{\frac{1}{2}} = V(I_q - ZZ^\top)^{-\frac{1}{2}},
\]
for some unique matrices, \( U \in O(p) \) and \( V \in O(q) \). Since \( Z = D^{-1}C \) and \( Z^\top = A^{-1}B \), we get \( C = DZ \) and \( B = AZ^\top \), but this is
\[
B = U(I_p - Z^\top Z)^{-\frac{1}{2}}Z^\top \\
C = V(I_q - ZZ^\top)^{-\frac{1}{2}}Z,
\]
as required. Therefore, the unique choice of \( Z = D^{-1}C = (A^{-1}B)^\top \), \( U \) and \( V \) does yield the formula of the proposition.

We next show that the matrix
\[
\begin{pmatrix}
\alpha^\frac{1}{2} & \alpha^\frac{1}{2}Z^\top \\
\delta^\frac{1}{2}Z & \delta^\frac{1}{2}
\end{pmatrix}
= \begin{pmatrix}
(I_p - Z^\top Z)^{-\frac{1}{2}} & (I_p - Z^\top Z)^{-\frac{1}{2}}Z^\top \\
(I_q - ZZ^\top)^{-\frac{1}{2}}Z & (I_q - ZZ^\top)^{-\frac{1}{2}}
\end{pmatrix}
\]
is symmetric. To prove this we use power series.
**Proposition 6.21.** For any \( q \times p \) matrix \( Z \) such that \( I_p - Z^\top Z \) and \( I_q - ZZ^\top \) are symmetric positive definite, the matrix

\[
S = \begin{pmatrix}
\alpha \frac{1}{2} \\
\delta \frac{1}{2} Z \\
\alpha \frac{1}{2} Z^\top \\
\delta \frac{1}{2}
\end{pmatrix}
\]

is symmetric, where \( \alpha = (I_p - Z^\top Z)^{-1} \) and \( \delta = (I_q - ZZ^\top)^{-1} \).

**Proof.** The matrix \( S \) is symmetric iff \( Z \alpha \frac{1}{2} = \delta \frac{1}{2} Z \), that is iff \( Z(I_p - Z^\top Z)^{-\frac{1}{2}} = (I_q - ZZ^\top)^{-\frac{1}{2}} Z \)

\[
(I_q - ZZ^\top)\frac{1}{2} Z = Z(I_p - Z^\top Z)\frac{1}{2}.
\]

If \( Z = 0 \), the equation holds trivially. If \( Z \neq 0 \), we know from linear algebra that \( ZZ^\top \) and \( Z^\top Z \) are symmetric positive semidefinite, and they have the same positive eigenvalues. Thus \( I_p - Z^\top Z \) is positive definite iff \( I_q - ZZ^\top \) is positive definite, and if so, we must have \( \rho(ZZ^\top) = \rho(Z^\top Z) < 1 \) (where \( \rho(ZZ^\top) \) denotes the largest modulus of the eigenvalues of \( ZZ^\top \), in this case, since the eigenvalues of \( ZZ^\top \) are nonnegative, this is the largest eigenvalue of \( ZZ^\top \)). If we use the spectral norm \( \| \cdot \| \) (the operator norm induced by the 2-norm), we have

\[
\| ZZ^\top \| = \sqrt{\rho((ZZ^\top)^\top ZZ)} = \rho(ZZ^\top) < 1,
\]

and similarly

\[
\| Z^\top Z \| = \rho(Z^\top Z) < 1.
\]

Therefore, the following series converge absolutely:

\[
(I_p - Z^\top Z)^{\frac{1}{2}} = 1 + \frac{1}{2} Z^\top Z - \frac{1}{8}(Z^\top Z)^2 + \cdots + \frac{1}{2} \left( \frac{1}{2} - 1 \right) \cdots \left( \frac{1}{2} - k + 1 \right) \frac{1}{k!} (Z^\top Z)^k + \cdots
\]

and

\[
(I_q - ZZ^\top)^{\frac{1}{2}} = 1 + \frac{1}{2} ZZ^\top - \frac{1}{8}(ZZ^\top)^2 + \cdots + \frac{1}{2} \left( \frac{1}{2} - 1 \right) \cdots \left( \frac{1}{2} - k + 1 \right) \frac{1}{k!} (ZZ^\top)^k + \cdots.
\]

We get

\[
Z(I_p - Z^\top Z)^{\frac{1}{2}} = Z + \frac{1}{2} ZZ^\top Z - \frac{1}{8} Z(Z^\top Z)^2 + \cdots + \frac{1}{2} \left( \frac{1}{2} - 1 \right) \cdots \left( \frac{1}{2} - k + 1 \right) \frac{1}{k!} Z(Z^\top Z)^k + \cdots
\]

and

\[
(I_q - ZZ^\top)^{\frac{1}{2}} Z = Z + \frac{1}{2} ZZ^\top Z - \frac{1}{8} (ZZ^\top)^2 Z + \cdots + \frac{1}{2} \left( \frac{1}{2} - 1 \right) \cdots \left( \frac{1}{2} - k + 1 \right) \frac{1}{k!} (ZZ^\top)^k Z + \cdots.
\]

However

\[
Z(Z^\top Z)^k = \underbrace{Z Z^\top Z \cdots Z^\top Z}_{k} = \underbrace{Z Z^\top \cdots Z Z^\top}_{k} Z = (ZZ^\top)^k Z,
\]

which proves that \( (I_q - ZZ^\top)^{\frac{1}{2}} Z = Z(I_p - Z^\top Z)^{\frac{1}{2}} \), as required. \( \square \)
Another proof of Proposition 6.21 can be given using the SVD of \( Z \). Indeed, we can write

\[
Z = PDQ^T
\]

where \( P \) is a \( q \times q \) orthogonal matrix, \( Q \) is a \( p \times p \) orthogonal matrix, and \( D \) is a \( q \times p \) matrix whose diagonal entries are (strictly) positive and all other entries zero. Then,

\[
I_p - Z^\top Z = I_p - QD^\top P^\top PDQ^\top = Q(I_p - D^\top D)Q^\top,
\]

a symmetric positive definite matrix by assumption. We also have

\[
I_q - ZZ^\top = I_q - PDQ^\top QD^\top P^\top = P(I_q - DD^\top)P^\top,
\]

another symmetric positive definite matrix by assumption. Then,

\[
Z(I_p - Z^\top Z)^{-\frac{1}{2}} = PDQ^\top Q(I_p - D^\top D)^{-\frac{1}{2}}Q^\top = PD(I_p - D^\top D)^{-\frac{1}{2}}Q^\top
\]

and

\[
(I_q - ZZ^\top)^{-\frac{1}{2}} = P(I_q - DD^\top)^{-\frac{1}{2}}P^\top PDQ^\top = P(I_q - DD^\top)^{-\frac{1}{2}}DQ^\top,
\]

so it suffices to prove that

\[
D(I_p - D^\top D)^{-\frac{1}{2}} = (I_q - DD^\top)^{-\frac{1}{2}}D.
\]

However, \( D \) is essentially a diagonal matrix and the above is easily verified, as the reader should check.

**Remark:** The polar form of matrices in \( O(p, q) \) can be obtained via the exponential map and the Lie algebra, \( \mathfrak{o}(p, q) \), of \( O(p, q) \), see Section 6.5. Indeed, every matrix \( X \in O(p, q) \) has a polar form of the form

\[
X = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix},
\]

with \( P \in O(p) \), \( Q \in O(q) \), and with \( \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix} \) symmetric positive definite. This implies that

\[
x^\top S_1 x = (x^\top 0) \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} > 0
\]

for all \( x \in \mathbb{R}^p \), \( x \neq 0 \), and that

\[
y^\top S_3 y = (0 y^\top) \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix} \begin{pmatrix} 0 \\ y \end{pmatrix} > 0
\]

for all \( y \in \mathbb{R}^q \), \( y \neq 0 \). Therefore, \( S_1 \) and \( S_3 \) are symmetric positive definite. But then, if we write

\[
X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},
\]
from
\[
\begin{pmatrix}
A & B \\
C & D
\end{pmatrix}
= \begin{pmatrix}
P & 0 \\
0 & Q
\end{pmatrix}
\begin{pmatrix}
S_1 & S_2 \\
S_2^\top & S_3
\end{pmatrix},
\]
we get \( A = PS_1 \) and \( D = QS_3 \), which are polar decompositions of \( A \) and \( D \) respectively. On the other hand, our factorization
\[
\begin{pmatrix}
A & B \\
C & D
\end{pmatrix}
= \begin{pmatrix}
U & 0 \\
0 & V
\end{pmatrix}
\begin{pmatrix}
\alpha_1 & \alpha_2 Z^\top \\
\delta_1 Z & \delta_2
\end{pmatrix}
\]
yields \( A = U \alpha_1 \) and \( D = V \delta_2 \), with \( U \in O(p) \), \( V \in O(q) \), and \( \alpha_1, \alpha_2, \delta_1, \delta_2 \) symmetric positive definite. By uniqueness of the polar form, \( P = U \), \( Q = V \), \( S_1 = \alpha_1 \) and \( S_3 = \delta_2 \), which shows that our factorization is the polar decomposition of \( X \) after all! This can also be proved more directly using the fact that \( I - Z^\top Z \) (and \( I - ZZ^\top \)) being positive definite implies that the spectral norms \( \|Z\| \) and \( \|Z^\top\| \) of \( Z \) and \( Z^\top \) are both strictly less than one.

We also have the following amusing property of the determinants of \( A \) and \( D \):

**Proposition 6.22.** For any matrix \( X \in O(p,q) \), if we write
\[
X = \begin{pmatrix}
A & B \\
C & D
\end{pmatrix},
\]
then
\[
det(X) = det(A) det(D)^{-1} \quad \text{and} \quad \|det(A)| = |det(D)| \geq 1.
\]

**Proof.** Using the identities \( A^\top B = C^\top D \) and \( D^\top D = I_q + B^\top B \) proven in Proposition 6.20, observe that
\[
\begin{pmatrix}
A^\top & 0 \\
B^\top & -D^\top
\end{pmatrix}
\begin{pmatrix}
A & B \\
C & D
\end{pmatrix}
= \begin{pmatrix}
A^\top A & A^\top B \\
B^\top A - D^\top C & B^\top B - D^\top D
\end{pmatrix}
= \begin{pmatrix}
A^\top A & A^\top B \\
0 & -I_q
\end{pmatrix}.
\]
If we compute determinants, we get
\[
det(A)(-1)^q det(D) det(X) = det(A)^2(-1)^q.
\]
It follows that
\[
det(X) = det(A) det(D)^{-1}.
\]
From \( A^\top A = I_p + C^\top C \) and \( D^\top D = I_q + B^\top B \), we conclude that \( |det(A)| \geq 1 \) and \( |det(D)| \geq 1 \). Since \( |det(X)| = 1 \), we have \( |det(A)| = |det(D)| \geq 1 \). \( \Box \)

**Remark:** It is easy to see that the equations relating \( A, B, C, D \) established in the proof of Proposition 6.20 imply that
\[
det(A) = \pm 1 \quad \text{iff} \quad C = 0 \quad \text{iff} \quad B = 0 \quad \text{iff} \quad det(D) = \pm 1.
\]
We end this section by exhibiting a bijection between \( \text{O}(p, q) \) and \( \text{O}(p) \times \text{O}(q) \times \mathbb{R}^{pq} \), and in essence justifying the statement that \( \text{SO}_0(p, q) \) is homeomorphic to \( \text{SO}(p) \times \text{SO}(q) \times \mathbb{R}^{pq} \).

The construction of the bijection begins with the following claim: For every \( q \times p \) matrix \( Y \), there is a unique \( q \times p \) matrix \( Z \) such that \( I_q - ZZ^\top \) is positive definite symmetric matrix and

\[
(I_q - ZZ^\top)^{-\frac{1}{2}} Z = Y,
\]

given by

\[
Z = (I_q + YY^\top)^{-\frac{1}{2}} Y.
\]

To verify the claim, we start with a given \( Y \) and define \( Z = (I_q + YY^\top)^{-\frac{1}{2}} Y \), and show that \( Z \) satisfies \((*)\). Indeed, \( I_q + YY^\top \) is symmetric positive definite, and we have

\[
ZZ^\top = (I_q + YY^\top)^{-\frac{1}{2}} YY^\top (I_q + YY^\top)^{-\frac{1}{2}}
= (I_q + YY^\top)^{-\frac{1}{2}} (I_q + YY^\top) (I_q + YY^\top)^{-\frac{1}{2}}
= I_q - (I + YY^\top)^{-1},
\]

so

\[
I_q - ZZ^\top = (I_q + YY^\top)^{-1},
\]

from which we deduce that \( I_q - ZZ^\top \) is positive definite (since it is the inverse of a positive definite matrix, and hence must have positive eigenvalues). Note that \( I_q - ZZ^\top \) is also symmetric since it is the inverse of a symmetric matrix. It follows that

\[
(I_q - ZZ^\top)^{-\frac{1}{2}} Z = (I_q + YY^\top)^{\frac{1}{2}} (I_q + YY^\top)^{-\frac{1}{2}} Y = Y,
\]

which shows that \( Z = (I_q + YY^\top)^{-\frac{1}{2}} Y \) is a solution of \((*)\).

We now verify the uniqueness of the solution. Assume that \( Z \) is a solution of \((*)\). Then, we have

\[
YY^\top = (I_q - ZZ^\top)^{-\frac{1}{2}} ZZ^\top (I_q - ZZ^\top)^{-\frac{1}{2}}
= (I_q - ZZ^\top)^{-\frac{1}{2}} (I_q - Z(I_q - ZZ^\top)) (I_q - ZZ^\top)^{-\frac{1}{2}}
= (I_q - ZZ^\top)^{-1} - I_q,
\]

so \( (I_q - ZZ^\top)^{-1} = I_q + YY^\top \), which implies that

\[
Z = (I_q - ZZ^\top)^{\frac{1}{2}} Y = (I_q + YY^\top)^{-\frac{1}{2}} Y.
\]

Therefore, the map \( Y \mapsto (I_q + YY^\top)^{-\frac{1}{2}} Y \) is a bijection between \( \mathbb{R}^{pq} \) and the set of \( q \times p \) matrices \( Z \) such that \( I_q - ZZ^\top \) is symmetric positive definite, whose inverse is the map

\[
Z \mapsto (I_q - ZZ^\top)^{-\frac{1}{2}} Z = \delta^\frac{1}{2} Z.
\]

As a corollary, there is a bijection between \( \text{O}(p, q) \) and \( \text{O}(p) \times \text{O}(q) \times \mathbb{R}^{pq} \).
6.4 Pseudo-Algebraic Groups

The topological structure of certain linear Lie groups determined by equations among the real and the imaginary parts of their entries can be determined by refining the polar form of matrices. Such groups are called pseudo-algebraic groups. For example, the groups $\text{SO}(p,q)$ and $\text{SU}(p,q)$ are pseudo-algebraic.

Consider the group $\text{GL}(n,\mathbb{C})$ of invertible $n \times n$ matrices with complex coefficients. If $A = (a_{kl})$ is such a matrix, denote by $x_{kl}$ the real part (resp. $y_{kl}$, the imaginary part) of $a_{kl}$ (so, $a_{kl} = x_{kl} + iy_{kl}$).

**Definition 6.2.** A subgroup $G$ of $\text{GL}(n,\mathbb{C})$ is pseudo-algebraic iff there is a finite set of polynomials in $2n^2$ variables with real coefficients $\{P_j(X_1,\ldots,X_{n^2},Y_1,\ldots,Y_{n^2})\}_{j=1}^t$, so that

$$A = (x_{kl} + iy_{kl}) \in G \iff P_j(x_{11},\ldots,x_{nn},y_{11},\ldots,y_{nn}) = 0, \quad \text{for } j = 1,\ldots,t.$$  

Since a pseudo-algebraic subgroup is the zero locus of a set of polynomials, it is a closed subgroup, and thus a Lie group.

Recall that if $A$ is a complex $n \times n$-matrix, its adjoint $A^*$ is defined by $A^* = (\overline{A})^\top$. Also, $\text{U}(n)$ denotes the group of unitary matrices, i.e., those matrices $A \in \text{GL}(n,\mathbb{C})$ so that $AA^* = A^*A = I$, and $\text{H}(n)$ denotes the vector space of Hermitian matrices i.e., those matrices $A$ so that $A^* = A$.

The following proposition is needed.

**Proposition 6.23.** Let $P(x_1,\ldots,x_n)$ be a polynomial with real coefficients. For any $(a_1,\ldots,a_n) \in \mathbb{R}^n$, assume that $P(e^{ka_1},\ldots,e^{ka_n}) = 0$ for all $k \in \mathbb{N}$. Then,

$$P(e^{ta_1},\ldots,e^{ta_n}) = 0 \quad \text{for all } t \in \mathbb{R}.$$  

**Proof.** Any monomial $a_1^{i_1}\cdots a_n^{i_n}$ in $P$ when evaluated at $(e^{ta_1},\ldots,e^{ta_n})$ becomes $ae^t\sum a_{ij}^t$. Collecting terms with the same exponential part, we may assume that we have an expression of the form

$$\sum_{k=1}^N \alpha_k e^{tb_k}$$  

which vanishes for all $t \in \mathbb{N}$. We may also assume that $\alpha_k \neq 0$ for all $k$ and that the $b_k$ are sorted so that $b_1 < b_2 < \cdots < b_N$. Assume by contradiction that $N > 0$. If we multiply the above expression by $e^{-tb_N}$, by relabeling the coefficients $b_k$ in the exponentials, we may assume that $b_1 < b_2 < \cdots < b_{N-1} < 0 = b_N$. Now, if we let $t$ go to $+\infty$, the terms $\alpha_k e^{tb_k}$ go to 0 for $k = 1,\ldots,N-1$, and we get $a_N = 0$, a contradiction. \[\square\]

Then, we have the following theorem which is essentially a refined version of the polar decomposition of matrices:
6.4. PSEUDO-ALGEBRAIC GROUPS

Theorem 6.24. Let $G$ be a pseudo-algebraic subgroup of $\text{GL}(n, \mathbb{C})$ stable under adjunction (i.e., we have $A^* \in G$ whenever $A \in G$). Then, there is some integer $d \in \mathbb{N}$ so that $G$ is homeomorphic to $(G \cap \text{U}(n)) \times \mathbb{R}^d$. Moreover, if $\mathfrak{g}$ is the Lie algebra of $G$, the map

$$(\text{U}(n) \cap G) \times (\text{H}(n) \cap \mathfrak{g}) \longrightarrow G \quad \text{given by} \quad (U, H) \mapsto Ue^H,$$ 

is a homeomorphism onto $G$.

Proof. We follow the proof in Mneimné and Testard [130] (Chapter 3); a similar proof is given in Knapp [106] (Chapter 1). First, we observe that for every invertible matrix $P$, the group $G$ is pseudo-algebraic iff $PGP^{-1}$ is pseudo-algebraic, since the map $X \mapsto PXP^{-1}$ is linear.

By the polar decomposition, every matrix $A \in G$ can be written uniquely as $A = US$, where $U \in \text{U}(n)$ and $S \in \text{HPD}(n)$. Furthermore, by Theorem 1.10, the matrix $S$ can be written (uniquely) as $S = e^H$, for some unique Hermitian matrix $H \in \text{H}(n)$, so we have $A = Ue^H$. We need to prove that $H \in \mathfrak{g}$ and that $U \in G$. Since $G$ is closed under adjunction, $A^* \in G$, that is $e^H U^* \in G$, so $e^H U^* U e^H = e^{2H} \in G$. If we can prove that $e^{tH} \in G$ for all $t \in \mathbb{R}$, then $H \in \mathfrak{g}$ and $e^H \in G$, so $U \in e^{-H} A \in G$.

Since $2H$ is Hermitian, it has real eigenvalues $\lambda_1, \ldots, \lambda_n$ and it can be diagonalized as $2H = VAV^{-1}$, where $V$ is unitary and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. By a previous observation, the group $VGV^{-1}$ is also pseudo-algebraic, so we may assume that $2H$ is a diagonal matrix with real entries, and to say that $e^{2H} \in G$ means that $e^{\lambda_1}, \ldots, e^{\lambda_n}$ satisfy a set of algebraic equations. Since $G$ is a group, for every $k \in \mathbb{Z}$, we have $e^{k2H} \in G$, so $e^{k\lambda_1}, \ldots, e^{k\lambda_n}$ satisfy the same set of algebraic equations. By Proposition 6.23, $e^{t\lambda_1}, \ldots, e^{t\lambda_n}$ satisfy the same set of algebraic equations for all $t \in \mathbb{R}$, which means that $e^{tH} \in G$ for all $t \in \mathbb{R}$. It follows that $H \in \mathfrak{g}$, $e^H \in G$, and thus $U \in e^{-H} A \in G$.

For invertible matrices, the polar decomposition is unique, so we found a unique $U \in \text{U}(n) \cap G$ and a unique matrix $H \in \text{H}(n) \cap \mathfrak{g}$ so that

$$A = Ue^H.$$ 

The fact that the map $(U, H) \mapsto Ue^H$ is a homeomorphism takes a little bit of work. This follows from the fact that polar decomposition and the bijection between $\text{H}(n)$ and $\text{HPD}(n)$ are homeomorphisms (see Section 1.5); these facts are proved in Mneimné and Testard [130]; see Theorem 1.6.3 for the first homeomorphism and Theorem 3.3.4 for the second homeomorphism. Since $\text{H}(n) \cap \mathfrak{g}$ is a real vector space, it is isomorphic to $\mathbb{R}^d$ for some $d \in \mathbb{N}$, and so $G$ is homeomorphic to $(G \cap \text{U}(n)) \times \mathbb{R}^d$. 

Observe that if $G$ is also compact then $d = 0$, and $G \subseteq \text{U}(n)$.

Remark: A subgroup $G$ of $\text{GL}(n, \mathbb{R})$ is called algebraic if there is a finite set of polynomials in $n^2$ variables with real coefficients $\{P_j(x_1, \ldots, x_{n^2})\}_{j=1}^t$, so that

$$A = (x_{kl}) \in G \quad \text{iff} \quad P_j(x_{11}, \ldots, x_{nm}) = 0, \quad \text{for} \quad j = 1, \ldots, t.$$
Then, it can be shown that every compact subgroup of $\text{GL}(n, \mathbb{R})$ is algebraic. The proof is quite involved and uses the existence of the Haar measure on a compact Lie group; see Mneimné and Testard [130] (Theorem 3.7).

6.5 More on the Topology of $\text{O}(p, q)$ and $\text{SO}(p, q)$

It turns out that the topology of the group $\text{O}(p, q)$ is completely determined by the topology of $\text{O}(p)$ and $\text{O}(q)$. This result can be obtained as a simple consequence of some standard Lie group theory. The key notion is that of a pseudo-algebraic group defined in Section 6.4.

We can apply Theorem 6.24 to determine the structure of the space $\text{O}(p, q)$. We know that $\text{O}(p, q)$ consists of the matrices $A$ in $\text{GL}(p+q, \mathbb{R})$ such that

$$A^\top I_{p,q} A = I_{p,q},$$

and so $\text{O}(p, q)$ is clearly pseudo-algebraic. Using the above equation, and the curve technique demonstrated at the beginning of Section 6.2, it is easy to determine the Lie algebra $\mathfrak{o}(p, q)$ of $\text{O}(p, q)$. We find that $\mathfrak{o}(p, q)$ is given by

$$\mathfrak{o}(p, q) = \left\{ \begin{pmatrix} X_1 & X_2 \\
X_2^\top & X_3 \end{pmatrix} \left| X_1^\top = -X_1, \ X_3^\top = -X_3, \ X_2 \text{ arbitrary} \right. \right\},$$

where $X_1$ is a $p \times p$ matrix, $X_3$ is a $q \times q$ matrix, and $X_2$ is a $p \times q$ matrix. Consequently, it immediately follows that

$$\mathfrak{o}(p, q) \cap H(p+q) = \left\{ \begin{pmatrix} 0 & X_2 \\
X_2^\top & 0 \end{pmatrix} \left| X_2 \text{ arbitrary} \right. \right\},$$

a vector space of dimension $pq$.

Some simple calculations also show that

$$\text{O}(p, q) \cap U(p+q) = \left\{ \begin{pmatrix} X_1 & 0 \\
0 & X_2 \end{pmatrix} \left| X_1 \in \text{O}(p), \ X_2 \in \text{O}(q) \right. \right\} \cong \text{O}(p) \times \text{O}(q).$$

Therefore, we obtain the structure of $\text{O}(p, q)$:

**Proposition 6.25.** The topological space $\text{O}(p, q)$ is homeomorphic to $\text{O}(p) \times \text{O}(q) \times \mathbb{R}^{pq}$.

Since $\text{O}(p)$ has two connected components when $p \geq 1$, we see that $\text{O}(p, q)$ has four connected components when $p, q \geq 1$. It is also obvious that

$$\text{SO}(p, q) \cap U(p+q) = \left\{ \begin{pmatrix} X_1 & 0 \\
0 & X_2 \end{pmatrix} \left| X_1 \in \text{O}(p), \ X_2 \in \text{O}(q), \ \det(X_1) \det(X_2) = 1 \right. \right\}.$$
6.5. MORE ON THE TOPOLOGY OF $O(P,Q)$ AND $SO(P,Q)$

Proposition 6.26. The topological space $SO(p,q)$ is homeomorphic to $S(O(p) \times O(q)) \times \mathbb{R}^{pq}$.

Observe that the dimension of all these spaces depends only on $p + q$: It is $(p + q)(p + q - 1)/2$. Also, $SO(p,q)$ has two connected components when $p, q \geq 1$. The connected component of $I_{p+q}$ is the group $SO_0(p,q)$. This latter space is homeomorphic to $SO(p) \times SO(q) \times \mathbb{R}^{pq}$. If we write

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix},$$

then it is shown in O’Neill [138] (Chapter 9, Lemma 6) that the connected component $SO_0(p,q)$ of $SO(p,q)$ containing $I$ is given by

$$SO_0(p,q) = \{A \in GL(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}, \det(P) > 0, \det(S) > 0\}.$$  

For both $SO(p,q)$ and $SO_0(p,q)$, the inverse is given by

$$A^{-1} = I_{p,q} A^\top I_{p,q}.$$  

We can show that $SO(p,q)$ and $SO(q,p)$ are isomorphic (similarly $SO_0(p,q)$ and $SO_0(q,p)$ are isomorphic) as follows. Let $J_{p,q}$ be the permutation matrix

$$J_{p,q} = \begin{pmatrix} 0 & I_q \\ I_p & 0 \end{pmatrix}.$$  

Observe that $J_{p,q} J_{q,p} = I_{p+q}$. Then, it is easy to check that the map $\psi$ given by

$$\psi(A) = J_{p,q} A J_{q,p}$$

is an isomorphism between $SO(p,q)$ and $SO(q,p)$, and an isomorphism between $SO_0(p,q)$ and $SO_0(q,p)$.

Theorem 6.24 gives the polar form of a matrix $A \in O(p,q)$: We have

$$A = U e^S,$$

with $U \in O(p) \times O(q)$ and $S \in so(p,q) \cap S(p+q)$,

where $U$ is of the form

$$U = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix},$$

with $P \in O(p)$ and $Q \in O(q)$,

and $so(p,q) \cap S(p+q)$ consists of all $(p+q) \times (p+q)$ symmetric matrices of the form

$$S = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix},$$

with $X$ an arbitrary $p \times q$ matrix.
It turns out that it is not very hard to compute explicitly the exponential $e^S$ of such matrices (see Mneimné and Testard [130]). Recall that the functions $\cosh$ and $\sinh$ also make sense for matrices (since the exponential makes sense) and are given by

$$
\cosh(A) = \frac{e^A + e^{-A}}{2} = I + \frac{A^2}{2!} + \cdots + \frac{A^{2k}}{(2k)!} + \cdots
$$

and

$$
\sinh(A) = \frac{e^A - e^{-A}}{2} = A + \frac{A^3}{3!} + \cdots + \frac{A^{2k+1}}{(2k+1)!} + \cdots.
$$

We also set

$$
\frac{\sinh(A)}{A} = I + \frac{A^2}{3!} + \cdots + \frac{A^{2k}}{(2k+1)!} + \cdots,
$$

which is defined for all matrices $A$ (even when $A$ is singular). Then, we have

**Proposition 6.27.** For any matrix $S$ of the form

$$
S = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix},
$$

we have

$$
e^S = \begin{pmatrix} \cosh((XX^\top)^{\frac{1}{2}}) & \frac{\sinh((XX^\top)^{\frac{1}{2}})X}{(XX^\top)^{\frac{1}{2}}} \\ \frac{\sinh(X^\top X)^{\frac{1}{2}}X^\top}{(X^\top X)^{\frac{1}{2}}} & \cosh((X^\top X)^{\frac{1}{2}}) \end{pmatrix}.
$$

**Proof.** By induction, it is easy to see that

$$
S^{2k} = \begin{pmatrix} (XX^\top)^k & 0 \\ 0 & (X^\top X)^k \end{pmatrix}
$$

and

$$
S^{2k+1} = \begin{pmatrix} 0 & (XX^\top)^kX \\ (X^\top X)^kX^\top & 0 \end{pmatrix}
$$

The rest is left as an exercise. \(\square\)

**Remark:** Although at first glance, $e^S$ does not look symmetric, but it is!

As a consequence of Proposition 6.27, every matrix $A \in O(p, q)$ has the polar form

$$
A = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} \cosh((XX^\top)^{\frac{1}{2}}) & \frac{\sinh((XX^\top)^{\frac{1}{2}})X}{(XX^\top)^{\frac{1}{2}}} \\ \frac{\sinh(X^\top X)^{\frac{1}{2}}X^\top}{(X^\top X)^{\frac{1}{2}}} & \cosh((X^\top X)^{\frac{1}{2}}) \end{pmatrix},
$$

with $P \in O(p)$, $Q \in O(q)$, and $X$ an arbitrary $p \times q$ matrix.
Chapter 7

Manifolds, Tangent Spaces, Cotangent Spaces, Submanifolds, Manifolds With Boundary

7.1 Charts and Manifolds

In Chapter 4 we defined the notion of a manifold embedded in some ambient space $\mathbb{R}^N$. In order to maximize the range of applications of the theory of manifolds, it is necessary to generalize the concept of a manifold to spaces that are not a priori embedded in some $\mathbb{R}^N$. The basic idea is still that, whatever a manifold is, it is a topological space that can be covered by a collection of open subsets $U_\alpha$, where each $U_\alpha$ is isomorphic to some "standard model," e.g., some open subset of Euclidean space $\mathbb{R}^n$. Of course, manifolds would be very dull without functions defined on them and between them. This is a general fact learned from experience: Geometry arises not just from spaces but from spaces and interesting classes of functions between them. In particular, we still would like to "do calculus" on our manifold and have good notions of curves, tangent vectors, differential forms, etc.

The small drawback with the more general approach is that the definition of a tangent vector is more abstract. We can still define the notion of a curve on a manifold, but such a curve does not live in any given $\mathbb{R}^n$, so it it not possible to define tangent vectors in a simple-minded way using derivatives. Instead, we have to resort to the notion of chart. This is not such a strange idea. For example, a geography atlas gives a set of maps of various portions of the earth and this provides a very good description of what the earth is, without actually imagining the earth embedded in 3-space.

The material of this chapter borrows from many sources, including Warner [175], Berger and Gostiaux [20], O'Neill [138], Do Carmo [60, 59], Gallot, Hulin and Lafontaine [73], Lang [114], Schwartz [156], Hirsch [91], Sharpe [162], Guillemin and Pollack [83], Lafontaine [110], Dubrovin, Fomenko and Novikov [63] and Boothby [22]. A nice (not very technical) exposition is given in Morita [133] (Chapter 1). The recent book by Tu [170] is also highly
recommended for its clarity. Among the many texts on manifolds and differential geometry, the book by Choquet-Bruhat, DeWitt-Morette and Dillard-Bleick [44] stands apart because it is one of the clearest and most comprehensive. (Many proofs are omitted, but this can be an advantage!) Being written for (theoretical) physicists, it contains more examples and applications than most other sources.

Given $\mathbb{R}^n$, recall that the projection functions $pr_i: \mathbb{R}^n \to \mathbb{R}$ are defined by

$$pr_i(x_1, \ldots, x_n) = x_i, \quad 1 \leq i \leq n.$$  

For technical reasons (in particular, to ensure the existence of partitions of unity, a crucial tool in manifold theory; see Sections 9.1 and 24.1) and to avoid “esoteric” manifolds that do not arise in practice, from now on, all topological spaces under consideration will be assumed to be Hausdorff and second-countable (which means that the topology has a countable basis).

The first step in generalizing the notion of a manifold is to define charts, a way to say that locally a manifold “looks like” an open subset of $\mathbb{R}^n$.

**Definition 7.1.** Given a topological space $M$, a chart (or local coordinate map) is a pair $(U, \varphi)$, where $U$ is an open subset of $M$ and $\varphi: U \to \Omega$ is a homeomorphism onto an open subset $\Omega = \varphi(U)$ of $\mathbb{R}^{n_\varphi}$ (for some $n_\varphi \geq 1$). For any $p \in M$, a chart $(U, \varphi)$ is a chart at $p$ iff $p \in U$. If $(U, \varphi)$ is a chart, then the functions $x_i = pr_i \circ \varphi$ are called local coordinates and for every $p \in U$, the tuple $(x_1(p), \ldots, x_n(p))$ is the set of coordinates of $p$ w.r.t. the chart. The inverse $(\Omega, \varphi^{-1})$ of a chart is called a local parametrization.

![Figure 7.1: A chart $(U, \varphi)$ on $M$](image)
7.1. CHARTS AND MANIFOLDS

Given any two charts \((U_i, \varphi_i)\) and \((U_j, \varphi_j)\), if \(U_i \cap U_j \neq \emptyset\), we have the transition maps \(\varphi_i^j: \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)\) and \(\varphi_j^i: \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)\), defined by

\[
\varphi_i^j = \varphi_j \circ \varphi_i^{-1} \quad \text{and} \quad \varphi_j^i = \varphi_i \circ \varphi_j^{-1}.
\]

![Diagram of transition maps](image)

**Figure 7.2:** The transition maps \(\varphi_i^j\) and \(\varphi_j^i\)

Clearly, \(\varphi_j^i = (\varphi_j^i)^{-1}\). Observe that the transition maps \(\varphi_i^j\) (resp. \(\varphi_j^i\)) are maps between open subsets of \(\mathbb{R}^n\). This is good news! Indeed, the whole arsenal of calculus is available for functions on \(\mathbb{R}^n\), and we will be able to promote many of these results to manifolds by imposing suitable conditions on transition functions.

As in Section 4.1, whatever our generalized notion of a manifold is, we would like to define the notion of tangent space at a point of manifold, the notion of smooth function between manifolds, and the notion of derivative of a function (at a point) between manifolds. Unfortunately, even though our parametrizations \(\varphi^{-1}: \Omega \to U\) are homeomorphisms, since \(U\) is a subset of a space \(M\) which is not assumed to be contained in \(\mathbb{R}^N\) (for any \(N\)), the derivative \(d\varphi_{\varphi^{-1}}\) does not make sense, unlike in the situation of Definition 4.1. Therefore, some extra conditions on the charts must be imposed in order to recapture the fact that for manifolds embedded in \(\mathbb{R}^N\), the parametrizations are immersions. An invaluable hint is provided by Lemma 4.2: we require the transition maps \(\varphi_i^j: \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)\) to be sufficiently differentiable. This makes perfect sense since the \(\varphi_i^j\) are functions between open
subsets of \( \mathbb{R}^n \). It also turns out that these conditions on transition maps guarantee that notions, such as tangent vectors, whose definition seems to depend on the choice of a chart, are in fact independent of the choice of charts. The above motivations suggest the following requirements on charts.

**Definition 7.2.** Given a topological space \( M \), given some integer \( n \geq 1 \) and given some \( k \) such that \( k \) is either an integer \( k \geq 1 \) or \( k = \infty \), a \( C^k \) \( n \)-atlas (or \( n \)-atlas of class \( C^k \)) \( \mathcal{A} \) is a family of charts \( \{(U_i, \varphi_i)\} \), such that

1. \( \varphi_i(U_i) \subseteq \mathbb{R}^n \) for all \( i \);
2. The \( U_i \) cover \( M \), i.e.,
   \[ M = \bigcup_i U_i; \]
3. Whenever \( U_i \cap U_j \neq \emptyset \), the transition map \( \varphi_j^i \) (and \( \varphi_i^j \)) is a \( C^k \)-diffeomorphism. When \( k = \infty \), the \( \varphi_i^j \) are smooth diffeomorphisms.

We must ensure that we have enough charts in order to carry out our program of generalizing calculus on \( \mathbb{R}^n \) to manifolds. For this, we must be able to add new charts whenever necessary, provided that they are consistent with the previous charts in an existing atlas. Technically, given a \( C^k \) \( n \)-atlas \( \mathcal{A} \) on \( M \), for any other chart \( (U, \varphi) \), we say that \( (U, \varphi) \) is compatible with the atlas \( \mathcal{A} \) iff every map \( \varphi \circ \varphi^{-1} \) and \( \varphi \circ \varphi_i^{-1} \) is \( C^k \) (whenever \( U \cap U_i \neq \emptyset \)).

Two atlases \( \mathcal{A} \) and \( \mathcal{A}' \) on \( M \) are compatible iff every chart of one is compatible with the other atlas. This is equivalent to saying that the union of the two atlases is still an atlas.

It is immediately verified that compatibility induces an equivalence relation on \( C^k \) \( n \)-atlases on \( M \). In fact, given an atlas \( \mathcal{A} \) for \( M \), the collection \( \tilde{\mathcal{A}} \) of all charts compatible with \( \mathcal{A} \) is a maximal atlas in the equivalence class of atlases compatible with \( \mathcal{A} \). Finally, we have our generalized notion of a manifold.

**Definition 7.3.** Given some integer \( n \geq 1 \) and given some \( k \) such that \( k \) is either an integer \( k \geq 1 \) or \( k = \infty \), a \( C^k \)-manifold of dimension \( n \) consists of a topological space \( M \) together with an equivalence class \( \tilde{\mathcal{A}} \) of \( C^k \) \( n \)-atlases on \( M \). Any atlas \( \mathcal{A} \) in the equivalence class \( \tilde{\mathcal{A}} \) is called a differentiable structure of class \( C^k \) (and dimension \( n \)) on \( M \). We say that \( M \) is modeled on \( \mathbb{R}^n \). When \( k = \infty \), we say that \( M \) is a smooth manifold.

**Remark:** It might have been better to use the terminology abstract manifold rather than manifold, to emphasize the fact that the space \( M \) is not a priori a subspace of \( \mathbb{R}^N \), for some suitable \( N \).

We can allow \( k = 0 \) in the above definitions. In this case, condition (3) in Definition 7.2 is void, since a \( C^0 \)-diffeomorphism is just a homeomorphism, but \( \varphi_i^j \) is always a homeomorphism. In this case, \( M \) is called a topological manifold of dimension \( n \). We do not require a manifold to be connected but we require all the components to have the same dimension \( n \).
Actually, on every connected component of $M$, it can be shown that the dimension $n_\varphi$ of the range of every chart is the same. This is quite easy to show if $k \geq 1$ but for $k = 0$, this requires a deep theorem of Brouwer. (Brouwer's Invariance of Domain Theorem states that if $U \subseteq \mathbb{R}^n$ is an open set and if $f: U \to \mathbb{R}^n$ is a continuous and injective map, then $f(U)$ is open in $\mathbb{R}^n$. Using Brouwer's Theorem, we can show the following fact: If $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ are two open subsets and if $f: U \to V$ is a homeomorphism between $U$ and $V$, then $m = n$. If $m > n$, then consider the injection, $i: \mathbb{R}^n \to \mathbb{R}^m$, where $i(x) = (x, 0_{m-n})$. Clearly, $i$ is injective and continuous, so $i \circ f: U \to i(V)$ is injective and continuous and Brouwer's Theorem implies that $i(V)$ is open in $\mathbb{R}^m$, which is a contradiction, as $i(V) = V \times \{0_{m-n}\}$ is not open in $\mathbb{R}^m$. If $m < n$, consider the homeomorphism $f^{-1}: V \to U$.)

What happens if $n = 0$? In this case, every one-point subset of $M$ is open, so every subset of $M$ is open; that is, $M$ is any (countable if we assume $M$ to be second-countable) set with the discrete topology!

Observe that since $\mathbb{R}^n$ is locally compact and locally connected, so is every manifold (check this!).

In order to get a better grasp of the notion of manifold it is useful to consider examples of non-manifolds. First, consider the curve in $\mathbb{R}^2$ given by the zero locus of the equation

$$y^2 = x^2 - x^3,$$

namely, the set of points

$$M_1 = \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^2 - x^3\}.$$

This curve showed in Figure 7.3 and called a nodal cubic is also defined as the parametric curve

$$x = 1 - t^2$$
$$y = t(1 - t^2).$$

We claim that $M_1$ is not even a topological manifold. The problem is that the nodal cubic has a self-intersection at the origin. If $M_1$ was a topological manifold, then there would be a
connected open subset $U \subseteq M_1$ containing the origin $O = (0, 0)$, namely the intersection of a small enough open disc centered at $O$ with $M_1$, and a local chart $\varphi: U \to \Omega$, where $\Omega$ is some connected open subset of $\mathbb{R}$ (that is, an open interval), since $\varphi$ is a homeomorphism. However, $U - \{O\}$ consists of four disconnected components, and $\Omega - \varphi(O)$ of two disconnected components, contradicting the fact that $\varphi$ is a homeomorphism.

Let us now consider the curve in $\mathbb{R}^2$ given by the zero locus of the equation

$$y^2 = x^3,$$

namely, the set of points

$$M_2 = \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^3\}.$$

This curve showed in Figure 7.4 and called a cuspidal cubic is also defined as the parametric curve

$$x = t^2, \quad y = t^3.$$

Consider the map, $\varphi: M_2 \to \mathbb{R}$, given by

$$\varphi(x, y) = y^{1/3}.$$ 

Since $x = y^{2/3}$ on $M_2$, we see that $\varphi^{-1}$ is given by

$$\varphi^{-1}(t) = (t^2, t^3)$$

and clearly $\varphi$ is a homeomorphism, so $M_2$ is a topological manifold. However, with the atlas consisting of the single chart $\{\varphi: M_2 \to \mathbb{R}\}$, the space $M_2$ is also a smooth manifold! Indeed, as there is a single chart, condition (3) of Definition 7.2 holds vacuously.

This fact is somewhat unexpected because the cuspidal cubic is not smooth at the origin, since the tangent vector of the parametric curve $c: t \mapsto (t^2, t^3)$ at the origin is the zero vector (the velocity vector at $t$ is $c'(t) = (2t, 3t^2)$). However, this apparent paradox has
to do with the fact that, as a parametric curve, $M_2$ is not immersed in $\mathbb{R}^2$ since $c'$ is not injective (see Definition 7.17 (a)), whereas as an abstract manifold, with this single chart, $M_2$ is diffeomorphic to $\mathbb{R}$.

Now, we also have the chart $\psi: M_2 \to \mathbb{R}$, given by $\psi(x, y) = y,$ with $\psi^{-1}$ given by $\psi^{-1}(u) = (u^{2/3}, u).$

With the atlas consisting of the single chart $\{\psi: M_2 \to \mathbb{R}\}$, the space $M_2$ is also a smooth manifold. Observe that $\varphi \circ \psi^{-1}(u) = u^{1/3},$

a map that is not differentiable at $u = 0$. Therefore, the atlas $\{\varphi: M_2 \to \mathbb{R}, \psi: M_2 \to \mathbb{R}\}$ is not $C^1$, and thus with respect to that atlas, $M_2$ is not a $C^1$-manifold. This example also shows that the atlases $\{\varphi: M_2 \to \mathbb{R}\}$ and $\{\psi: M_2 \to \mathbb{R}\}$ are inequivalent.

The example of the cuspidal cubic reveals one of the subtleties of the definition of a $C^k$ (or $C^\infty$) manifold: whether a topological space is a $C^k$-manifold or a smooth manifold depends on the choice of atlas. As a consequence, if a space $M$ happens to be a topological manifold because it has an atlas consisting of a single chart, or more generally if it has an atlas whose transition functions “avoid” singularities, then it is automatically a smooth manifold. In particular, if $f: U \to \mathbb{R}^m$ is any continuous function from some open subset $U$ of $\mathbb{R}^n$ to $\mathbb{R}^m$, then the graph $\Gamma(f) \subseteq \mathbb{R}^{n+m}$ of $f$ given by $\Gamma(f) = \{(x, f(x)) \in \mathbb{R}^{n+m} \mid x \in U\}$

is a smooth manifold of dimension $n$ with respect to the atlas consisting of the single chart $\varphi: \Gamma(f) \to U$, given by $\varphi(x, f(x)) = x,$ with its inverse $\varphi^{-1}: U \to \Gamma(f)$ given by $\varphi^{-1}(x) = (x, f(x)).$

The notion of a submanifold using the concept of “adapted chart” (see Definition 7.16 in Section 7.6) gives a more satisfactory treatment of $C^k$ (or smooth) submanifolds of $\mathbb{R}^n$.

It should also be noted that determining the number of inequivalent differentiable structures on a topological space is a very difficult problem, even for $\mathbb{R}^n$. In the case of $\mathbb{R}^n$, it turns out that any two smooth differentiable structures are diffeomorphic, except for $n = 4$. For $n = 4$, it took some very hard and deep work to show that there are uncountably many distinct diffeomorphism classes of differentiable structures. The case of the spheres $S^n$ is even more mysterious. It is known that there is a single diffeomorphism class for $n = 1, 2, 3$, but for $n = 4$ the answer is unknown! For $n = 15$, there are 16,256 distinct classes; for
more about these issues, see Conlon [45] (Chapter 3). It is also known that every topological manifold admits a smooth structure for \( n = 1, 2, 3 \). However, for \( n = 4 \), there exist nonsmoothable manifolds; see Conlon [45] (Chapter 3).

In some cases, \( M \) does not come with a topology in an obvious (or natural) way and a slight variation of Definition 7.2 is more convenient in such a situation:

**Definition 7.4.** Given a set \( M \), given some integer \( n \geq 1 \) and given some \( k \) such that \( k \) is either an integer \( k \geq 1 \) or \( k = \infty \), a \( C^k \) \( n \)-atlas (or \( n \)-atlas of class \( C^k \)) \( A \) is a family of charts \( \{(U_i, \varphi_i)\} \), such that

1. Each \( U_i \) is a subset of \( M \) and \( \varphi_i : U_i \to \varphi_i(U_i) \) is a bijection onto an open subset \( \varphi_i(U_i) \subseteq \mathbb{R}^n \), for all \( i \);
2. The \( U_i \) cover \( M \); that is,
   \[
   M = \bigcup_i U_i;
   \]
3. Whenever \( U_i \cap U_j \neq \emptyset \), the sets \( \varphi_i(U_i \cap U_j) \) and \( \varphi_j(U_i \cap U_j) \) are open in \( \mathbb{R}^n \) and the transition maps \( \varphi^i_j \) and \( \varphi^j_i \) are \( C^k \)-diffeomorphisms.

Then, the notion of a chart being compatible with an atlas and of two atlases being compatible is just as before, and we get a new definition of a manifold analogous to Definition 7.3. But, this time we give \( M \) the topology in which the open sets are arbitrary unions of domains of charts \( U_i \), more precisely, the \( U_i \)'s of the maximal atlas defining the differentiable structure on \( M \).

It is not difficult to verify that the axioms of a topology are verified, and \( M \) is indeed a topological space with this topology. It can also be shown that when \( M \) is equipped with the above topology, then the maps \( \varphi_i : U_i \to \varphi_i(U_i) \) are homeomorphisms, so \( M \) is a manifold according to Definition 7.3. We also require that under this topology, \( M \) is Hausdorff and second-countable. A sufficient condition (in fact, also necessary!) for being second-countable is that some atlas be countable. A sufficient condition of \( M \) to be Hausdorff is that for all \( p, q \in M \) with \( p \neq q \), either \( p, q \in U_i \) for some \( U_i \), or \( p \in U_i \) and \( q \in U_j \) for some disjoint \( U_i, U_j \). Thus, we are back to the original notion of a manifold where it is assumed that \( M \) is already a topological space.

One can also define the topology on \( M \) in terms of any of the atlases \( A \) defining \( M \) (not only the maximal one) by requiring \( U \subseteq M \) to be open iff \( \varphi_i(U \cap U_i) \) is open in \( \mathbb{R}^n \), for every chart \( (U_i, \varphi_i) \) in the atlas \( A \). Then, one can prove that we obtain the same topology as the topology induced by the maximal atlas. For details, see Berger and Gostiaux [20], Chapter 2.

If the underlying topological space of a manifold is compact, then \( M \) has some finite atlas. Also, if \( A \) is some atlas for \( M \) and \( (U, \varphi) \) is a chart in \( A \), for any (nonempty) open subset \( V \subseteq U \), we get a chart \( (V, \varphi \upharpoonright V) \), and it is obvious that this chart is compatible with
Thus, \((V, \varphi \upharpoonright V)\) is also a chart for \(M\). This observation shows that if \(U\) is any open subset of a \(C^k\)-manifold \(M\), then \(U\) is also a \(C^k\)-manifold whose charts are the restrictions of charts on \(M\) to \(U\).

We are now fully prepared to present a variety of examples.

**Example 7.1.** The sphere \(S^n\).

Using the stereographic projections (from the north pole and the south pole), we can define two charts on \(S^n\) and show that \(S^n\) is a smooth manifold. Let \(\sigma_N: S^n - \{N\} \to \mathbb{R}^n\) and \(\sigma_S: S^n - \{S\} \to \mathbb{R}^n\), where \(N = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1}\) (the north pole) and \(S = (0, \ldots, 0, -1) \in \mathbb{R}^{n+1}\) (the south pole) be the maps called respectively stereographic projection from the north pole and stereographic projection from the south pole, given by

\[
\sigma_N(x_1, \ldots, x_{n+1}) = \frac{1}{1 - x_{n+1}} (x_1, \ldots, x_n) \quad \text{and} \quad \sigma_S(x_1, \ldots, x_{n+1}) = \frac{1}{1 + x_{n+1}} (x_1, \ldots, x_n).
\]

The inverse stereographic projections are given by

\[
\sigma_N^{-1}(x_1, \ldots, x_n) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right) + 1} \left(2x_1, \ldots, 2x_n, \left(\sum_{i=1}^{n} x_i^2\right) - 1\right)
\]

and

\[
\sigma_S^{-1}(x_1, \ldots, x_n) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right) + 1} \left(2x_1, \ldots, 2x_n, -\left(\sum_{i=1}^{n} x_i^2\right) + 1\right).
\]

Thus, if we let \(U_N = S^n - \{N\}\) and \(U_S = S^n - \{S\}\), we see that \(U_N\) and \(U_S\) are two open subsets covering \(S^n\), both homeomorphic to \(\mathbb{R}^n\). Furthermore, it is easily checked that on the overlap \(U_N \cap U_S = S^n - \{N, S\}\), the transition maps

\[
\mathcal{I} = \sigma_S \circ \sigma_N^{-1} = \sigma_N \circ \sigma_S^{-1}
\]

defined on \(\varphi_N(U_N \cap U_S) = \varphi_S(U_N \cap U_S) = \mathbb{R}^n - \{0\}\), are given by

\[
(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^{n} x_i^2} (x_1, \ldots, x_n);
\]

that is, the inversion \(\mathcal{I}\) of center \(O = (0, \ldots, 0)\) and power 1. Clearly, this map is smooth on \(\mathbb{R}^n - \{O\}\), so we conclude that \((U_N, \sigma_N)\) and \((U_S, \sigma_S)\) form a smooth atlas for \(S^n\).

**Example 7.2.** Smooth manifolds in \(\mathbb{R}^N\).

Any \(m\)-dimensional manifold \(M\) in \(\mathbb{R}^N\) is a smooth manifold, because by Lemma 4.2, the inverse maps \(\varphi^{-1}: U \to \Omega\) of the parametrizations \(\varphi: \Omega \to U\) are charts that yield smooth transition functions. In particular, by Theorem 4.8, any linear Lie group is a smooth manifold.
Example 7.3. The projective space $\mathbb{RP}^n$.

To define an atlas on $\mathbb{RP}^n$, it is convenient to view $\mathbb{RP}^n$ as the set of equivalence classes of vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation

$$u \sim v \text{ iff } v = \lambda u, \text{ for some } \lambda \neq 0 \in \mathbb{R}.$$ 

Given any $p = [x_1, \ldots, x_{n+1}] \in \mathbb{RP}^n$, we call $(x_1, \ldots, x_{n+1})$ the homogeneous coordinates of $p$. It is customary to write $(x_1: \cdots: x_{n+1})$ instead of $[x_1, \ldots, x_{n+1}]$. (Actually, in most books, the indexing starts with 0, i.e., homogeneous coordinates for $\mathbb{RP}^n$ are written as $(x_0: \cdots: x_n)$.) Now, $\mathbb{RP}^n$ can also be viewed as the quotient of the sphere $S^n$ under the equivalence relation where any two antipodal points $x$ and $-x$ are identified. It is not hard to show that the projection $\pi: S^n \to \mathbb{RP}^n$ is both open and closed. Since $S^n$ is compact and second-countable, we can apply Propositions 3.31 and 3.33 to prove that under the quotient topology, $\mathbb{RP}^n$ is Hausdorff, second-countable, and compact.

We define charts in the following way. For any $i$, with $1 \leq i \leq n+1$, let

$$U_i = \{(x_1: \cdots: x_{n+1}) \in \mathbb{RP}^n \mid x_i \neq 0\}.$$ 

Observe that $U_i$ is well defined, because if $(y_1: \cdots: y_{n+1}) = (x_1: \cdots: x_{n+1})$, then there is some $\lambda \neq 0$ so that $y_j = \lambda x_j$, for $j = 1, \ldots, n+1$. We can define a homeomorphism $\varphi_i$ of $U_i$ onto $\mathbb{R}^n$ as follows:

$$\varphi_i(x_1: \cdots: x_{n+1}) = \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{n+1}}{x_i}\right),$$

where the $i$th component is omitted. Again, it is clear that this map is well defined since it only involves ratios. We can also define the maps $\psi_i$ from $\mathbb{R}^n$ to $U_i \subseteq \mathbb{RP}^n$, given by

$$\psi_i(x_1, \ldots, x_n) = (x_1: \cdots: x_{i-1}: 1: x_i: \cdots: x_n),$$

where the 1 goes in the $i$th slot, for $i = 1, \ldots, n+1$.

One easily checks that $\varphi_i$ and $\psi_i$ are mutual inverses, so the $\varphi_i$ are homeomorphisms. On the overlap $U_i \cap U_j$, (where $i \neq j$), as $x_j \neq 0$, we have

$$(\varphi_j \circ \varphi_i^{-1})(x_1, \ldots, x_n) = \left(\frac{x_1}{x_j}, \ldots, \frac{x_{i-1}}{x_j}, \frac{1}{x_j}, \frac{x_i}{x_j}, \ldots, \frac{x_{j-1}}{x_j}, \frac{x_{j+1}}{x_j}, \ldots, \frac{x_{n}}{x_j}\right).$$

(We assumed that $i < j$; the case $j < i$ is similar.) This is clearly a smooth function from $\varphi_i(U_i \cap U_j)$ to $\varphi_j(U_i \cap U_j)$. As the $U_i$ cover $\mathbb{RP}^n$, we conclude that the $(U_i, \varphi_i)$ are $n+1$ charts making a smooth atlas for $\mathbb{RP}^n$. Intuitively, the space $\mathbb{RP}^n$ is obtained by gluing the open subsets $U_i$ on their overlaps. Even for $n = 3$, this is not easy to visualize!
Example 7.4. The Grassmannian $G(k, n)$.

Recall that $G(k, n)$ is the set of all $k$-dimensional linear subspaces of $\mathbb{R}^n$, also called $k$-planes. Every $k$-plane $W$ is the linear span of $k$ linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$; furthermore, $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ both span $W$ iff there is an invertible $k \times k$-matrix $\Lambda = (\lambda_{ij})$ such that

$$v_j = \sum_{i=1}^{k} \lambda_{ij} u_i, \quad 1 \leq j \leq k.$$

Obviously, there is a bijection between the collection of $k$ linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$ and the collection of $n \times k$ matrices of rank $k$. Furthermore, two $n \times k$ matrices $A$ and $B$ of rank $k$ represent the same $k$-plane iff

$$B = A\Lambda, \quad \text{for some invertible } k \times k \text{ matrix, } \Lambda.$$

(Note the analogy with projective spaces where two vectors $u, v$ represent the same point iff $v = \lambda u$ for some invertible $\lambda \in \mathbb{R}$.)

The set of $n \times k$ matrices of rank $k$ is a subset of $\mathbb{R}^{n \times k}$, in fact an open subset.

One can show that the equivalence relation on $n \times k$ matrices of rank $k$ given by

$$B = A\Lambda, \quad \text{for some invertible } k \times k \text{ matrix, } \Lambda,$$

is open, and that the graph of this equivalence relation is closed. For some help proving these facts, see Problem 7.2 in Tu [170]. By Proposition 3.32, the Grassmannian $G(k, n)$ is Hausdorff and second-countable.

We can define the domain of charts (according to Definition 7.2) on $G(k, n)$ as follows: For every subset $S = \{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$, let $U_S$ be the subset of $n \times k$ matrices $A$ of rank $k$ whose rows of index in $S = \{i_1, \ldots, i_k\}$ form an invertible $k \times k$ matrix denoted $A_S$. Note $U_S$ is open in quotient topology of $G(k, n)$ since the existence of an invertible $k \times k$ matrix is equivalent to the open condition of $\det A_S \neq 0$. Observe that the $k \times k$ matrix consisting of the rows of the matrix $AA_S^{-1}$ whose index belong to $S$ is the identity matrix $I_k$. Therefore, we can define a map $\varphi_S: U_S \to \mathbb{R}^{(n-k) \times k}$ where $\varphi_S(A)$ is equal to the $(n - k) \times k$ matrix obtained by deleting the rows of index in $S$ from $AA_S^{-1}$.

We need to check that this map is well defined, i.e., that it does not depend on the matrix $A$ representing $W$. Let us do this in the case where $S = \{1, \ldots, k\}$, which is notionally simpler. The general case can be reduced to this one using a suitable permutation.

If $B = A\Lambda$, with $\Lambda$ invertible, if we write

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

where $A_1$ and $B_1$ are $k \times k$ matrices and $A_2$ and $B_2$ are $(n-k) \times k$ matrices, as $B = A\Lambda$, we get $B_1 = A_1\Lambda$ and $B_2 = A_2\Lambda$, from which we deduce that

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} B_1^{-1} = \begin{pmatrix} I_k \\ A_2\Lambda^{-1}A_1^{-1} \end{pmatrix} = \begin{pmatrix} I_k \\ A_2\Lambda^{-1} \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} A_1^{-1}. $$
Therefore, our map is indeed well-defined.

Here is an example for \( n = 6 \) and \( k = 3 \). Let \( A \) be the matrix

\[
\begin{pmatrix}
2 & 3 & 5 \\
1 & 1 & 1 \\
1 & 1 & 0 \\
1 & -1 & 2 \\
1 & 0 & 0 \\
2 & -1 & 2
\end{pmatrix}
\]

and let

\( S = \{2, 3, 5\} \).

Then, we have

\[
A_{\{2,3,5\}} = \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 0 \\
1 & 0 & 0
\end{pmatrix},
\]

and we find that

\[
A_{\{2,3,5\}}^{-1} = \begin{pmatrix}
0 & 0 & 1 \\
0 & 1 & -1 \\
1 & -1 & 0
\end{pmatrix},
\]

and

\[
AA_{\{2,3,5\}}^{-1} = \begin{pmatrix}
5 & -2 & -1 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
2 & -3 & 2 \\
0 & 0 & 1 \\
2 & -3 & 3
\end{pmatrix}.
\]

Therefore,

\[
\varphi_{\{2,3,5\}}(A) = \begin{pmatrix}
5 & -2 & -1 \\
2 & -3 & 2 \\
2 & -3 & 3
\end{pmatrix}.
\]

We can define its inverse \( \psi_S \) as follows: Let \( \pi_S \) be the permutation of \( \{1, \ldots, n\} \) sending \( \{1, \ldots, k\} \) to \( S \) defined such that if \( S = \{i_1 < \cdots < i_k\} \), then \( \pi_S(j) = i_j \) for \( j = 1, \ldots, k \), and if \( \{h_1 < \cdots < h_{n-k}\} = \{1, \ldots, n\} - S \), then \( \pi_S(k+j) = h_j \) for \( j = 1, \ldots, n-k \) (this is a \( k \)-shuffle). If \( P_S \) is the permutation matrix associated with \( \pi_S \), for any \( (n-k) \times k \) matrix \( M \), let

\[
\psi_S(M) = P_S \left( \begin{pmatrix} I_k \\ M \end{pmatrix} \right).
\]
The effect of $\psi_S$ is to “insert into $M$” the rows of the identity matrix $I_k$ as the rows of index from $S$. Using our previous example where $n = 6, k = 3$ and $S = \{2, 3, 5\}$, we have

$$M = \begin{pmatrix} 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix},$$

the permutation $\pi_S$ is given by

$$\pi_S = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 5 & 1 & 4 & 6 \end{pmatrix},$$

whose permutation matrix is

$$P_{\{2, 3, 5\}} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\psi_{\{2, 3, 5\}}(M) = P_{\{2, 3, 5\}} \left( I_3 \begin{pmatrix} M \end{pmatrix} \right) = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix}.$$

Since the permutation $\pi_S$ is a $k$-shuffle that sends $\{1, \ldots, k\}$ to $S$, we see that $\varphi_S(A)$ is also obtained by first forming $P_S^{-1}A$, which brings the rows of index in $S$ to the first $k$ rows, then forming $P_S^{-1}A(P_S^{-1}A)^{-1}_{\{1, \ldots, k\}}$, and finally deleting the first $k$ rows. If we write $A$ and $P_S^{-1}$ in block form as

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad P_S^{-1} = \begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix},$$

with $A_1$ a $k \times k$ matrix, $A_2$ a $(n-k) \times k$ matrix, $P_1$ a $k \times k$ matrix, $P_4$ a $(n-k) \times (n-k)$ matrix, $P_2$ a $k \times (n-k)$ matrix, and $P_3$ a $(n-k) \times k$ matrix, then

$$P_S^{-1}A = \begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} P_1A_1 + P_2A_2 \\ P_3A_1 + P_4A_2 \end{pmatrix},$$

so

$$P_S^{-1}A(P_S^{-1}A)^{-1}_{\{1, \ldots, k\}} = \begin{pmatrix} P_1A_1 + P_2A_2 \\ P_3A_1 + P_4A_2 \end{pmatrix}(P_1A_1 + P_2A_2)^{-1} = \begin{pmatrix} I_k \\ (P_3A_1 + P_4A_2)(P_1A_1 + P_2A_2)^{-1} \end{pmatrix}.$$
and thus, \( \varphi_S(A) = (P_3A_1 + P_4A_2)(P_1A_1 + P_2A_2)^{-1} \).

With the above example,

\[
P_{\{2,3,5\}}^{-1} = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix},
\]

and then

\[
P_{\{2,3,5\}}^{-1} A_{\{2,3,5\}} = \begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 0 \\
1 & 0 & 0 \\
2 & 3 & 5 \\
1 & -1 & 2 \\
2 & -1 & 2
\end{pmatrix},
\]

\[
(P_{\{2,3,5\}}^{-1} A_{\{2,3,5\}})^{-1}_{\{1,2,3\}} = \begin{pmatrix}
0 & 0 & 1 \\
0 & 1 & -1 \\
1 & -1 & 0
\end{pmatrix},
\]

and

\[
P_{\{2,3,5\}}^{-1} A(P_{\{2,3,5\}}^{-1} A_{\{2,3,5\}})^{-1}_{\{1,2,3\}} = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
5 & -2 & -1 \\
2 & -3 & 2 \\
2 & -3 & 3
\end{pmatrix},
\]

which does yield

\[
\varphi_{\{2,3,5\}}(A) = \begin{pmatrix}
5 & -2 & -1 \\
2 & -3 & 2 \\
2 & -3 & 3
\end{pmatrix}.
\]

At this stage, we have charts that are bijections from subsets \( U_S \) of \( G(k,n) \) to open subsets, namely, \( \mathbb{R}^{(n-k)\times k} \). Then, the reader can check that the transition map \( \varphi_T \circ \varphi_S^{-1} \) from \( \varphi_S(U_S \cap U_T) \) to \( \varphi_T(U_S \cap U_T) \) is given by

\[
M \mapsto (P_3 + P_4M)(P_1 + P_2M)^{-1},
\]

where

\[
\begin{pmatrix}
P_1 & P_2 \\
P_3 & P_4
\end{pmatrix} = P_T^{-1} P_S
\]

is the matrix of the permutation \( \pi_T^{-1} \circ \pi_S \). This map is smooth, as the inversion of a matrix uses the cofactor matrix which relies on the smoothness of the determinant. and so the
Definition 7.5. Given any two \( a \in \mathbb{C} \) and \( \phi \in M \) on \( \mathbb{C} \) of the form \( V = S \) obtained by making \( S \in \text{Obtained} \) and \( SO \), observe that if \( U \in \mathbb{R} \), then using the Gram-Schmidt orthonormalization procedure, every basis \( B = (b_1, \ldots, b_k) \) for \( W \) yields an orthonormal basis \( U = (u_1, \ldots, u_k) \), and there is an invertible matrix \( \Lambda \) such that \( U = B\Lambda \), where the columns of \( B \) are the \( b_j \)’s and the columns of \( U \) are the \( u_j \)’s. Thus we may assume that the representatives of \( W \) are matrices \( U \) which have orthonormal columns and are characterized by the equation \( U^TU = I_k \).

The space of such matrices is closed and clearly bounded in \( \mathbb{R}^{n \times k} \), and thus compact. In fact, the space of \( n \times k \) matrices \( U \) satisfying \( U^TU = I \) is the Stiefel manifold \( St(k, n) \). Observe that if \( U \) and \( V \) are two \( n \times k \) matrices such that \( U^TU = I \) and \( V^TV = I \) and if \( V = UA \) for some invertible \( k \times k \) matrix \( \Lambda \), then \( \Lambda \in O(k) \). Then \( G(k, n) \) is the orbit space obtained by making \( O(k) \) act on \( St(k, n) \) on the right, i.e. \( S(k, n)/O(k) \cong G(k, n) \), and since \( S(k, n) \) is compact, we conclude that \( G(k, n) \) is also compact as it is the continuous image of a projection map.

Remark: The reader should have no difficulty proving that the collection of \( k \)-planes represented by matrices in \( U_S \) is precisely the set of \( k \)-planes \( W \) supplementary to the \( (n-k) \)-plane spanned by the canonical basis vectors \( e_{j_{k+1}}, \ldots, e_{j_n} \) (i.e., \( \text{span}(W \cup \{e_{j_{k+1}}, \ldots, e_{j_n}\}) = \mathbb{R}^n \), where \( S = \{i_1, \ldots, i_k\} \) and \( \{j_{k+1}, \ldots, j_n\} = \{1, \ldots, n\} - S \).

Example 7.5. Product Manifolds.

Let \( M_1 \) and \( M_2 \) be two \( C^k \)-manifolds of dimension \( n_1 \) and \( n_2 \), respectively. The topological space \( M_1 \times M_2 \) with the product topology (the opens of \( M_1 \times M_2 \) are arbitrary unions of sets of the form \( U \times V \), where \( U \) is open in \( M_1 \) and \( V \) is open in \( M_2 \)) can be given the structure of a \( C^k \)-manifold of dimension \( n_1 + n_2 \) by defining charts as follows: For any two charts \( (U_i, \varphi_i) \) on \( M_1 \) and \( (V_j, \psi_j) \) on \( M_2 \), we declare that \( (U_i \times V_j, \varphi_i \times \psi_j) \) is a chart on \( M_1 \times M_2 \), where \( \varphi_i \times \psi_j : U_i \times V_j \to \mathbb{R}^{n_1 + n_2} \) is defined so that

\[
\varphi_i \times \psi_j(p, q) = (\varphi_i(p), \psi_j(q)), \quad \text{for all } (p, q) \in U_i \times V_j.
\]

We define \( C^k \)-maps between manifolds as follows:

Definition 7.5. Given any two \( C^k \)-manifolds \( M \) and \( N \) of dimension \( m \) and \( n \) respectively, a \( C^k \)-map is a continuous function \( h : M \to N \) satisfying the following property: For every
$p \in M$, there is some chart $(U, \varphi)$ at $p$ and some chart $(V, \psi)$ at $q = h(p)$, with $h(U) \subseteq V$ and
\[
\psi \circ h \circ \varphi^{-1} : \varphi(U) \longrightarrow \psi(V)
\]
a $C^k$-function.

Figure 7.5: The $C^k$ map from $M$ to $N$, where the dimension of $M$ is 2 and the dimension of $N$ is 3.

It is easily shown that Definition 7.5 does not depend on the choice of charts.

The requirement in Definition 7.5 that $h : M \rightarrow N$ should be continuous is actually redundant. Indeed, since $\varphi$ and $\psi$ are homeomorphisms, $\varphi$ and $\psi^{-1}$ are continuous, and since $\varphi(U)$ is an open subset of $\mathbb{R}^m$ and $\psi(V)$ is an open subset of $\mathbb{R}^n$, the function $\psi \circ h \circ \varphi^{-1} : \varphi(U) \rightarrow \psi(V)$ being a $C^k$-function is continuous, so the restriction of $h$ to $U$ being equal to the composition of the three continuous maps
\[
\psi^{-1} \circ (\psi \circ h \circ \varphi^{-1}) \circ \varphi
\]
is also continuous on $U$. Since this holds on some open subset containing $p$, for every $p \in M$, the function $h$ is continuous on $M$. 
Other definitions of a smooth map appear in the literature, some requiring continuity. The following proposition from Berger and Gostiaux [20] (Theorem 2.3.3) helps clarifying how these definition relate.

**Proposition 7.1.** Let \( h: M \to N \) be a function between two manifolds \( M \) and \( N \). The following equivalences hold.

1. The map \( h \) is continuous, and for every \( p \in M \), for every chart \( (U, \varphi) \) at \( p \) and every chart \( (V, \psi) \) at \( h(p) \), the function \( \psi \circ h \circ \varphi^{-1} \) from \( \varphi(U \cap h^{-1}(V)) \) to \( \psi(V) \) is a \( C^k \)-function.

2. The map \( h \) is continuous, and for every \( p \in M \), for every chart \( (U, \varphi) \) at \( p \) and every chart \( (V, \psi) \) at \( h(p) \), if \( h(U) \subseteq V \), then the function \( \psi \circ h \circ \varphi^{-1} \) from \( \varphi(U) \) to \( \psi(V) \) is a \( C^k \)-function.

3. For every \( p \in M \), there is some chart \( (U, \varphi) \) at \( p \) and some chart \( (V, \psi) \) at \( q = h(p) \) with \( h(U) \subseteq V \), such that the function \( \psi \circ h \circ \varphi^{-1} \) from \( \varphi(U) \) to \( \psi(V) \) is a \( C^k \)-function.

Observe that (3) states exactly the conditions of Definition 7.5, with the continuity requirement omitted. Condition (1) is used by many texts. The continuity of \( h \) is required to ensure that \( h^{-1}(V) \) is an open set. The implication \((ii) \Rightarrow (iii)\) also requires the continuity of \( h \).

Even though the continuity requirement in Definition 7.5 is redundant, it seems to us that it does not hurt to emphasize that smooth maps are continuous.

In the special case where \( N = \mathbb{R} \), we obtain the notion of a \( C^k \)-function on \( M \). One checks immediately that a function \( f: M \to \mathbb{R} \) is a \( C^k \)-map iff for every \( p \in M \), there is some chart \( (U, \varphi) \) at \( p \) so that

\[
f \circ \varphi^{-1} : \varphi(U) \to \mathbb{R}
\]

is a \( C^k \)-function. If \( U \) is an open subset of \( M \), the set of \( C^k \)-functions on \( U \) is denoted by \( C^k(U) \). In particular, \( C^k(M) \) denotes the set of \( C^k \)-functions on the manifold, \( M \). Observe that \( C^k(U) \) is a commutative ring.

On the other hand, if \( M \) is an open interval of \( \mathbb{R} \), say \( M = (a, b) \), then \( \gamma: (a, b) \to N \) is called a \( C^k \)-curve in \( N \). One checks immediately that a function \( \gamma: (a, b) \to N \) is a \( C^k \)-map iff for every \( q \in N \), there is some chart \( (V, \psi) \) at \( q \) and some open subinterval \( (c, d) \) of \( (a, b) \), so that \( \gamma((c, d)) \subseteq V \) and

\[
\psi \circ \gamma: (c, d) \to \psi(V)
\]

is a \( C^k \)-function.

It is clear that the composition of \( C^k \)-maps is a \( C^k \)-map. A \( C^k \)-map \( h: M \to N \) between two manifolds is a \( C^k \)-diffeomorphism iff \( h \) has an inverse \( h^{-1}: N \to M \) (i.e., \( h^{-1} \circ h = \text{id}_M \) and \( h \circ h^{-1} = \text{id}_N \)), and both \( h \) and \( h^{-1} \) are \( C^k \)-maps (in particular, \( h \) and \( h^{-1} \) are homeomorphisms). Next, we define tangent vectors.
7.2 Tangent Vectors, Tangent Spaces

Let $M$ be a $C^k$ manifold of dimension $n$, with $k \geq 1$. The purpose of the next three sections is to define the tangent space $T_p(M)$, at a point $p$ of a manifold $M$. We provide three definitions of the notion of a tangent vector to a manifold and prove their equivalence.

The first definition uses equivalence classes of curves on a manifold and is the most intuitive.

The second definition makes heavy use of the charts and of the transition functions. It is also quite intuitive and it is easy to see that it is equivalent to the first definition. The second definition is the most convenient one to define the manifold structure of the tangent bundle $T(M)$ (see Section 8.1).

The third definition (given in the next section) is based on the view that a tangent vector $v$, at $p$, induces a differential operator on real-valued functions $f$, defined locally near $p$; namely, the map $f \mapsto v(f)$ is a linear form satisfying an additional property akin to the rule for taking the derivative of a product (the Leibniz property). Such linear forms are called point-derivations. This third definition is more intrinsic than the first two but more abstract. However, for any point $p$ on the manifold $M$ and for any chart whose domain contains $p$, there is a convenient basis of the tangent space $T_p(M)$. The third definition is also the most convenient one to define vector fields. A few technical complications arise when $M$ is not a smooth manifold (when $k \neq \infty$) but these are easily overcome using “stationary germs.”

As pointed out by Serre in [159] (Chapter III, Section 8), the relationship between the first definition and the third definition of the tangent space at $p$ is best described by a nondegenerate pairing which shows that $T_p(M)$ is the dual of the space of point-derivations at $p$ that vanish on stationary germs. This pairing is presented in Section 7.4.

The most intuitive method to define tangent vectors is to use curves. Let $p \in M$ be any point on $M$ and let $\gamma: (-\epsilon, \epsilon) \to M$ be a $C^1$-curve passing through $p$, that is, with $\gamma(0) = p$. Unfortunately, if $M$ is not embedded in any $\mathbb{R}^N$, the derivative $\gamma'(0)$ does not make sense. However, for any chart, $(U, \varphi)$, at $p$, the map $\varphi \circ \gamma$ is a $C^1$-curve in $\mathbb{R}^n$ and the tangent vector $v = (\varphi \circ \gamma)'(0)$ is well defined. The trouble is that different curves may yield the same $v$!

To remedy this problem, we define an equivalence relation on curves through $p$ as follows:

**Definition 7.6.** Given a $C^k$ manifold, $M$, of dimension $n$, for any $p \in M$, two $C^1$-curves, $\gamma_1: (-\epsilon_1, \epsilon_1) \to M$ and $\gamma_2: (-\epsilon_2, \epsilon_2) \to M$, through $p$ (i.e., $\gamma_1(0) = \gamma_2(0) = p$) are equivalent iff there is some chart, $(U, \varphi)$, at $p$ so that

$$(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0).$$

See Figure 7.6.
Now, the problem is that this definition seems to depend on the choice of the chart. Fortunately, this is not the case. For, if \((V, \psi)\) is another chart at \(p\), as \(p\) belongs both to \(U\) and \(V\), we have \(U \cap V \neq \emptyset\), so the transition function \(\eta = \psi \circ \varphi^{-1}\) is \(C^k\) and, by the chain rule, we have

\[
(\psi \circ \gamma_1)'(0) = (\eta \circ \varphi \circ \gamma_1)'(0) \\
= \eta'(\varphi(p))((\varphi \circ \gamma_1)'(0)) \\
= \eta'(\varphi(p))((\varphi \circ \gamma_2)'(0)) \\
= (\eta \circ \varphi \circ \gamma_2)'(0) \\
= (\psi \circ \gamma_2)'(0).
\]

This leads us to the first definition of a tangent vector.

**Definition 7.7.** (Tangent Vectors, Version 1) Given any \(C^k\)-manifold, \(M\), of dimension \(n\), with \(k \geq 1\), for any \(p \in M\), a **tangent vector to \(M\) at \(p\)** is any equivalence class \(u = [\gamma]\) of \(C^1\)-curves \(\gamma\) through \(p\) on \(M\), modulo the equivalence relation defined in Definition 7.6. The set of all tangent vectors at \(p\) is denoted by \(T_p(M)\) (or \(T_pM\)).

In order to make \(T_pM\) into a vector space, given a chart \((U, \varphi)\) with \(p \in U\), we observe that the map \(\varphi_U : T_pM \to \mathbb{R}^n\) given by

\[
\varphi_U([\gamma]) = (\varphi \circ \gamma)'(0)
\]
is a bijection, where \([\gamma]\) is the equivalence class of a curve \(\gamma\) in \(M\) through \(p\) (with \(\gamma(0) = p\)). The map \(\varphi_U\) is injective by definition of the equivalence relation on curves; it is surjective, because for every vector \(v \in \mathbb{R}^n\), if \(\gamma_v\) is the curve given by \(\gamma_v(t) = \varphi^{-1}(\varphi(p) + tv)\), then \((\varphi \circ \gamma_v)'(0) = v\), and so \(\varphi_U([\gamma_v]) = v\).

Observe that for any chart \((U, \varphi)\) at \(p\), the equivalence class \([\gamma]\) of all curves through \(p\) such that \((\varphi \circ \gamma)'(0) = v\) for some given vector \(v \in \mathbb{R}^n\) is determined by the special curve \(\gamma_v\) defined above.

The vector space structure on \(T_pM\) is defined as follows. For any chart \((U, \varphi)\) at \(p\), given any two equivalences classes \([\gamma_1]\) and \([\gamma_2]\) in \(T_pM\), for any real \(\lambda\), we set
\[
[\gamma_1] + [\gamma_2] = \varphi_U^{-1}(\varphi_U([\gamma_1]) + \varphi_U([\gamma_2]))
\]
\[
\lambda [\gamma_1] = \varphi_U^{-1}(\lambda \varphi_U([\gamma_1])).
\]
If \((V, \psi)\) is any other chart at \(p\), since by the chain rule
\[
(\psi \circ \gamma)'(0) = (\psi \circ \varphi^{-1})'_{\varphi(p)} \circ (\varphi \circ \gamma)'(0),
\]
it follows that
\[
\overline{\psi}_N = (\psi \circ \varphi^{-1})'_{\varphi(p)} \circ \varphi_U.
\]
Since \((\psi \circ \varphi^{-1})'_{\varphi(p)}\) is a linear isomorphism, we see that the vector space structure defined above does not depend on the choice of chart at \(p\). Therefore, with this vector space structure on \(T_pM\), the map \(\varphi_U: T_pM \to \mathbb{R}^n\) is a linear isomorphism. This shows that \(T_pM\) is a vector space of dimension \(n = \text{dimension of } M\).

In particular, if \(M\) is an \(n\)-dimensional smooth manifold in \(\mathbb{R}^N\) and if \(\gamma\) is a curve in \(M\) through \(p\), then \(\gamma'(0) = u\) is well defined as a vector in \(\mathbb{R}^N\), and the equivalence class of all curves \(\gamma\) through \(p\) such that \((\varphi \circ \gamma)'(0)\) is the same vector in some chart \(\varphi: U \to \Omega\) can be identified with \(u\). Thus, the tangent space \(T_pM\) to \(M\) at \(p\) is isomorphic to
\[
\{\gamma'(0) \mid \gamma: (-\epsilon, \epsilon) \to M\ \text{is a C}^1\text{-curve with } \gamma(0) = p\}.
\]

In the special case of a linear Lie group \(G\), Proposition 4.10 shows that the exponential map \(\exp: \mathfrak{g} \to G\) is a diffeomorphism from some open subset of \(\mathfrak{g}\) containing 0 to some open subset of \(G\) containing \(I\). For every \(g \in G\), since \(L_g\) is a diffeomorphism, the map \(L_g \circ \exp: L_g(\mathfrak{g}) \to G\) is a diffeomorphism from some open subset of \(L_g(\mathfrak{g})\) containing 0 to some open subset of \(G\) containing \(g\). Furthermore,
\[
L_g(\mathfrak{g}) = g \mathfrak{g} = \{gX \mid X \in \mathfrak{g}\}.
\]
Thus, we obtain smooth parametrizations of \(G\) whose inverses are charts on \(G\), and since by definition of \(\mathfrak{g}\), for every \(X \in \mathfrak{g}\), the curve \(\gamma(t) = ge^{tX}\) is a curve through \(g\) in \(G\) such that \(\gamma'(0) = gX\), we see that the tangent space \(T_gG\) to \(G\) at \(g\) is isomorphic to \(gg\).
One should observe that unless $M = \mathbb{R}^n$, in which case, for any $p, q \in \mathbb{R}^n$, the tangent space $T_q(M)$ is naturally isomorphic to the tangent space $T_p(M)$ by the translation $q - p$, for an arbitrary manifold, there is no relationship between $T_p(M)$ and $T_q(M)$ when $p \neq q$.

The second way of defining tangent vectors has the advantage that it makes it easier to define tangent bundles (see Section 8.1).

**Definition 7.8.** (Tangent Vectors, Version 2) Given any $C^k$-manifold, $M$, of dimension $n$, with $k \geq 1$, for any $p \in M$, consider the triples, $(U, \varphi, u)$, where $(U, \varphi)$ is any chart at $p$ and $u$ is any vector in $\mathbb{R}^n$. Say that two such triples $(U, \varphi, u)$ and $(V, \psi, v)$ are equivalent iff 

$$(\psi \circ \varphi^{-1})'_{\varphi(p)}(u) = v.$$

See Figure 7.7. A tangent vector to $M$ at $p$ is an equivalence class of triples, $[(U, \varphi, u)]$, for the above equivalence relation.

![Figure 7.7: Two equivalent tangent vector $u$ and $v$](image)
for simplicity of notation if we denote the equivalence class of the triple \((U, \varphi, u)\) by \([u]\), we set
\[
[u] + [v] = \theta_{U,\varphi,p}^{-1}(\theta_{U,\varphi,p}^{-1}([u]) + \theta_{U,\varphi,p}^{-1}([v]))
\]
\[
\lambda[u] = \theta_{U,\varphi,p}^{-1}(\lambda \theta_{U,\varphi,p}^{-1}([u])).
\]
Since the equivalence between triples \((U, \varphi, u)\) and \((V, \psi, v)\) is given by
\[
(\psi \circ \varphi^{-1})'(0)(u) = v,
\]
we have
\[
\theta_{V,\psi,p}^{-1} = (\psi \circ \varphi^{-1})'(0) \circ \theta_{U,\varphi,p}^{-1},
\]
so the vector space structure on \(T_p M\) does not depend on the choice of chart at \(p\).

The equivalence of this definition with the definition in terms of curves (Definition 7.7) is easy to prove.

**Proposition 7.2.** Let \(M\) be any \(C^k\)-manifold of dimension \(n\), with \(k \geq 1\). For every \(p \in M\), for every chart, \((U, \varphi)\), at \(p\), if \(x = [\gamma]\) is any tangent vector (version 1) given by some equivalence class of \(C^1\)-curves \(\gamma: (-\epsilon, +\epsilon) \to M\) through \(p\) (i.e., \(p = \gamma(0)\)), then the map
\[
x \mapsto [(U, \varphi, (\varphi \circ \gamma)'(0))]
\]
is an isomorphism between \(T_p(M)\)-version 1 and \(T_p(M)\)-version 2.

**Proof.** If \(\sigma\) is another curve equivalent to \(\gamma\), then \((\varphi \circ \gamma)'(0) = (\varphi \circ \sigma)'(0)\), so the map is well-defined. It is clearly injective. As for surjectivity, define the curve \(\gamma_v\) on \(M\) through \(p\) by
\[
\gamma_v(t) = \varphi^{-1}(\varphi(p) + tu);
\]
see Figure 7.8. Then, \((\varphi \circ \gamma_v)(t) = \varphi(p) + tu\) and
\[
(\varphi \circ \gamma_v)'(0) = u,
\]
as desired. \(\square\)

### 7.3 Tangent Vectors as Derivations

One of the defects of the above definitions of a tangent vector is that it has no clear relation to the \(C^k\)-differential structure of \(M\). In particular, the definition does not seem to have anything to do with the functions defined locally at \(p\). There is another way to define tangent vectors that reveals this connection more clearly. Moreover, such a definition is more intrinsic, i.e., does not refer explicitly to charts. Our presentation of this second approach is heavily inspired by Schwartz [156] (Chapter 3, Section 9) but also by Warner [175] and Serre [159] (Chapter III, Sections 7 and 8).
Figure 7.8: The tangent vector \( u \) is in one-to-one correspondence with the line through \( \varphi(p) \) with direction \( u \).

As a first step, consider the following: Let \((U, \varphi)\) be a chart at \( p \in M \) (where \( M \) is a \( C^k \)-manifold of dimension \( n \), with \( k \geq 1 \) and let \( x_i = pr_i \circ \varphi \), the \( i \)th local coordinate \( (1 \leq i \leq n) \). For any real-valued function \( f \) defined on \( p \in U \), set

\[
\left( \frac{\partial}{\partial x_i} \right)_p f = \frac{\partial(f \circ \varphi^{-1})}{\partial X_i} \bigg|_{\varphi(p)}, \quad 1 \leq i \leq n.
\]

(Here, \((\partial g/\partial X_i)\big|_y \) denotes the partial derivative of a function \( g: \mathbb{R}^n \to \mathbb{R} \) with respect to the \( i \)th coordinate, evaluated at \( y \).)

We would expect that the function that maps \( f \) to the above value is a linear map on the set of functions defined locally at \( p \), but there is technical difficulty: The set of real-valued functions defined locally at \( p \) is not a vector space! To see this, observe that if \( f \) is defined on an open \( p \in U \) and \( g \) is defined on a different open \( p \in V \), then we do not know how to define \( f + g \). The problem is that we need to identify functions that agree on a smaller open subset. This leads to the notion of **germs**.

**Definition 7.9.** Given any \( C^k \)-manifold \( M \) of dimension \( n \), with \( k \geq 1 \), for any \( p \in M \), a **locally defined function at \( p \)** is a pair \((U, f)\), where \( U \) is an open subset of \( M \) containing \( p \) and \( f \) is a real-valued function defined on \( U \). Two locally defined functions \((U, f)\) and \((V, g)\) at \( p \) are **equivalent** iff there is some open subset \( W \subseteq U \cap V \) containing \( p \), so that

\[
f \upharpoonright W = g \upharpoonright W.
\]
The equivalence class of a locally defined function at \( p \), denoted \([f] \) or \( f \), is called a germ at \( p \).

One should check that the relation of Definition 7.9 is indeed an equivalence relation. Of course, the value at \( p \) of all the functions \( f \) in any germ \( f \), is \( f(p) \). Thus, we set \( f(p) = f(p) \), for any \( f \in \mathfrak{f} \).

For example, for every \( a \in (-1, 1) \), the locally defined functions \((\mathbb{R} - \{1\}, 1/(1 - x)) \) and \((-1, 1), \sum_{n=0}^\infty x^n \) at \( a \) are equivalent.

We can define addition of germs, multiplication of a germ by a scalar and multiplication of germs, as follows. If \((U, f)\) and \((V, g)\) are two locally defined functions at \( p \), we define \((U \cap V, f + g), (U \cap V, fg)\) and \((U, \lambda f)\) as the locally defined functions at \( p \) given by \((f + g)(q) = f(q) + g(q)\) and \((fg)(q) = f(q)g(q)\) for all \( q \in U \cap V \), and \((\lambda f)(q) = \lambda f(q)\) for all \( q \in U \), with \( \lambda \in \mathbb{R} \). Then, if \( f = [f] \) and \( g = [g] \) are two germs at \( p \), we define

\[
[f] + [g] = [f + g] \\
\lambda[f] = [\lambda f] \\
[f][g] = [fg].
\]

However, we have to check that these definitions make sense, that is, that they don’t depend on the choice of representatives chosen in the equivalence classes \([f]\) and \([g]\). Let us give the details of this verification for the sum of two germs, \([f]\) and \([g]\).

We need to check that for any locally defined functions \((U_1, f_1), (U_2, f_2), (V_1, g_1)\), and \((V_2, g_2)\), at \( p \), if \((U_1, f_1)\) and \((U_2, f_2)\) are equivalent and if \((V_1, g_1)\) and \((V_2, g_2)\) are equivalent, then \((U_1 \cap V_1, f_1 + g_1)\) and \((U_2 \cap V_2, f_2 + g_2)\) are equivalent. However, as \((U_1, f_1)\) and \((U_2, f_2)\) are equivalent, there is some \( W_1 \subseteq U_1 \cap U_2 \) so that \( f_1 \upharpoonright W_1 = f_2 \upharpoonright W_1 \) and as \((V_1, g_1)\) and \((V_2, g_2)\) are equivalent, there is some \( W_2 \subseteq V_1 \cap V_2 \) so that \( g_1 \upharpoonright W_2 = g_2 \upharpoonright W_2 \). Then, observe that \((f_1 + g_1) \upharpoonright (W_1 \cap W_2) = (f_2 + g_2) \upharpoonright (W_1 \cap W_2)\), which means that \([f_1 + g_1] = [f_2 + g_2]\). Therefore, \([f + g]\) does not depend on the representatives chosen in the equivalence classes \([f]\) and \([g]\) and it makes sense to set

\[
[f] + [g] = [f + g].
\]

We can proceed in a similar fashion to define \(\lambda[f]\) and \([f][g]\). Therefore, the germs at \( p \) form a ring.

The commutative ring of germs of \(C^k\)-functions at \( p \) is denoted \(\mathcal{O}^{(k)}_{M,p} \). When \( k = \infty \), we usually drop the superscript \( \infty \).

Remark: Most readers will most likely be puzzled by the notation \(\mathcal{O}^{(k)}_{M,p}\). In fact, it is standard in algebraic geometry, but it is not as commonly used in differential geometry. For any open subset \( U \) of a manifold \( M \), the ring \(C^k(U)\) of \(C^k\)-functions on \( U \) is also denoted \(\mathcal{O}^{(k)}_M(U)\) (certainly by people with an algebraic geometry bent!). Then, it turns out that the
7.3. TANGENT VECTORS AS DERIVATIONS

map $U \mapsto \mathcal{O}^{(k)}_M(U)$ is a sheaf, denoted $\mathcal{O}^{(k)}_M$, and the ring $\mathcal{O}^{(k)}_{M,p}$ is the stalk of the sheaf $\mathcal{O}^{(k)}_M$ at $p$. Such rings are called local rings. Roughly speaking, all the “local” information about $M$ at $p$ is contained in the local ring $\mathcal{O}^{(k)}_{M,p}$. (This is to be taken with a grain of salt. In the $C^k$-case where $k < \infty$, we also need the “stationary germs,” as we will see shortly.)

Now that we have a rigorous way of dealing with functions locally defined at $p$, observe that the map

$$v_i : f \mapsto \left( \frac{\partial}{\partial x_i} \right)_p f$$

yields the same value for all functions $f$ in a germ $f$ at $p$. Furthermore, the above map is linear on $\mathcal{O}^{(k)}_{M,p}$. More is true:

1. For any two functions $f, g$ locally defined at $p$, we have

$$\left( \frac{\partial}{\partial x_i} \right)_p (fg) = \left( \left( \frac{\partial}{\partial x_i} \right)_p f \right) g(p) + f(p) \left( \frac{\partial}{\partial x_i} \right)_p g.$$

2. If $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, then

$$\left( \frac{\partial}{\partial x_i} \right)_p f = 0.$$

The first property says that $v_i$ is a point derivation; it is also known as the Leibniz property. As to the second property, when $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, we say that $f$ is stationary at $p$.

It is easy to check (using the chain rule) that being stationary at $p$ does not depend on the chart $(U, \varphi)$ at $p$ or on the function chosen in a germ $f$. Therefore, the notion of a stationary germ makes sense.

**Definition 7.10.** We say that a germ $f$ at $p \in M$ is a stationary germ iff $(f \circ \varphi^{-1})'(\varphi(p)) = 0$ for some chart $(U, \varphi)$ at $p$ and some function $f$ in the germ $f$. The $C^k$-stationary germs form a subring of $\mathcal{O}^{(k)}_{M,p}$ (but not an ideal) denoted $\mathcal{S}^{(k)}_{M,p}$.

Remarkably, it turns out that the set of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$ is isomorphic to the tangent space $T_p(M)$. First, we prove that this space has $(\frac{\partial}{\partial x_1})_p, \ldots, (\frac{\partial}{\partial x_n})_p$ as a basis.

**Proposition 7.3.** Given any $C^k$-manifold $M$ of dimension $n$, with $k \geq 1$, for any $p \in M$ and any chart $(U, \varphi)$ at $p$, the $n$ functions $(\frac{\partial}{\partial x_1})_p, \ldots, (\frac{\partial}{\partial x_n})_p$ defined on $\mathcal{O}^{(k)}_M$ by

$$\left( \frac{\partial}{\partial x_i} \right)_p f = \left. \frac{\partial(f \circ \varphi^{-1})}{\partial X_i} \right|_{\varphi(p)} \quad 1 \leq i \leq n,$$
are linear forms that vanish on $S_{M,p}^{(k)}$. Every linear form $L$ on $O_{M,p}^{(k)}$ that vanishes on $S_{M,p}^{(k)}$ can be expressed in a unique way as

$$L = \sum_{i=1}^{n} \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p,$$

where $\lambda_i \in \mathbb{R}$. Therefore, the linear forms

$$\left( \frac{\partial}{\partial x_1} \right)_p, \ldots, \left( \frac{\partial}{\partial x_n} \right)_p$$

form a basis of the vector space of linear forms on $O_{M,p}^{(k)}$ that vanish on $S_{M,p}^{(k)}$.

**Proof.** The first part of the proposition is trivial by definition of $(\partial/\partial x_i)_p$, since for a stationary germ $f$, we have $(f \circ \varphi^{-1})'(\varphi(p)) = 0$.

Next, assume that $L$ is a linear form on $O_{M,p}^{(k)}$ that vanishes on $S_{M,p}^{(k)}$. For any function $(U,f)$ locally defined at $p$, consider the function $(U,g)$ locally defined at $p$ given by

$$g(q) = f(q) - \sum_{i=1}^{n} (pr_i \circ \varphi)(q) \left( \frac{\partial}{\partial x_i} \right)_p f, \quad q \in U.$$

Observe that the germ of $g$ is stationary at $p$. Indeed, if we let $X = \varphi(q)$, then $q = \varphi^{-1}(X)$, and we can write

$$(g \circ \varphi^{-1})(X) = (f \circ \varphi^{-1})(X) - \sum_{i=1}^{n} pr_i(X) \left( \frac{\partial}{\partial x_i} \right)_p f$$

$$= (f \circ \varphi^{-1})(X_1, \ldots, X_n) - \sum_{i=1}^{n} X_i \left( \frac{\partial}{\partial x_i} \right)_p f.$$

By definition it follows that

$$\left. \frac{\partial (g \circ \varphi^{-1})}{\partial X_i} \right|_{\varphi(p)} = \left. \frac{\partial (f \circ \varphi^{-1})}{\partial X_i} \right|_{\varphi(p)} = \left( \frac{\partial}{\partial x_i} \right)_p f = 0.$$

But then, as $L$ vanishes on stationary germs and the germ of

$$g = f - \sum_{i=1}^{n} (pr_i \circ \varphi) \left( \frac{\partial}{\partial x_i} \right)_p f$$

is stationary at $p$, we have $L(g) = 0$, so

$$L(f) = \sum_{i=1}^{n} L(pr_i \circ \varphi) \left( \frac{\partial}{\partial x_i} \right)_p f,$$
as desired. We still have to prove linear independence. If
\[ \sum_{i=1}^{n} \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p = 0, \]
then if we apply this relation to the functions \( x_i = pr_i \circ \varphi \), as \( \left( \frac{\partial}{\partial x_i} \right)_p x_j = \delta_{ij} \),
we get \( \lambda_i = 0 \), for \( i = 1, \ldots, n \).

To define our third version of tangent vectors, we need to define point-derivations.

**Definition 7.11.** Given any \( C^k \)-manifold \( M \) of dimension \( n \), with \( k \geq 1 \), for any \( p \in M \), a \textit{derivation at} \( p \) \textit{in} \( M \) or \textit{point-derivation on} \( \mathcal{O}^{(k)}_{M,p} \) is a linear form \( v \) on \( \mathcal{O}^{(k)}_{M,p} \), such that

\[ v(fg) = v(f)g(p) + f(p)v(g), \]
for all germs \( f, g \in \mathcal{O}^{(k)}_{M,p} \). The above is called the \textit{Leibniz property}. Let \( \mathcal{D}^{(k)}_p(M) \) denote the set of point-derivations on \( \mathcal{O}^{(k)}_{M,p} \).

As expected, point-derivations vanish on constant functions.

**Proposition 7.4.** Every point-derivation \( v \) on \( \mathcal{O}^{(k)}_{M,p} \) vanishes on germs of constant functions.

**Proof.** If \( g \) is a germ of constant functions at \( p \), then there is some \( \lambda \in \mathbb{R} \) so that \( g = \lambda \) (a constant function with value \( \lambda \)) for all \( g \in g \). Since \( v \) is linear,
\[ v(g) = v(\lambda 1) = \lambda v(1), \]
where \( 1 \) is the germ of constant functions with value 1, so we just have to show that \( v(1) = 0 \). However, because \( 1 = 1 \cdot 1 \) and \( v \) is a point-derivation, we get
\[ v(1) = v(1 \cdot 1) = v(1)1(p) + 1(p)v(1) = v(1)1 + 1v(1) = 2v(1) \]
from which we conclude that \( v(1) = 0 \), as claimed.

Recall that we observed earlier that the \( \left( \frac{\partial}{\partial x_i} \right)_p \) are point-derivations at \( p \). Therefore, we have

**Proposition 7.5.** Given any \( C^k \)-manifold \( M \) of dimension \( n \), with \( k \geq 1 \), for any \( p \in M \), the linear forms on \( \mathcal{O}^{(k)}_{M,p} \) that vanish on \( \mathcal{S}^{(k)}_{M,p} \) are exactly the point-derivations on \( \mathcal{O}^{(k)}_{M,p} \) that vanish on \( \mathcal{S}^{(k)}_{M,p} \).
Proof. By Proposition 7.3, 
\[ \left( \frac{\partial}{\partial x_1} \right)_p, \ldots, \left( \frac{\partial}{\partial x_n} \right)_p \]
form a basis of the linear forms on \( O_{M,p}^{(k)} \) that vanish on \( S_{M,p}^{(k)} \). Since each \( \left( \frac{\partial}{\partial x_i} \right)_p \) is also a point-derivation at \( p \), the result follows.

Remark: Proposition 7.5 says that any linear form on \( O_{M,p}^{(k)} \) that vanishes on \( S_{M,p}^{(k)} \) belongs to \( D_{p}^{(k)}(M) \), the set of point-derivations on \( O_{M,p}^{(k)} \). However, in general, when \( k \neq \infty \), a point-derivation on \( O_{M,p}^{(k)} \) does not necessarily vanish on \( S_{M,p}^{(k)} \). We will see in Proposition 7.9 that this is true for \( k = \infty \).

Here is now our third definition of a tangent vector.

Definition 7.12. (Tangent Vectors, Version 3) Given any \( C^k \)-manifold \( M \) of dimension \( n \), with \( k \geq 1 \), for any \( p \in M \), a tangent vector to \( M \) at \( p \) is any point-derivation on \( O_{M,p}^{(k)} \) that vanishes on \( S_{M,p}^{(k)} \), the subspace of stationary germs.

Let us consider the simple case where \( M = \mathbb{R} \). In this case, for every \( x \in \mathbb{R} \), the tangent space \( T_x(\mathbb{R}) \) is a one-dimensional vector space isomorphic to \( \mathbb{R} \) and \( \left( \frac{\partial}{\partial t} \right)_x = \frac{d}{dt}|_x \) is a basis vector of \( T_x(\mathbb{R}) \). For every \( C^k \)-function \( f \) locally defined at \( x \), we have
\[ \left( \frac{\partial}{\partial t} \right)_x f = \left. \frac{df}{dt} \right|_x = f'(x). \]
Thus, \( \left( \frac{\partial}{\partial t} \right)_x \) is: compute the derivative of a function at \( x \).

We now prove the equivalence of version 1 and version 3 of a tangent vector.

Proposition 7.6. Let \( M \) be any \( C^k \)-manifold of dimension \( n \), with \( k \geq 1 \). For any \( p \in M \), let \( u \) be any tangent vector (version 1) given by some equivalence class of \( C^1 \)-curves \( \gamma: (-\epsilon, +\epsilon) \rightarrow M \) through \( p \) (i.e., \( p = \gamma(0) \)). Then, the map \( L_u \) defined on \( O_{M,p}^{(k)} \) by
\[ L_u(f) = (f \circ \gamma)'(0) \]
is a point-derivation that vanishes on \( S_{M,p}^{(k)} \). Furthermore, the map \( u \mapsto L_u \) defined above is an isomorphism between \( T_p(M) \) and the space of linear forms on \( O_{M,p}^{(k)} \) that vanish on \( S_{M,p}^{(k)} \).

Proof. (After L. Schwartz) Clearly, \( L_u(f) \) does not depend on the representative \( f \) chosen in the germ \( f \). If \( \gamma \) and \( \sigma \) are equivalent curves defining \( u \), then \( (\varphi \circ \sigma)'(0) = (\varphi \circ \gamma)'(0) \), so from the chain rule we get
\[ (f \circ \sigma)'(0) = (f \circ \varphi^{-1})'(\varphi(p))((\varphi \circ \sigma)'(0)) = (f \circ \varphi^{-1})'(\varphi(p))((\varphi \circ \gamma)'(0)) = (f \circ \gamma)'(0), \]
which shows that $L_u(f)$ does not depend on the curve $\gamma$ defining $u$. If $f$ is a stationary germ, then pick any chart $(U, \varphi)$ at $p$, and let $\psi = \varphi \circ \gamma$. We have

$$L_u(f) = (f \circ \gamma)'(0) = ((f \circ \varphi^{-1}) \circ (\varphi \circ \gamma))'(0) = (f \circ \varphi^{-1})'(\varphi(p))(\psi'(0)) = 0,$$

since $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, as $f$ is a stationary germ. The definition of $L_u$ makes it clear that $L_u$ is a point-derivation at $p$. If $u \neq v$ are two distinct tangent vectors, then there exist some curves $\gamma$ and $\sigma$ through $p$ so that

$$(\varphi \circ \gamma)'(0) \neq (\varphi \circ \sigma)'(0).$$

Thus, there is some $i$, with $1 \leq i \leq n$, so that if we let $f = pr_i \circ \varphi$, then

$$(f \circ \gamma)'(0) \neq (f \circ \sigma)'(0),$$

and so, $L_u \neq L_v$. This proves that the map $u \mapsto L_u$ is injective.

For surjectivity, recall that every linear map $L$ on $O^{(k)}_{M,p}$ that vanishes on $S^{(k)}_{M,p}$ can be uniquely expressed as

$$L = \sum_{i=1}^{n} \lambda_i \left( \frac{\partial}{\partial x_i} \right)_{p}. $$

Define the curve $\gamma$ on $M$ through $p$ by

$$\gamma(t) = \varphi^{-1}(\varphi(p) + t(\lambda_1, \ldots, \lambda_n)),$$

for $t$ in a small open interval containing 0. See Figure 7.8. Then, we have

$$f(\gamma(t)) = (f \circ \varphi^{-1})(\varphi(p) + t(\lambda_1, \ldots, \lambda_n)),$$

and by the chain rule we get

$$(f \circ \gamma)'(0) = (f \circ \varphi^{-1})'(\varphi(p))(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} \lambda_i \frac{\partial(f \circ \varphi^{-1})}{\partial X_i} \bigg|_{\varphi(p)} = L(f).$$

This proves that $T_p(M)$ is isomorphic to the space of linear forms on $O^{(k)}_{M,p}$ that vanish on $S^{(k)}_{M,p}$.

We show in the next section that the the space of linear forms on $O^{(k)}_{M,p}$ that vanish on $S^{(k)}_{M,p}$ is isomorphic to $(O^{(k)}_{M,p}/S^{(k)}_{M,p})^*$ (the dual of the quotient space $O^{(k)}_{M,p}/S^{(k)}_{M,p}$).

Even though this is just a restatement of Proposition 7.3, we state the following proposition because of its practical usefulness:
Proposition 7.7. Given any $C^k$-manifold $M$ of dimension $n$, with $k \geq 1$, for any $p \in M$ and any chart $(U, \varphi)$ at $p$, the $n$ tangent vectors
\[
\left( \frac{\partial}{\partial x_1} \right)_p, \ldots, \left( \frac{\partial}{\partial x_n} \right)_p
\]
form a basis of $T_pM$.

When $M$ is a smooth manifold, things get a little simpler. Indeed, it turns out that in this case, every point-derivation vanishes on stationary germs. To prove this, we recall the following result from calculus (see Warner [175]):

Proposition 7.8. If $g : \mathbb{R}^n \to \mathbb{R}$ is a $C^k$-function ($k \geq 2$) on a convex open $U$ about $p \in \mathbb{R}^n$, then for every $q \in U$, we have
\[
g(q) = g(p) + \sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(q)(x_i(q) - x_i(p)) + \sum_{i,j=1}^{n} (x_i(q) - x_i(p))(x_j(q) - x_j(p)) \int_0^1 (1 - t) \frac{\partial^2 g}{\partial x_i \partial x_j}(1-t)p+ tq dt.
\]
In particular, if $g \in C^\infty(U)$, then the integral as a function of $q$ is $C^\infty$.

Proposition 7.9. Let $M$ be any $C^\infty$-manifold of dimension $n$. For any $p \in M$, any point-derivation on $\mathcal{O}^{(\infty)}_{M,p}$ vanishes on $\mathcal{S}^{(\infty)}_{M,p}$, the ring of stationary germs. Consequently, $T_p(M) = \mathcal{D}^{(\infty)}_{p}(M)$.

Proof. Pick some chart $(U, \varphi)$ at $p$, where $U$ is convex (for instance, an open ball) and let $f$ be any stationary germ. If we apply Proposition 7.8 to $f \circ \varphi^{-1}$ (for any $f \in \mathcal{f}$) and then compose $f \circ \varphi^{-1}$ with $\varphi$, we get
\[
f(q) = f(p) + \sum_{i=1}^{n} \frac{\partial(f \circ \varphi^{-1})}{\partial x_i}(\varphi(q))(x_i(q) - x_i(p)) + \sum_{i,j=1}^{n} (x_i(q) - x_i(p))(x_j(q) - x_j(p)) h,
\]
near $p$, where $h$ is $C^\infty$ and $x_i = pr_i \circ \varphi$. Since $f$ is a stationary germ, this yields
\[
f(q) = f(p) + \sum_{i,j=1}^{n} (x_i(q) - x_i(p))(x_j(q) - x_j(p)) h.
\]
If $v$ is any point-derivation and $f(p)$ is constant, Proposition 7.4 implies $v(f(p)) = 0$, and we get
\[
v(f) = v(f(p)) + \sum_{i,j=1}^{n} [(x_i(q) - x_i(p))(x_j(q) - x_j(p)) h](p)v(h)
+ (x_i(q) - x_i(p))(x_j(q) - x_j(p)) h(p) + (x_i(q) - x_i(p))(x_j(q) - x_j(p)) h(p) = 0.
\]
Since $x_i(q) - x_i(p))(p) = (x_j(q) - x_j(p))(p) = 0$, we conclude that $v$ vanishes on stationary germs. \(\square\)
Proposition 7.9 shows that in the case of a smooth manifold, in Definition 7.12, we can omit the requirement that point-derivations vanish on stationary germs, since this is automatic.

**Remark:** In the case of smooth manifolds \( k = \infty \) some authors, including Morita [133] (Chapter 1, Definition 1.32) and O’Neil [138] (Chapter 1, Definition 9), define derivations as linear derivations with domain \( C^\infty(M) \), the set of all smooth functions on the entire manifold, \( M \). This definition is simpler in the sense that it does not require the definition of the notion of germ but it is not local, because it is not obvious that if \( v \) is a point-derivation at \( p \), then \( v(f) = v(g) \) whenever \( f, g \in C^\infty(M) \) agree locally at \( p \). In fact, if two smooth locally defined functions agree near \( p \) it may not be possible to extend both of them to the whole of \( M \). However, it can be proved that this property is local because on smooth manifolds, “bump functions” exist (see Section 9.1, Proposition 9.2). Unfortunately, this argument breaks down for \( C^k \)-manifolds with \( k < \infty \) and in this case the ring of germs at \( p \) can’t be avoided.

### 7.4 Tangent and Cotangent Spaces Revisited ⊗

The space of linear forms on \( \mathcal{O}_{M,p}^{(k)} \) that vanish on \( \mathcal{S}_{M,p}^{(k)} \) turns out to be isomorphic to the dual of the quotient space \( \mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)} \), and this fact shows that the dual \( (T_pM)^* \) of the tangent space \( T_pM \), called the **cotangent space** to \( M \) at \( p \), can be viewed as the quotient space \( \mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)} \). This provides a fairly intrinsic definition of the cotangent space to \( M \) at \( p \). For notational simplicity, we write \( T_p^*M \) instead of \( (T_pM)^* \). This section is quite technical and can be safely skipped upon first (or second!) reading.

Let us refresh the reader’s memory and review quotient vector spaces. If \( E \) is a vector space, the set of all linear forms \( f: E \to \mathbb{R} \) on \( E \) is a vector space called the **dual** of \( E \) and denoted by \( E^* \). If \( H \subseteq E \) is any subspace of \( E \), we define the equivalence relation \( \sim \) so that for all \( u, v \in E \),

\[
u \sim v \quad \text{iff} \quad u - v \in H.
\]

Every equivalence class \( [u] \), is equal to the subset \( u + H = \{u + h \mid h \in H\} \), called a **coset**, and the set of equivalence classes \( E/H \) modulo \( \sim \) is a vector space under the operations

\[
[u] + [v] = [u + v]
\]

\[
\lambda[u] = [\lambda u].
\]

The space \( E/H \) is called the **quotient of \( E \) by \( H \)** or for short, a **quotient space**.

Denote by \( \mathcal{L}(E/H) \) the set of linear forms \( f: E \to \mathbb{R} \) that vanish on \( H \) (this means that for every \( f \in \mathcal{L}(E/H) \), we have \( f(h) = 0 \) for all \( h \in H \)). We claim that there is an isomorphism

\[
\mathcal{L}(E/H) \cong (E/H)^*
\]
between $\mathcal{L}(E/H)$ and the dual of the quotient space $E/H$.

To see this, define the map $f \mapsto \hat{f}$ from $\mathcal{L}(E/H)$ to $(E/H)^*$ as follows: For any $f \in \mathcal{L}(E/H)$,

$$\hat{f}([u]) = f(u), \quad [u] \in E/H.$$ 

This function is well-defined because it does not depend on the representative $u$, chosen in the equivalence class $[u]$. Indeed, if $v \sim u$, then $v = u + h$ some $h \in H$ and so

$$f(v) = f(u + h) = f(u) + f(h) = f(u),$$

since $f(h) = 0$ for all $h \in H$. The formula $\hat{f}([u]) = f(u)$ makes it obvious that $\hat{f}$ is linear since $f$ is linear. The mapping $f \mapsto \hat{f}$ is injective. This is because if $\hat{f}_1 = \hat{f}_2$, then

$$\hat{f}_1([u]) = \hat{f}_2([u])$$

for all $u \in E$, and because $\hat{f}_1([u]) = f_1(u)$ and $\hat{f}_2([u]) = f_2(u)$, we get $f_1(u) = f_2(u)$ for all $u \in E$, that is, $f_1 = f_2$. The mapping $f \mapsto \hat{f}$ is surjective because given any linear form $\varphi \in (E/H)^*$, if we define $f$ by

$$f(u) = \varphi([u])$$

for all $u \in E$, then $f$ is linear, vanishes on $H$ and clearly, $\hat{f} = \varphi$. Therefore, we have the isomorphism,

$$\mathcal{L}(E/H) \cong (E/H)^*,$$

as claimed.

As the subspace of linear forms on $O^{(k)}_{M,p}$ that vanish on $S^{(k)}_{M,p}$ is isomorphic to the dual $(O^{(k)}_{M,p}/S^{(k)}_{M,p})^*$ of the space $O^{(k)}_{M,p}/S^{(k)}_{M,p}$, we see that the linear forms

$$\left(\frac{\partial}{\partial x_1} \right)_p, \ldots, \left(\frac{\partial}{\partial x_n} \right)_p$$

also form a basis of $(O^{(k)}_{M,p}/S^{(k)}_{M,p})^*$.

There is a conceptually clearer way to define a canonical isomorphism between $T_p(M)$ and the dual of $O^{(k)}_{M,p}/S^{(k)}_{M,p}$ in terms of a nondegenerate pairing between $T_p(M)$ and $O^{(k)}_{M,p}/S^{(k)}_{M,p}$ (for the notion of a pairing, see Definition 21.1 and Proposition 21.1). This pairing is described by Serre in [159] (Chapter III, Section 8) for analytic manifolds and can be adapted to our situation.

Define the map $\omega: T_p(M) \times O^{(k)}_{M,p} \to \mathbb{R}$, so that

$$\omega([\gamma], f) = (f \circ \gamma)'(0),$$

for all $[\gamma] \in T_p(M)$ and all $f \in O^{(k)}_{M,p}$ (with $f \in \mathfrak{f}$). It is easy to check that the above expression does not depend on the representatives chosen in the equivalences classes $[\gamma]$, and
7.4. TANGENT AND COTANGENT SPACES REVISITED

Proposition 7.10. The map \( \omega: T_p(M) \times (\mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)}) \rightarrow \mathbb{R} \) defined so that
\[
\omega([\gamma], [f]) = (f \circ \gamma)'(0),
\]
for all \( [\gamma] \in T_p(M) \) and all \( [f] \in \mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)} \), is a nondegenerate pairing (with \( f \in f \)). Consequently, there is a canonical isomorphism between \( T_p(M) \) and \( (\mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)})^* \) and a canonical isomorphism between \( T_p^*(M) \) and \( \mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)} \).

Proof. This is basically a replay of the proof of Proposition 7.6. First, assume that given some \( [\gamma] \in T_p(M) \), we have \( \omega([\gamma], [f]) = 0 \) for all \( [f] \in \mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)} \). Pick a chart \((U, \varphi)\), with \( p \in U \) and let \( x_i = pr_i \circ \varphi \). Then, the \( x_i \)'s are not stationary germs, since \( x_i \circ \varphi^{-1} = pr_i \circ \varphi \circ \varphi^{-1} = pr_i \) and \( (pr_i)'(0) = pr_i \) (because \( pr_i \) is a linear form). By hypothesis, \( \omega([\gamma], [x_i]) = 0 \) for \( i = 1, \ldots, n \), which means that
\[
(x_i \circ \gamma)'(0) = (pr_i \circ \varphi \circ \gamma)'(0) = 0
\]
for \( i = 1, \ldots, n \), namely, \( pr_i((\varphi \circ \gamma)'(0)) = 0 \) for \( i = 1, \ldots, n \); that is,
\[
(\varphi \circ \gamma)'(0) = 0_n,
\]
proving that \( [\gamma] = 0 \).

Next, assume that given some \( [f] \in \mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)} \), we have \( \omega([\gamma], [f]) = 0 \) for all \( [\gamma] \in T_p(M) \). Again, pick a chart \((U, \varphi)\). For every \( z \in \mathbb{R}^n \), we have the curve \( \gamma_z \) given by
\[
\gamma_z(t) = \varphi^{-1}(\varphi(p) + tz)
\]
for all \( t \) in a small open interval containing 0. See Figure 7.8. Then, by hypothesis,
\[
\omega([\gamma_z], [f]) = (f \circ \gamma_z)'(0) = (f \circ \varphi^{-1})'(\varphi(p))(z) = 0
\]
for all \( z \in \mathbb{R}^n \), which means that
\[
(f \circ \varphi^{-1})'(\varphi(p)) = 0.
\]
But then, \( f \) is a stationary germ and so, \( [f] = 0 \). Therefore, we proved that \( \omega \) is a nondegenerate pairing. Since \( T_p(M) \) and \( \mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)} \) have finite dimension \( n \), it follows by Proposition 21.1 that there is are canonical isomorphisms between \( T_p(M) \) and \( (\mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)})^* \) and between \( T_p^*(M) \) and \( \mathcal{O}_{M,p}^{(k)}/S_{M,p}^{(k)} \). \( \square \)
In view of Proposition 7.10, we can identify $T_p(M)$ with $(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^*$ and $T_p^*(M)$ with $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$.

**Remark:** Also recall that if $E$ is a finite dimensional space, the map $\text{eval}_E: E \to E^{**}$ defined so that,

$$\text{eval}_E(v)(f) = f(v), \quad \text{for all } v \in E \text{ and for all } f \in E^*,$$

is a linear isomorphism.

Observe that we can view $\omega(u,f) = \omega(\gamma, [f])$ as the result of computing the directional derivative of the locally defined function $f \in \mathfrak{f}$ in the direction $u$ (given by a curve $\gamma$).

Proposition 7.10 suggests the following definition:

**Definition 7.13.** (Tangent and Cotangent Spaces, Version 3) Given any $C^k$-manifold $M$ of dimension $n$, with $k \geq 1$, for any $p \in M$, the tangent space at $p$ denoted $T_p(M)$ is the space of point-derivations on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$. Thus, $T_p(M)$ can be identified with $(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^*$, the dual of the quotient space $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$. The space $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ is called the cotangent space at $p$; it is isomorphic to the dual $T_p^*(M)$, of $T_p(M)$. (For simplicity of notation we also denote $T_p(M)$ by $T_p M$ and $T_p^*(M)$ by $T_p^* M$.)

We can consider any $C^k$-function $f$ on some open subset $U$ of $M$ as a representative of the germ $f \in \mathcal{O}_{M,p}^{(k)}$, so the image of $f$ in $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ under the canonical projection of $\mathcal{O}_{M,p}^{(k)}$ onto $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ makes sense. Observe that if $x_i = pr_i \circ \varphi$, as

$$\left(\frac{\partial}{\partial x_i}\right)_p x_j = \delta_{i,j},$$

the images of $x_1, \ldots, x_n$ in $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ form the dual basis of the basis $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ of $T_p(M)$.

Given any $C^k$-function $f$ on $U$, we denote the image of $f$ in $T_p^*(M) = \mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ by $df_p$. This is the differential of $f$ at $p$. Using the isomorphism between $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ and $(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^{**}$ described above, $df_p$ corresponds to the linear map in $T_p^*(M)$ defined by

$$df_p(v) = v(f),$$

for all $v \in T_p(M)$. With this notation, we see that $(dx_1)_p, \ldots, (dx_n)_p$ is a basis of $T_p^*(M)$, and this basis is dual to the basis $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ of $T_p(M)$. For simplicity of notation, we often omit the subscript $p$ unless confusion arises.

**Remark:** Strictly speaking, a tangent vector $v \in T_p(M)$ is defined on the space of germs $\mathcal{O}_{M,p}^{(k)}$, at $p$. However, it is often convenient to define $v$ on $C^k$-functions $f \in C^k(U)$, where $U$ is some open subset containing $p$. This is easy: Set

$$v(f) = v(f).$$
7.4. TANGENT AND COTANGENT SPACES REVISITED

Given any chart \((U, \varphi)\) at \(p\), since \(v\) can be written in a unique way as

\[
v = \sum_{i=1}^{n} \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p,
\]

we get

\[
v(f) = \sum_{i=1}^{n} \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p f.
\]

This shows that \(v(f)\) is the directional derivative of \(f\) in the direction \(v\). The directional derivative, \(v(f)\), is also denoted \(v[f]\).

It is also possible to define \(T_p(M)\) just in terms of \(\mathcal{O}^{(\infty)}_{M,p}\), and we get a fourth definition of \(T_p M\). Let \(\mathfrak{m}_{M,p} \subseteq \mathcal{O}^{(\infty)}_{M,p}\) be the ideal of germs that vanish at \(p\). Then, we also have the ideal \(\mathfrak{m}^2_{M,p}\), which consists of all finite sums of products of two elements in \(\mathfrak{m}_{M,p}\) and it turns out that \(T^*_p(M)\) is isomorphic to \(\mathfrak{m}_{M,p}/\mathfrak{m}^2_{M,p}\) (see Warner [175], Lemma 1.16).

Actually, if we let \(\mathfrak{m}^{(k)}_{M,p} \subseteq \mathcal{O}^{(k)}_{M,p}\) denote the ideal of \(C^k\)-germs that vanish at \(p\) and \(\mathfrak{s}^{(k)}_{M,p} \subseteq \mathcal{S}^{(k)}_{M,p}\) denote the ideal of stationary \(C^k\)-germs that vanish at \(p\), adapting Warner’s argument, we can prove the following proposition:

**Proposition 7.11.** We have the inclusion, \((\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}\) and the isomorphism

\[\left(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}\right)^* \cong \left(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}\right)^* .\]

As a consequence, \(T^*_p(M) \cong \left(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}\right)^*\) and \(T^*_p(M) \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}\).

**Proof.** Given any two germs, \(f, g \in \mathfrak{m}^{(k)}_{M,p}\), for any two locally defined functions, \(f \in \mathfrak{f}\) and \(g \in \mathfrak{g}\), since \(f(p) = g(p) = 0\), for any chart \((U, \varphi)\) with \(p \in U\), by definition of the product \(fg\) of two functions, for any \(q \in M\) near \(p\), we have

\[
(fg \circ \varphi^{-1})(q) = (fg)(\varphi^{-1}(q))
= f(\varphi^{-1}(q))g(\varphi^{-1}(q))
= (f \circ \varphi^{-1})(q)(g \circ \varphi^{-1})(q),
\]

so

\[
fg \circ \varphi^{-1} = (f \circ \varphi^{-1})(g \circ \varphi^{-1}),
\]

and by the product rule for derivatives, we get

\[
(fg \circ \varphi^{-1})'(0) = (f \circ \varphi^{-1})'(0)(g \circ \varphi^{-1})(0) + (f \circ \varphi^{-1})(0)(g \circ \varphi^{-1})'(0) = 0,
\]

because \((g \circ \varphi^{-1})(0) = g(\varphi^{-1}(0)) = g(p) = 0\) and \((f \circ \varphi^{-1})(0) = f(\varphi^{-1}(0)) = f(p) = 0\). Therefore, \(fg\) is stationary at \(p\) and since \(fg(p) = 0\), we have \(fg \in \mathfrak{s}^{(k)}_{M,p}\), which implies the inclusion \((\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}\).
Now, the key point is that any constant germ is stationary, since the derivative of a constant function is zero. Consequently, if \( v \) is a linear form on \( \mathcal{O}_{M,p}^{(k)} \) vanishing on \( \mathcal{S}_{M,p}^{(k)} \), then

\[
v(f) = v(f - f(p)),
\]

for all \( f \in \mathcal{O}_{M,p}^{(k)} \), where \( f(p) \) denotes the germ of constant functions with value \( f(p) \). We use this fact to define two functions between \((\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \) and \((\mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)})^* \) which are mutual inverses.

The map from \((\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \) to \((\mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)})^* \) is restriction to \( \mathfrak{m}_{M,p}^{(k)} \); every linear form \( v \) on \( \mathcal{O}_{M,p}^{(k)} \) vanishing on \( \mathcal{S}_{M,p}^{(k)} \) yields a linear form on \( \mathfrak{m}_{M,p}^{(k)} \) that vanishes on \( \mathfrak{s}_{M,p}^{(k)} \).

Conversely, for any linear form \( \ell \) on \( \mathfrak{m}_{M,p}^{(k)} \) vanishing on \( \mathfrak{s}_{M,p}^{(k)} \), define the function \( v_\ell \) so that

\[
v_\ell(f) = \ell(f - f(p)),
\]

for any germ \( f \in \mathcal{O}_{M,p}^{(k)} \). Since \( \ell \) is linear, it is clear that \( v_\ell \) is also linear. If \( f \) is stationary at \( p \), then \( f - f(p) \) is also stationary at \( p \) because the derivative of a constant is zero. Obviously, \( f - f(p) \) vanishes at \( p \). It follows that \( v_\ell \) vanishes on stationary germs at \( p \).

Using the fact that \( v(f) = v(f - f(p)) \), it is easy to check that the above maps between \((\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \) and \((\mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)})^* \) are mutual inverses, establishing the desired isomorphism. Because \((\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \) is finite-dimensional, we also have the isomorphism

\[
\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)} \cong \mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)}
\]

which yields the isomorphisms \( T^*_p(M) \cong \mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)} \) and \( T^*_p(M) \cong \mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)} \). \( \square \)

When \( k = \infty \), Proposition 7.8 shows that every stationary germ that vanishes at \( p \) belongs to \( \mathfrak{m}_{M,p}^{2} \). Therefore, when \( k = \infty \), we have \( \mathfrak{s}_{M,p}^{(\infty)} = \mathfrak{m}_{M,p}^{2} \) and so, we obtain the result quoted above (from Warner):

\[
T^*_p(M) = \mathcal{O}_{M,p}^{(\infty)}/\mathcal{S}_{M,p}^{(\infty)} \cong \mathfrak{m}_{M,p}^{2}/\mathfrak{m}_{M,p}^{2}.
\]

Remarks:

(1) The isomorphism

\[
(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \cong (\mathfrak{m}_{M,p}^{(k)}/\mathfrak{s}_{M,p}^{(k)})^*
\]

yields another proof that the linear forms in \((\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^* \) are point-derivations, using the argument from Warner [175] (Lemma 1.16). It is enough to prove that every linear
After having explored thoroughly the notion of tangent vector, we show how a tangent map can be a point-derivation on $\mathcal{O}_{M,p}$, and is defined functions at $M$.

7.5 Tangent Maps

(2) The ideal $m_{M,p}^{(k)}$ is in fact the unique maximal ideal of $\mathcal{O}_{M,p}^{(k)}$. This is because if $f \in \mathcal{O}_{M,p}^{(k)}$ does not vanish at $p$, then $1/f$ belongs to $\mathcal{O}_{M,p}^{(k)}$ (because if $f$ does not vanish at $p$, then by continuity, $f$ does not vanish in some open subset containing $p$, for all $f \in f$), and any proper ideal containing $m_{M,p}^{(k)}$ and $f$ would be equal to $\mathcal{O}_{M,p}^{(k)}$, which is absurd. Thus, $\mathcal{O}_{M,p}^{(k)}$ is a local ring (in the sense of commutative algebra) called the local ring of germs of $C^k$-functions at $p$. These rings play a crucial role in algebraic geometry.

(3) Using the map $f \mapsto f - f(p)$, it is easy to see that

$$\mathcal{O}_{M,p}^{(k)} \cong \mathbb{R} \oplus m_{M,p}^{(k)}$$

and

$$\mathcal{S}_{M,p}^{(k)} \cong \mathbb{R} \oplus s_{M,p}^{(k)}.$$

7.5 Tangent Maps

After having explored thoroughly the notion of tangent vector, we show how a $C^k$-map $h : M \to N$, between $C^k$ manifolds, induces a linear map $dh_p : T_p(M) \to T_{h(p)}(N)$, for every $p \in M$. We find it convenient to use Version 3 of the definition of a tangent vector. So, let $u \in T_p(M)$ be a point-derivation on $\mathcal{O}_{M,p}^{(k)}$ that vanishes on $\mathcal{S}_{M,p}^{(k)}$. We would like $dh_p(u)$ to be a point-derivation on $\mathcal{O}_{h(M),h(p)}^{(k)}$ that vanishes on $\mathcal{S}_{h(M),h(p)}^{(k)}$. Now, for every germ $g \in \mathcal{O}_{h(M),h(p)}^{(k)}$, if $g \in \mathcal{g}$ is any locally defined function at $h(p)$, it is clear that $g \circ h$ is locally defined at $p$ and is $C^k$, and that if $g_1, g_2 \in \mathcal{g}$ then $g_1 \circ h$ and $g_2 \circ h$ are equivalent. The germ of all locally defined functions at $p$ of the form $g \circ h$, with $g \in \mathcal{g}$, will be denoted $g \circ h$. Then, we set

$$dh_p(u)(g) = u(g \circ h).$$

In any chart $(U, \varphi)$ at $p$, if $u = \sum_{i=1}^n \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p$, then

$$dh_p(u)(g) = \sum_{i=1}^n \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p g \circ h$$
for any \( g \in \mathfrak{g} \). Moreover, if \( g \) is a stationary germ at \( h(p) \), then for some chart \((V, \psi)\) on \( N \) at \( q = h(p) \), we have \((g \circ \psi^{-1})'(\psi(q)) = 0\) and, for any chart \((U, \varphi)\) at \( p \) on \( M \), we use the chain rule to obtain

\[
(g \circ h \circ \varphi^{-1})'(\varphi(p)) = (g \circ \psi^{-1})'(\psi(q))((\psi \circ h \circ \varphi^{-1})'(\varphi(p))) = 0,
\]

which means that \( g \circ h \) is stationary at \( p \). Therefore, \( dh_p(u) \in T_{h(p)}(M) \). It is also clear that \( dh_p \) is a linear map. We summarize all this in the following definition:

**Definition 7.14.** (Using Version 3 of a tangent vector) Given any two \( C^k \)-manifolds \( M \) and \( N \), of dimension \( m \) and \( n \) respectively, for any \( C^k \)-map \( h: M \to N \) and for every \( p \in M \), the differential of \( h \) at \( p \) or tangent map \( dh_p: T_p(M) \to T_{h(p)}(N) \) (also denoted \( T_p h: T_p(M) \to T_{h(p)}(N) \)), is the linear map defined so that

\[
dh_p(u)(g) = T_p h(u)(g) = u(g \circ h)
\]

for every \( u \in T_p(M) \) and every germ \( g \in O_{N,h(p)}^{(k)} \). The linear map \( dh_p \) \((= T_p h)\) is sometimes denoted \( h_p' \) or \( D_p h \).

![Figure 7.9: The tangent map \( dh_p(u)(g) = \sum_{i=1}^{2} \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p g \circ h \)](image)

The chain rule is easily generalized to manifolds.

**Proposition 7.12.** Given any two \( C^k \)-maps \( f: M \to N \) and \( g: N \to P \) between smooth \( C^k \)-manifolds, for any \( p \in M \), we have

\[
d(g \circ f)_p = dg_{f(p)} \circ df_p.
\]
7.5. **TANGENT MAPS**

In the special case where \( N = \mathbb{R} \), a \( C^k \)-map between the manifolds \( M \) and \( \mathbb{R} \) is just a \( C^k \)-function on \( M \). It is interesting to see what \( T_p f \) is explicitly. Since \( N = \mathbb{R} \), germs (of functions on \( \mathbb{R} \)) at \( t_0 = f(p) \) are just germs of \( C^k \)-functions \( g : \mathbb{R} \to \mathbb{R} \) locally defined at \( t_0 \). Then, for any \( u \in T_p(M) \) and every germ \( g \) at \( t_0 \),

\[
T_p f(u)(g) = u(g \circ f).
\]

If we pick a chart \((U, \varphi)\) on \( M \) at \( p \), we know that the \((\frac{\partial}{\partial x_i})_p \) form a basis of \( T_p(M) \), with \( 1 \leq i \leq n \). Therefore, it is enough to figure out what \( T_p f(u)(g) \) is when \( u = \left(\frac{\partial}{\partial x_i}\right)_p \). In this case,

\[
T_p f \left( \left( \frac{\partial}{\partial x_i} \right)_p \right)(g) = \left. \frac{\partial(g \circ f \circ \varphi^{-1})}{\partial X_i} \right|_{\varphi(p)}.
\]

Using the chain rule, we find that

\[
T_p f \left( \left( \frac{\partial}{\partial x_i} \right)_p \right)(g) = \left. \left( \frac{\partial}{\partial x_i} \right)_p f \frac{dg}{dt} \right|_{t_0}.
\]

Therefore, we have

\[
T_p f(u) = u(f) \left. \frac{d}{dt} \right|_{t_0}.
\]

This shows that we can identify \( T_p f \) with the linear form in \( T^*_p(M) \) defined by

\[
df_p(u) = u(f), \quad u \in T_p M,
\]

by identifying \( T_{t_0} \mathbb{R} \) with \( \mathbb{R} \). This is consistent with our previous definition of \( df_p \) as the image of \( f \) in \( T^*_p(M) = \mathcal{O}^{(k)}_{M,p} / \mathcal{S}^{(k)}_{M,p} \) (as \( T_p(M) \) is isomorphic to \((\mathcal{O}^{(k)}_{M,p} / \mathcal{S}^{(k)}_{M,p})^*\)).

Again, even though this is just a restatement of facts we already showed, we state the following proposition because of its practical usefulness:

**Proposition 7.13.** Given any \( C^k \)-manifold \( M \) of dimension \( n \), with \( k \geq 1 \), for any \( p \in M \) and any chart \((U, \varphi)\) at \( p \), the \( n \) linear maps

\[
(dx_1)_p, \ldots, (dx_n)_p
\]

form a basis of \( T^*_p M \), where \((dx_i)_p\), the differential of \( x_i \) at \( p \), is identified with the linear form in \( T^*_p M \) such that \((dx_i)_p(v) = v(x_1)\), for every \( v \in T_p M \) (by identifying \( T_{t_0} \mathbb{R} \) with \( \mathbb{R} \)).

In preparation for the definition of the flow of a vector field (which will be needed to define the exponential map in Lie group theory), we need to define the tangent vector to a curve on a manifold. Given a \( C^k \)-curve \( \gamma : (a, b) \to M \) on a \( C^k \)-manifold \( M \), for any \( t_0 \in (a, b) \), we would like to define the tangent vector to the curve \( \gamma \) at \( t_0 \) as a tangent vector
to $M$ at $p = \gamma(t_0)$. We do this as follows: Recall that $\frac{d}{dt}\big|_{t_0}$ is a basis vector of $T_{t_0}(\mathbb{R}) = \mathbb{R}$. So, define the tangent vector to the curve $\gamma$ at $t_0$, denoted $\dot{\gamma}(t_0)$ (or $\gamma'(t_0)$, or $\frac{d}{dt}(t_0)$), by

$$
\dot{\gamma}(t_0) = d\gamma_{t_0}\left(\frac{d}{dt}\big|_{t_0}\right).
$$

We find it necessary to define curves (in a manifold) whose domain is not an open interval. A map $\gamma: [a, b] \to M$, is a $C^k$-curve in $M$ if it is the restriction of some $C^k$-curve $\tilde{\gamma}: (a - \epsilon, b + \epsilon) \to M$, for some (small) $\epsilon > 0$. Note that for such a curve (if $k \geq 1$) the tangent vector $\dot{\gamma}(t)$ is defined for all $t \in [a, b]$. A continuous curve $\gamma: [a, b] \to M$ is piecewise $C^k$ iff there a sequence $a_0 = a, a_1, \ldots, a_m = b$, so that the restriction $\gamma_i$ of $\gamma$ to each $[a_i, a_{i+1}]$ is a $C^k$-curve, for $i = 0, \ldots, m - 1$. This implies that $\gamma'_i(a_{i+1})$ and $\gamma'_{i+1}(a_{i+1})$ are defined for $i = 0, \ldots, m - 1$, but there may be a jump in the tangent vector to $\gamma$ at $a_{i+1}$, that is, we may have $\gamma'_i(a_{i+1}) \neq \gamma'_{i+1}(a_{i+1})$.

Sometimes, especially in the case of a linear Lie group, it is more convenient to define the tangent map in terms of Version 1 of a tangent vector. Given any two $C^k$-manifolds $M$ and $N$, of dimension $m$ and $n$ respectively, for any $C^k$-map $h: M \to N$ and for every $p \in M$, the differential of $h$ at $p$ or tangent map $dh_p: T_p(M) \to T_{h(p)}(N)$ (also denoted $T_p h: T_p(M) \to T_{h(p)}(N)$), is the linear map defined such that for every equivalence class $u = [\gamma]$ of curves $\gamma$ in $M$ with $\gamma(0) = p$,

$$
dh_p(u) = T_p h(u) = v,
$$

where $v$ is the equivalence class of all curves through $h(p)$ in $N$ of the form $h \circ \gamma$, with $\gamma \in u$. 

**Definition 7.15.** (Using Version 1 of a tangent vector) Given any two $C^k$-manifolds $M$ and $N$, of dimension $m$ and $n$ respectively, for any $C^k$-map $h: M \to N$ and for every $p \in M$, the differential of $h$ at $p$ or tangent map $dh_p: T_p(M) \to T_{h(p)}(N)$ (also denoted $T_p h: T_p(M) \to T_{h(p)}(N)$), is the linear map defined such that for every equivalence class $u = [\gamma]$ of curves $\gamma$ in $M$ with $\gamma(0) = p$,
Figure 7.10: The tangent map $dh_p(u) = v$ defined via equivalent curves

If $M$ is a manifold in $\mathbb{R}^{N_1}$ and $N$ is a manifold in $\mathbb{R}^{N_2}$ (for some $N_1, N_2 \geq 1$), then $\gamma'(0) \in \mathbb{R}^{N_1}$ and $(h \circ \gamma)'(0) \in \mathbb{R}^{N_2}$, so in this case the definition of $dh_p = Th_p$ is just Definition 4.9; namely, for any curve $\gamma$ in $M$ such that $\gamma(0) = p$ and $\gamma'(0) = u$,

$$dh_p(u) = Th_p(u) = (h \circ \gamma)'(0).$$

For example, consider the linear Lie group $\text{SO}(3)$, pick any vector $u \in \mathbb{R}^3$, and let $f: \text{SO}(3) \to \mathbb{R}^3$ be given by

$$f(R) = Ru, \quad R \in \text{SO}(3).$$

To compute $df_R: T_R\text{SO}(3) \to T_{Ru}\mathbb{R}^3$, since $T_R\text{SO}(3) = R\mathfrak{so}(3)$ and $T_{Ru}\mathbb{R}^3 = \mathbb{R}^3$, pick any tangent vector $RB \in R\mathfrak{so}(3) = T_R\text{SO}(3)$ (where $B$ is any $3 \times 3$ skew symmetric matrix), let $\gamma(t) = Re^{tB}$ be the curve through $R$ such that $\gamma'(0) = RB$, and compute

$$df_R(RB) = (f(\gamma(t))')'(0) = (Re^{tB}u)'(0) = RBu.$$

Therefore, we see that

$$df_R(X) = Xu, \quad X \in T_R\text{SO}(3) = R\mathfrak{so}(3).$$

If we express the skew symmetric matrix $B \in \mathfrak{so}(3)$ as $B = \omega \times$ for some vector $\omega \in \mathbb{R}^3$, then we have

$$df_R(R\omega \times) = R\omega \times u = R(\omega \times u).$$

Using the isomorphism of the Lie algebras $(\mathbb{R}^3, \times)$ and $\mathfrak{so}(3)$, the tangent map $df_R$ is given by

$$df_R(R\omega) = R(\omega \times u).$$
Here is another example inspired by an optimization problem investigated by Taylor and Kriegman. Pick any two vectors \( u, v \in \mathbb{R}^3 \), and let \( f : \text{SO}(3) \to \mathbb{R} \) be the function given by

\[
f(R) = (u^\top R v)^2.
\]

To compute \( df_R : T_R \text{SO}(3) \to T_{f(R)} \mathbb{R} \), since \( T_R \text{SO}(3) = R \text{so}(3) \) and \( T_{f(R)} \mathbb{R} = \mathbb{R} \), again pick any tangent vector \( RB \in R \text{so}(3) = T_R \text{SO}(3) \) (where \( B \) is any \( 3 \times 3 \) skew symmetric matrix), let \( \gamma(t) = R e^{tB} \) be the curve through \( R \) such that \( \gamma'(0) = RB \), and compute via the product rule

\[
df_R(RB) = (f(\gamma(t)))'(0) = ((u^\top R e^{tB} v)^2)'(0) = u^\top RB vu^\top R v + u^\top R vu^\top RB v = 2u^\top RB vu^\top R v,
\]

where the last equality used the observation that \( u^\top RBv \) and \( u^\top R v \) are real numbers. Therefore,

\[
df_R(X) = 2u^\top X vu^\top R v, \quad X \in R \text{so}(3).
\]

Unlike the case of functions defined on vector spaces, in order to define the gradient of \( f \), a function defined on \( \text{SO}(3) \), a “nonflat” manifold, we need to pick a Riemannian metric on \( \text{SO}(3) \). We will explain how to do this in Chapter 10.

### 7.6 Submanifolds, Immersions, Embeddings

Although the notion of submanifold is intuitively rather clear, technically, it is a bit tricky. In fact, the reader may have noticed that many different definitions appear in books and that it is not obvious at first glance that these definitions are equivalent. What is important is that a submanifold \( N \) of a given manifold \( M \) has the topology induced by \( M \) but also that the charts of \( N \) are somehow induced by those of \( M \).

Given \( m, n \), with \( 0 \leq m \leq n \), we can view \( \mathbb{R}^m \) as a subspace of \( \mathbb{R}^n \) using the inclusion

\[
\mathbb{R}^m \cong \mathbb{R}^m \times \{(0, \ldots, 0)\} \hookrightarrow \mathbb{R}^m \times \mathbb{R}^{n-m} = \mathbb{R}^n, \quad (x_1, \ldots, x_m) \mapsto (x_1, \ldots, x_m, 0, \ldots, 0).
\]

**Definition 7.16.** Given a \( C^k \)-manifold \( N \) of dimension \( n \), a subset \( M \) of \( N \) is an \( m \)-dimensional submanifold of \( N \) (where \( 0 \leq m \leq n \)) iff for every point \( p \in M \), there is a chart \((U, \varphi)\) of \( N \) (in the maximal atlas for \( N \)), with \( p \in U \), so that

\[
\varphi(U \cap M) = \varphi(U) \cap (\mathbb{R}^m \times \{0_{n-m}\}).
\]

(We write \( 0_{n-m} = (0, \ldots, 0) \).)
The subset $U \cap M$ of Definition 7.16 is sometimes called a *slice of $(U, \varphi)$* and we say that $(U, \varphi)$ is *adapted to $M$* (See O’Neill [138] or Warner [175]).

Other authors, including Warner [175], use the term submanifold in a broader sense than us and they use the word *embedded submanifold* for what is defined in Definition 7.16.

The following proposition has an almost trivial proof but it justifies the use of the word submanifold:

**Proposition 7.14.** Given a $C^k$-manifold $N$ of dimension $n$, for any submanifold $M$ of $N$ of dimension $m \leq n$, the family of pairs $(U \cap M, \varphi \upharpoonright U \cap M)$, where $(U, \varphi)$ ranges over the charts over any atlas for $N$, is an atlas for $M$, where $M$ is given the subspace topology. Therefore, $M$ inherits the structure of a $C^k$-manifold.

In fact, every chart on $M$ arises from a chart on $N$ in the following precise sense:

**Proposition 7.15.** Given a $C^k$-manifold $N$ of dimension $n$ and a submanifold $M$ of $N$ of dimension $m \leq n$, for any $p \in M$ and any chart $(W, \eta)$ of $M$ at $p$, there is some chart $(U, \varphi)$ of $N$ at $p$, so that

$$
\varphi(U \cap M) = \varphi(U) \cap (\mathbb{R}^m \times \{0_{n-m}\}) \quad \text{and} \quad \varphi \upharpoonright U \cap M = \eta \upharpoonright U \cap M,
$$

where $p \in U \cap M \subseteq W$.

**Proof.** See Berger and Gostiaux [20] (Chapter 2).
It is also useful to define more general kinds of “submanifolds.”

**Definition 7.17.** Let $h: M \to N$ be a $C^k$-map of manifolds.

(a) The map $h$ is an **immersion** of $M$ into $N$ iff $dh_p$ is injective for all $p \in M$.

(b) The set $h(M)$ is an **immersed submanifold** of $N$ iff $h$ is an injective immersion.

(c) The map $h$ is an **embedding** of $M$ into $N$ iff $h$ is an injective immersion such that the induced map, $M \to h(M)$, is a homeomorphism, where $h(M)$ is given the subspace topology (equivalently, $h$ is an open map from $M$ into $h(M)$ with the subspace topology). We say that $h(M)$ (with the subspace topology) is an **embedded submanifold** of $N$.

(d) The map $h$ is a **submersion** of $M$ into $N$ iff $dh_p$ is surjective for all $p \in M$.

Again, we warn our readers that certain authors (such as Warner [175]) call $h(M)$, in (b), a submanifold of $N$! We prefer the terminology **immersed submanifold**.

The notion of immersed submanifold arises naturally in the framework of Lie groups. Indeed, the fundamental correspondence between Lie groups and Lie algebras involves Lie subgroups that are not necessarily closed. But, as we will see later, subgroups of Lie groups that are also submanifolds are always closed. It is thus necessary to have a more inclusive notion of submanifold for Lie groups and the concept of immersed submanifold is just what’s needed.

Immersions of $\mathbb{R}$ into $\mathbb{R}^3$ are parametric curves and immersions of $\mathbb{R}^2$ into $\mathbb{R}^3$ are parametric surfaces. These have been extensively studied, for example, see DoCarmo [59], Berger and Gostiaux [20] or Gallier [72].

Immersions (i.e., subsets of the form $h(M)$, where $M$ is an immersion) are generally neither injective immersions (i.e., subsets of the form $h(M)$, where $M$ is an injective immersion) nor embeddings (or submanifolds). For example, immersions can have self-intersections, as the plane curve (nodal cubic) shown in Figure 7.12 and given by: $x = t^2 - 1; y = t(t^2 - 1)$.

Note that the cuspidal cubic, $t \mapsto (t^2, t^3)$, is an injective map, but it is not an immersion since its derivative at the origin is zero.

Injective immersions are generally not embeddings (or submanifolds) because $h(M)$ may not be homeomorphic to $M$. An example is given by the Lemniscate of Bernoulli shown in Figure 7.13, an injective immersion of $\mathbb{R}$ into $\mathbb{R}^2$:

\[
\begin{align*}
x &= \frac{t(1 + t^2)}{1 + t^4}, \\
y &= \frac{t(1 - t^2)}{1 + t^4}.
\end{align*}
\]

When $t = 0$, the curve passes through the origin. When $t \to -\infty$, the curve tends to the origin from the left and from above, and when $t \to +\infty$, the curve tends to the origin.
from the right and from below. Therefore, the inverse of the map defining the Lemniscate of Bernoulli is not continuous at the origin.

Another interesting example is the immersion of $\mathbb{R}$ into the 2-torus, $T^2 = S^1 \times S^1 \subseteq \mathbb{R}^4$, given by

$$t \mapsto (\cos t, \sin t, \cos ct, \sin ct),$$

where $c \in \mathbb{R}$. One can show that the image of $\mathbb{R}$ under this immersion is closed in $T^2$ iff $c$ is rational. Moreover, the image of this immersion is dense in $T^2$ but not closed iff $c$ is irrational. The above example can be adapted to the torus in $\mathbb{R}^3$: One can show that the immersion given by

$$t \mapsto ((2 + \cos t) \cos(\sqrt{2}t), (2 + \cos t) \sin(\sqrt{2}t), \sin t),$$

is dense but not closed in the torus (in $\mathbb{R}^3$) given by

$$(s, t) \mapsto ((2 + \cos s) \cos t, (2 + \cos s) \sin t, \sin s),$$

where $s, t \in \mathbb{R}$.

There is, however, a close relationship between submanifolds and embeddings.

**Proposition 7.16.** If $M$ is a submanifold of $N$, then the inclusion map $j: M \to N$ is an embedding. Conversely, if $h: M \to N$ is an embedding, then $h(M)$ with the subspace topology is a submanifold of $N$ and $h$ is a diffeomorphism between $M$ and $h(M)$. 
Proof. See O’Neill [138] (Chapter 1) or Berger and Gostiaux [20] (Chapter 2).

In summary, embedded submanifolds and (our) submanifolds coincide. Some authors refer to spaces of the form \( h(M) \), where \( h \) is an injective immersion, as \textit{immersed submanifolds} and we have adopted this terminology. However, in general, an immersed submanifold is \textit{not} a submanifold. One case where this holds is when \( M \) is compact, since then, a bijective continuous map is a homeomorphism. For yet a notion of submanifold intermediate between immersed submanifolds and (our) submanifolds, see Sharpe [162] (Chapter 1).
Chapter 8

Vector Fields, Lie Derivatives, Integral Curves, Flows

Our goal in this chapter is to generalize the concept of a vector field to manifolds and to promote some standard results about ordinary differential equations to manifolds.

8.1 Tangent and Cotangent Bundles

Let $M$ be a $C^k$-manifold (with $k \geq 2$). Roughly speaking, a vector field on $M$ is the assignment $p \mapsto X(p)$, of a tangent vector $X(p) \in T_p(M)$, to a point $p \in M$. Generally, we would like such assignments to have some smoothness properties when $p$ varies in $M$, for example, to be $C^l$, for some $l$ related to $k$. Now, if the collection $T(M)$ of all tangent spaces $T_p(M)$ was a $C^l$-manifold, then it would be very easy to define what we mean by a $C^l$-vector field: We would simply require the map $X: M \to T(M)$ to be $C^l$.

If $M$ is a $C^k$-manifold of dimension $n$, then we can indeed make $T(M)$ into a $C^{k-1}$-manifold of dimension $2n$ and we now sketch this construction.

We find it most convenient to use Version 2 of the definition of tangent vectors, i.e., as equivalence classes of triples $(U, \varphi, x)$, where $(U, \varphi)$ is a chart at $p$ and $x \in \mathbb{R}^n$. Recall that $(U, \varphi, x)$ and $(V, \psi, y)$ are equivalent iff

$$(\psi \circ \varphi^{-1})'(\varphi(p))(x) = y.$$ 

First, we let $T(M)$ be the disjoint union of the tangent spaces $T_p(M)$, for all $p \in M$. Formally,

$$T(M) = \{(p, v) \mid p \in M, v \in T_p(M)\}.$$ 

See Figure 8.1.
There is a **natural projection**

\[ \pi: T(M) \to M, \quad \text{with} \quad \pi(p, v) = p. \]

We still have to give \( T(M) \) a topology and to define a \( C^{k-1} \)-atlas. For every chart \((U, \varphi)\) of \( M \) (with \( U \) open in \( M \)), we define the function \( \tilde{\varphi}: \pi^{-1}(U) \to \mathbb{R}^{2n} \), by

\[ \tilde{\varphi}(p, v) = (\varphi(p), \theta^{-1}_{U,\varphi,p}(v)), \]

where \((p, v) \in \pi^{-1}(U)\) and \( \theta_{U,\varphi,p} \) is the isomorphism between \( \mathbb{R}^{n} \) and \( T_p(M) \) described just after Definition 7.8. It is obvious that \( \tilde{\varphi} \) is a bijection between \( \pi^{-1}(U) \) and \( \varphi(U) \times \mathbb{R}^{n} \), an open subset of \( \mathbb{R}^{2n} \). See Figure 8.2.

---

**Figure 8.1:** The tangent bundle of \( S^1 \)

**Figure 8.2:** A chart for \( T(S^1) \)
We give \( T(M) \) the weakest topology that makes all the \( \tilde{\varphi} \) continuous, i.e., we take the collection of subsets of the form \( \tilde{\varphi}^{-1}(W) \), where \( W \) is any open subset of \( \varphi(U) \times \mathbb{R}^n \), as a basis of the topology of \( T(M) \). One may check that \( T(M) \) is Hausdorff and second-countable in this topology. If \((U, \varphi)\) and \((V, \psi)\) are two overlapping charts of \( M \), then the definition of the equivalence relation on triples \((U, \varphi, x)\) and \((V, \psi, y)\) immediately implies that

\[
\theta_{(V, \psi, p)}^{-1} \circ \theta_{(U, \varphi, p)} = (\psi \circ \varphi^{-1})'_z
\]

for all \( p \in U \cap V \), with \( z = \varphi(p) \), so the transition map,

\[
\tilde{\psi} \circ \tilde{\varphi}^{-1} : \varphi(U \cap V) \times \mathbb{R}^n \longrightarrow \psi(U \cap V) \times \mathbb{R}^n
\]

is given by

\[
\tilde{\psi} \circ \tilde{\varphi}^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), (\psi \circ \varphi^{-1})'_z(x)), \quad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.
\]

It is clear that \( \tilde{\psi} \circ \tilde{\varphi}^{-1} \) is a \( C^k \)-map. Therefore, \( T(M) \) is indeed a \( C^k \)-manifold of dimension \( 2n \), called the tangent bundle.

**Remark:** Even if the manifold \( M \) is naturally embedded in \( \mathbb{R}^N \) (for some \( N \geq n = \text{dim}(M) \)), it is not at all obvious how to view the tangent bundle \( T(M) \) as embedded in \( \mathbb{R}^{N'} \), for some suitable \( N' \). Hence, we see that the definition of an abstract manifold is unavoidable.

A similar construction can be carried out for the cotangent bundle. In this case, we let \( T^*(M) \) be the disjoint union of the cotangent spaces \( T^*_p(M) \), that is,

\[
T^*(M) = \{ (p, \omega) \mid p \in M, \omega \in T^*_p(M) \}.
\]

We also have a natural projection \( \pi : T^*(M) \to M \) with \( \pi(p, \omega) = p \), and we can define charts in several ways. One method used by Warner [175] goes as follows: For any chart, \((U, \varphi)\), on \( M \), we define the function, \( \tilde{\varphi} : \pi^{-1}(U) \to \mathbb{R}^{2n} \), by

\[
\tilde{\varphi}(p, \omega) = \left( \varphi(p), \omega \left( \frac{\partial}{\partial x_1} \right)_p, \ldots, \omega \left( \frac{\partial}{\partial x_n} \right)_p \right),
\]

where \( (p, \omega) \in \pi^{-1}(U) \) and the \( \left( \frac{\partial}{\partial x_i} \right)_p \) are the basis of \( T_p(M) \) associated with the chart \((U, \varphi)\). Again, one can make \( T^*(M) \) into a \( C^{k-1} \)-manifold of dimension \( 2n \), called the cotangent bundle. We leave the details as an exercise to the reader (Or, look at Berger and Gostiaux [20]). Another method using Version 3 of the definition of tangent vectors is presented in Section 28.2. For each chart \((U, \varphi)\) on \( M \), we obtain a chart

\[
\tilde{\varphi}^* : \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n \subseteq \mathbb{R}^{2n}
\]
on $T^*(M)$ given by
\[ \tilde{\varphi}^*(p, \omega) = (\varphi(p), \theta_{U,\varphi,p}^*(\omega)) \]
for all $(p, \omega) \in \pi^{-1}(U)$, where
\[ \theta_{U,\varphi,p}^* = \iota \circ \theta_{U,\varphi,p}^T : T^*_p(M) \to \mathbb{R}^n. \]

Here, $\theta_{U,\varphi,p}^T : T^*_p(M) \to (\mathbb{R}^n)^*$ is obtained by dualizing the map, $\theta_{U,\varphi,p} : \mathbb{R}^n \to T_p(M)$ and $\iota : (\mathbb{R}^n)^* \to \mathbb{R}^n$ is the isomorphism induced by the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$ and its dual basis.

For simplicity of notation, we also use the notation $TM$ for $T(M)$ (resp. $T^*M$ for $T^*(M)$).

Observe that for every chart $(U, \varphi)$ on $M$, there is a bijection
\[ \tau_U : \pi^{-1}(U) \to U \times \mathbb{R}^n, \]
given by
\[ \tau_U(p, v) = (p, \theta_{U,\varphi,p}^{-1}(v)). \]

Clearly, $pr_1 \circ \tau_U = \pi$ on $\pi^{-1}(U)$, as illustrated by the following commutative diagram:

\[ \pi^{-1}(U) \xrightarrow{\tau_U} U \times \mathbb{R}^n \]

Thus locally, that is over $U$, the bundle $T(M)$ looks like the product manifold $U \times \mathbb{R}^n$. We say that $T(M)$ is **locally trivial** (over $U$) and we call $\tau_U$ a **trivializing map**. For any $p \in M$, the vector space $\pi^{-1}(p) = \{p\} \times T_p(M) \cong T_p(M)$ is called the **fibre above** $p$. Observe that the restriction of $\tau_U$ to $\pi^{-1}(p)$ is a linear isomorphism between $\{p\} \times T_p(M) \cong T_p(M)$ and $\{p\} \times \mathbb{R}^n \cong \mathbb{R}^n$, for any $p \in M$. Furthermore, for any two overlapping charts $(U, \varphi)$ and $(V, \psi)$, there is a function $g_{UV} : U \cap V \to \text{GL}(n, \mathbb{R})$ such that
\[ (\tau_U \circ \tau_V^{-1})(p, x) = (p, g_{UV}(p)(x)) \]
for all $p \in U \cap V$ and all $x \in \mathbb{R}^n$, with $g_{UV}(p)$ given by
\[ g_{UV}(p) = (\varphi \circ \psi^{-1})'(p). \]

Obviously, $g_{UV}(p)$ is a linear isomorphism of $\mathbb{R}^n$ for all $p \in U \cap V$. The maps $g_{UV}(p)$ are called the **transition functions** of the tangent bundle.

For example, if $M = S^n$, the $n$-sphere in $\mathbb{R}^{n+1}$, we have two charts given by the stereographic projection $(U_N, \sigma_N)$ from the north pole, and the stereographic projection $(U_S, \sigma_S)$ from the south pole (with $U_N = S^n - \{N\}$ and $U_S = S^n - \{S\}$), and on the overlap, $U_N \cap U_S = S^n - \{N, S\}$, the transition maps
\[ I = \sigma_S \circ \sigma_N^{-1} = \sigma_N \circ \sigma_S^{-1} \]
defined on \( \varphi_N(U_N \cap U_S) = \varphi_S(U_N \cap U_S) = \mathbb{R}^n - \{0\} \), are given by

\[
(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^n x_i^2} (x_1, \ldots, x_n);
\]

that is, the inversion \( I \) of center \( O = (0, \ldots, 0) \) and power 1. We leave it as an exercise to prove that for every point \( u \in \mathbb{R}^n - \{0\} \), we have

\[
dI_u(h) = \|u\|^{-2} \left( h - \frac{2 \langle u, h \rangle}{\|u\|^2} u \right),
\]

the composition of the hyperplane reflection about the hyperplane \( u^\perp \subseteq \mathbb{R}^n \) with the magnification of center \( O \) and ratio \( \|u\|^{-2} \). This is a similarity transformation. Therefore, the transition function \( g_{NS} \) (defined on \( U_N \cap U_S \)) of the tangent bundle \( TS^n \) is given by

\[
g_{NS}(p)(h) = \|\sigma_S(p)\|^{-2} \left( h - \frac{2 \langle \sigma_S(p), h \rangle}{\|\sigma_S(p)\|^2} \sigma_S(p) \right).
\]

All these ingredients are part of being a vector bundle. For more on bundles, see Chapter 28, in particular, Section 28.2 on vector bundles where the construction of the bundles \( TM \) and \( T^*M \) is worked out in detail. See also the references in Chapter 28.

When \( M = \mathbb{R}^n \), observe that \( T(M) = M \times \mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n \), i.e., the bundle \( T(M) \) is (globally) trivial.

Given a \( C^k \)-map \( h: M \rightarrow N \) between two \( C^k \)-manifolds, we can define the function \( dh: T(M) \rightarrow T(N) \) (also denoted \( Th \), or \( h_* \), or \( Dh \)), by setting

\[
dh(u) = dh_p(u), \quad \text{iff} \quad u \in T_p(M).
\]

We leave the next proposition as an exercise to the reader (A proof can be found in Berger and Gostiaux [20]).

**Proposition 8.1.** Given a \( C^k \)-map \( h: M \rightarrow N \) between two \( C^k \)-manifolds \( M \) and \( N \) (with \( k \geq 1 \)), the map \( dh: T(M) \rightarrow T(N) \) is a \( C^{k-1} \) map.

We are now ready to define vector fields.

### 8.2 Vector Fields, Lie Derivative

In Section 2.3 we introduced the notion of a vector field in \( \mathbb{R}^n \). We now generalize the notion of a vector field to a manifold. Let \( M \) be a \( C^{k+1} \) manifold. A \( C^k \)-vector field on \( M \) is an assignment \( p \mapsto X(p) \) of a tangent vector \( X(p) \in T_p(M) \) to a point \( p \in M \), so that \( X(p) \) varies in a \( C^k \)-fashion in terms of \( p \). This notion is captured rigorously by the following definition:
Definition 8.1. Let $M$ be a $C^{k+1}$ manifold, with $k \geq 1$. For any open subset $U$ of $M$, a \textit{vector field on $U$} is any \textit{section $X$ of $T(M)$ over $U$}, that is, any function $X : U \to T(M)$, such that $\pi \circ X = \text{id}_U$ (i.e., $X(p) \in T_p(M)$, for every $p \in U$). We also say that $X$ is a \textit{lifting of $U$ into $T(M)$}. We say that $X$ is a $C^k$-vector field on $U$ iff $X$ is a section over $U$ and a $C^k$-map. The set of $C^k$-vector fields over $U$ is denoted $\Gamma^{(k)}(U,T(M))$; see Figure 8.3.

Given a curve, $\gamma : [a,b] \to M$, a \textit{vector field $X$ along $\gamma$} is any section of $T(M)$ over $\gamma$, i.e., a $C^k$-function $X : [a,b] \to T(M)$, such that $\pi \circ X = \gamma$. We also say that $X$ lifts $\gamma$ into $T(M)$.

![Figure 8.3: A vector field on $S^1$ represented as the section $X$ in $T(S^1)$.](image)

Clearly, $\Gamma^{(k)}(U,T(M))$ is a real vector space. For short, the space $\Gamma^{(k)}(M,T(M))$ is also denoted by $\Gamma^{(k)}(T(M))$ (or $\mathfrak{X}^{(k)}(M)$, or even $\Gamma(T(M))$ or $\mathfrak{X}(M)$).

Remark: We can also define a $C^j$-vector field on $U$ as a section, $X$, over $U$ which is a $C^j$-map, where $0 \leq j \leq k$. Then, we have the vector space $\Gamma^{(j)}(U,T(M))$, etc.

If $M = \mathbb{R}^n$ and $U$ is an open subset of $M$, then $T(M) = \mathbb{R}^n \times \mathbb{R}^n$ and a section of $T(M)$ over $U$ is simply a function, $X$, such that

$$X(p) = (p, u), \quad \text{with} \quad u \in \mathbb{R}^n,$$

for all $p \in U$. In other words, $X$ is defined by a function, $f : U \to \mathbb{R}^n$ (namely, $f(p) = u$). This corresponds to the “old” definition of a vector field in the more basic case where the manifold, $M$, is just $\mathbb{R}^n$.

For any vector field $X \in \Gamma^{(k)}(U,T(M))$ and for any $p \in U$, we have $X(p) = (p, v)$ for some $v \in T_p(M)$, and it is convenient to denote the vector $v$ by $X_p$ so that $X(p) = (p, X_p)$. In fact, in most situations it is convenient to identify $X(p)$ with $X_p \in T_p(M)$, and we will do so from now on. This amounts to identifying the isomorphic vector spaces $\{p\} \times T_p(M)$ and $T_p(M)$, which we always do. Let us illustrate the advantage of this convention with the next definition.
Given any $C^k$-function $f \in C^k(U)$ and a vector field $X \in \Gamma^{(k)}(U, T(M))$, we define the vector field $fX$ by

$$(fX)_p = f(p)X_p, \quad p \in U.$$ 

Obviously, $fX \in \Gamma^{(k)}(U, T(M))$, which shows that $\Gamma^{(k)}(U, T(M))$ is also a $C^k(U)$-module.

For any chart $(U, \varphi)$ on $M$ it is easy to check that the map

$$p \mapsto \left( \frac{\partial}{\partial x_i} \right)_p, \quad p \in U,$$

is a $C^k$-vector field on $U$ (with $1 \leq i \leq n$). This vector field is denoted $\left( \frac{\partial}{\partial x_i} \right)$ or $\frac{\partial}{\partial x_i}$.

**Definition 8.2.** Let $M$ be a $C^{k+1}$ manifold and let $X$ be a $C^k$ vector field on $M$. If $U$ is any open subset of $M$ and $f$ is any function in $C^k(U)$, then the **Lie derivative of $f$ with respect to $X$**, denoted $X(f)$ or $L_X f$, is the function on $U$ given by

$$X(f)(p) = X_p(f) = X_p(f), \quad p \in U.$$ 

In particular, if $(U, \varphi)$ is any chart at $p$ and $X_p = \sum_{i=1}^n \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p$, then

$$X_p(f) = \sum_{i=1}^n \lambda_i \left( \frac{\partial}{\partial x_i} \right)_p f.$$ 

Observe that

$$X(f)(p) = df_p(X_p),$$

where $df_p$ is identified with the linear form in $T^*_p(M)$ defined by

$$df_p(v) = v(f), \quad v \in T_p M,$$

by identifying $T_0 \mathbb{R}$ with $\mathbb{R}$ (see the discussion following Proposition 7.12). The Lie derivative, $L_X f$, is also denoted $X[f]$.

As a special case, when $(U, \varphi)$ is a chart on $M$, the vector field, $\frac{\partial}{\partial x_i}$, just defined above induces the function

$$p \mapsto \left( \frac{\partial}{\partial x_i} \right)_p f, \quad p \in U,$$

denoted $\frac{\partial}{\partial x_i}(f)$ or $\left( \frac{\partial}{\partial x_i} \right)f$.

It is easy to check that $X(f) \in C^{k-1}(U)$. As a consequence, every vector field $X \in \Gamma^{(k)}(U, T(M))$ induces a linear map,

$$L_X : C^k(U) \longrightarrow C^{k-1}(U),$$
given by $f \mapsto X(f)$. It is immediate to check that $L_X$ has the Leibniz property, i.e.,

$$L_X(fg) = L_X(f)g + fL_X(g).$$

Linear maps with this property are called derivations. Thus, we see that every vector field induces some kind of differential operator, namely, a linear derivation. Unfortunately, not every linear derivation of the above type arises from a vector field, although this turns out to be true in the smooth case i.e., when $k = \infty$ (for a proof, see Gallot, Hulin and Lafontaine [73] or Lafontaine [110]).

In the rest of this section, unless stated otherwise, we assume that $k \geq 1$. The following easy proposition holds (c.f. Warner [175]):

**Proposition 8.2.** Let $X$ be a vector field on the $C^{k+1}$-manifold $M$, of dimension $n$. Then, the following are equivalent:

(a) $X$ is $C^k$.

(b) If $(U, \varphi)$ is a chart on $M$ and if $f_1, \ldots, f_n$ are the functions on $U$ uniquely defined by

$$X \restriction_U = \sum_{i=1}^n f_i \frac{\partial}{\partial x_i},$$

then each $f_i$ is a $C^k$-map.

(c) Whenever $U$ is open in $M$ and $f \in C^k(U)$, then $X(f) \in C^{k-1}(U)$.

Given any two $C^k$-vector field $X, Y$ on $M$, for any function $f \in C^k(M)$, we defined above the function $X(f)$ and $Y(f)$. Thus, we can form $X(Y(f))$ (resp. $Y(X(f))$), which are in $C^{k-2}(M)$. Unfortunately, even in the smooth case, there is generally no vector field $Z$ such that

$$Z(f) = X(Y(f)), \quad \text{for all } f \in C^k(M).$$

This is because $X(Y(f))$ (and $Y(X(f))$) involve second-order derivatives. However, if we consider $X(Y(f)) - Y(X(f))$, then second-order derivatives cancel out and there is a unique vector field inducing the above differential operator. Intuitively, $XY - YX$ measures the “failure of $X$ and $Y$ to commute.”

**Proposition 8.3.** Given any $C^{k+1}$-manifold $M$, of dimension $n$, for any two $C^k$-vector fields $X, Y$ on $M$, there is a unique $C^{k-1}$-vector field $[X, Y]$, such that

$$[X, Y](f) = X(Y(f)) - Y(X(f)), \quad \text{for all } f \in C^{k-1}(M).$$

**Proof.** First we prove uniqueness. For this it is enough to prove that $[X, Y]$ is uniquely defined on $C^k(U)$, for any chart, $(U, \varphi)$. Over $U$, we know that

$$X = \sum_{i=1}^n X_i \frac{\partial}{\partial x_i} \quad \text{and} \quad Y = \sum_{i=1}^n Y_i \frac{\partial}{\partial x_i},$$
where $X_i, Y_i \in \mathcal{C}^k(U)$. Then, for any $f \in \mathcal{C}^k(M)$, we have

$$X(Y(f)) = X \left( \sum_{j=1}^{n} Y_j \frac{\partial}{\partial x_j} (f) \right) = \sum_{i,j=1}^{n} X_i \frac{\partial}{\partial x_i} (Y_j) \frac{\partial}{\partial x_j} (f) + \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_j \partial x_i} (f)$$

$$Y(X(f)) = Y \left( \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i} (f) \right) = \sum_{i,j=1}^{n} Y_j \frac{\partial}{\partial x_j} (X_i) \frac{\partial}{\partial x_i} (f) + \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_i \partial x_j} (f).$$

However, as $f \in \mathcal{C}^k(M)$, with $k \geq 2$, we have

$$\sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_j \partial x_i} (f) = \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_i \partial x_j} (f),$$

and we deduce that

$$X(Y(f)) - Y(X(f)) = \sum_{i,j=1}^{n} \left( X_i \frac{\partial}{\partial x_i} (Y_j) - Y_i \frac{\partial}{\partial x_i} (X_j) \right) \frac{\partial}{\partial x_j} (f).$$

This proves that $[X,Y] = XY - YX$ is uniquely defined on $U$ and that it is $C^{k-1}$. Thus, if $[X,Y]$ exists, it is unique.

To prove existence, we use the above expression to define $[X,Y]_U$, locally on $U$, for every chart, $(U, \varphi)$. On any overlap, $U \cap V$, by the uniqueness property that we just proved, $[X,Y]_U$ and $[X,Y]_V$ must agree. Then we can define the vector field $[X,Y]$ as follows: for every chart $(U, \varphi)$, the restriction $[X,Y]$ to $U$ is equal to $[X,Y]_U$. This well defined because whenever two charts with domains $U$ and $V$ overlap, we know that $[X,Y]_U = [X,Y]_V$ agree. Therefore, $[X,Y]$ is a $C^{k-1}$-vector field defined on the whole of $M$.

**Definition 8.3.** Given any $C^{k+1}$-manifold $M$, of dimension $n$, for any two $C^k$-vector fields $X, Y$ on $M$, the *Lie bracket* $[X,Y]$ of $X$ and $Y$, is the $C^{k-1}$ vector field defined so that

$$[X,Y](f) = X(Y(f)) - Y(X(f)), \quad \text{for all} \quad f \in \mathcal{C}^{k-1}(M).$$

An an example, in $\mathbb{R}^3$, if $X$ and $Y$ are the two vector fields,

$$X = \frac{\partial}{\partial x} + y \frac{\partial}{\partial z} \quad \text{and} \quad Y = \frac{\partial}{\partial y},$$

then to compute $[X,Y]$, set $g = Y(f) = \frac{\partial f}{\partial y}$ and observe that

$$X(Y(f)) = X(g) = \frac{\partial g}{\partial x} + y \frac{\partial g}{\partial z} = \frac{\partial^2 f}{\partial x \partial y} + y \frac{\partial^2 f}{\partial z \partial y}.$$
Next set $h = X(f) = \frac{\partial f}{\partial x} + y \frac{\partial f}{\partial z}$ and calculate

$$Y(X(f)) = Y(h) = \frac{\partial}{\partial y} \left( \frac{\partial f}{\partial x} + y \frac{\partial f}{\partial z} \right) = \frac{\partial^2 f}{\partial y \partial x} + \frac{\partial f}{\partial z} + y \frac{\partial^2 f}{\partial y \partial z}.$$ 

Then

$$[X,Y](f) = X(Y(f)) - Y(X(f))$$

$$= \frac{\partial^2 f}{\partial x \partial y} + y \frac{\partial^2 f}{\partial z \partial y} - \frac{\partial f}{\partial y \partial x} - y \frac{\partial^2 f}{\partial y \partial z}$$

$$= -\frac{\partial f}{\partial z}.$$ 

Hence

$$[X,Y] = -\frac{\partial}{\partial z}.$$

We also have the following simple proposition whose proof is left as an exercise (or, see Do Carmo [60]):

**Proposition 8.4.** Given any $C^{k+1}$-manifold $M$, of dimension $n$, for any $C^k$-vector fields $X, Y, Z$ on $M$, for all $f, g \in C^k(M)$, we have:

(a) $[[X,Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$  
   (Jacobi identity).

(b) $[X, X] = 0$.

(c) $[fX, gY] = fg[X,Y] + fX(g)Y - gY(f)X$.

(d) $[-, -]$ is bilinear.

As a consequence, for smooth manifolds ($k = \infty$), the space of vector fields $\Gamma^{(\infty)}(T(M))$ is a vector space equipped with a bilinear operation $[-, -]$ that satisfies the Jacobi identity. This makes $\Gamma^{(\infty)}(T(M))$ a Lie algebra.

Let $h: M \to N$ be a diffeomorphism between two manifolds. Then, vector fields can be transported from $N$ to $M$ and conversely.

**Definition 8.4.** Let $h: M \to N$ be a diffeomorphism between two $C^{k+1}$ manifolds. For every $C^k$ vector field $Y$ on $N$, the **pull-back of $Y$ along $h$** is the vector field $h^*Y$ on $M$, given by

$$(h^*Y)_p = dh_{h(p)}^{-1}(Y_{h(p)}), \quad p \in M.$$ 

See Figure 8.4. For every $C^k$ vector field $X$ on $M$, the **push-forward of $X$ along $h$** is the vector field $h_*X$ on $N$, given by

$$h_*X = (h^{-1})^*X,$$

that is, for every $p \in M$, $(h_*X)_{h(p)} = dh_p(X_p)$, or equivalently,

$$(h_*X)_q = dh_{h^{-1}(q)}(X_{h^{-1}(q)}), \quad q \in N.$$ 

See Figure 8.5.
8.2. VECTOR FIELDS, LIE DERIVATIVE

Figure 8.4: The pull-back of the vector field $Y$

Figure 8.5: The push-forward of the vector field $X$
It is not hard to check that
\[ L_{h_*} f = L_X (f \circ h) \circ h^{-1}, \]
for any function \( f \in C^k(N) \). This is because
\[
(L_{h_*} f)_{h(p)} = (h_* X)_{h(p)}(f) \quad \text{by Definition 8.2}
\]
\[
= df_{h(p)} ((h_* X)_{h(p)}) \quad \text{by remark after Definition 8.2}
\]
\[
= df_{h(p)} (dh_p(X_p)) \quad \text{by Definition 8.4}
\]
\[
= d(f \circ h)_p (X_p) \quad \text{by the chain rule}
\]
\[
= d(f \circ h)_{h^{-1}(q)} (X_{h^{-1}(q)}) \quad p = h^{-1}(q)
\]
\[
= X_{h^{-1}(q)} (f \circ h) \quad \text{by remark after Definition 8.2}
\]
\[
= (L_X (f \circ h))_{h^{-1}(q)}. \]

One more notion will be needed when we deal with Lie algebras.

**Definition 8.5.** Let \( h: M \to N \) be a \( C^{k+1} \)-map of manifolds. If \( X \) is a \( C^k \) vector field on \( M \) and \( Y \) is a \( C^k \) vector field on \( N \), we say that \( X \) and \( Y \) are \( h \)-related iff

\[ dh \circ X = Y \circ h. \]

The basic result about \( h \)-related vector fields is:

**Proposition 8.5.** Let \( h: M \to N \) be a \( C^{k+1} \)-map of manifolds, let \( X \) and \( Y \) be \( C^k \) vector fields on \( M \) and let \( X_1, Y_1 \) be \( C^k \) vector fields on \( N \). If \( X \) is \( h \)-related to \( X_1 \) and \( Y \) is \( h \)-related to \( Y_1 \), then \( [X, Y] \) is \( h \)-related to \( [X_1, Y_1] \).

**Proof.** Basically, one needs to unwind the definitions, see Warner [175], Chapter 1. \( \square \)

If \( h: M \to N \) is a diffeomorphism, then for every vector field \( X \) on \( M \), since by definition of \( h_* X \), we have
\[
(h_* X)_{h(p)} = dh_p(X_p),
\]
the vector fields \( X \) and \( h_* X \) are \( h \)-related. Thus, as a corollary of Proposition 8.5, for any two vector fields \( X, Y \) on \( M \), we get
\[
h_* [X, Y] = [h_* X, h_* Y];
\]
that is,
\[
dh_p([X, Y]_p) = [dh_p(X_p), dh_p(Y_p)].\]
8.3 Integral Curves, Flow of a Vector Field, One-Parameter Groups of Diffeomorphisms

We begin with integral curves and (local) flows of vector fields on a manifold.

**Definition 8.6.** Let $X$ be a $C^{k-1}$ vector field on a $C^k$-manifold $M$ ($k \geq 2$), and let $p_0$ be a point on $M$. An integral curve (or trajectory) for $X$ with initial condition $p_0$ is a $C^{k-1}$-curve $\gamma: I \to M$, so that

$$\dot{\gamma}(t) = X_{\gamma(t)}$$

for all $t \in I$, and $\gamma(0) = p_0$,

where $I = (a, b) \subseteq \mathbb{R}$ is an open interval containing 0. See Figure 2.1.

What Definition 8.6 says is that an integral curve $\gamma$ with initial condition $p_0$ is a curve on the manifold $M$ passing through $p_0$, and such that for every point $p = \gamma(t)$ on this curve, the tangent vector to this curve at $p$, that is $\dot{\gamma}(t)$, coincides with the value $X_p$ of the vector field $X$ at $p$.

Given a vector field $X$ as above, and a point $p_0 \in M$, is there an integral curve through $p_0$? Is such a curve unique? If so, how large is the open interval $I$? We provide some answers to the above questions below.

**Definition 8.7.** Let $X$ be a $C^{k-1}$ vector field on a $C^k$-manifold $M$ ($k \geq 2$), and let $p_0$ be a point on $M$. A local flow for $X$ at $p_0$ is a map

$$\varphi: J \times U \to M,$$

where $J \subseteq \mathbb{R}$ is an open interval containing 0 and $U$ is an open subset of $M$ containing $p_0$, so that for every $p \in U$, the curve $t \mapsto \varphi(t, p)$ is an integral curve of $X$ with initial condition $p$. See Figure 2.2.

Thus, a local flow for $X$ is a family of integral curves for all points in some small open set around $p_0$ such that these curves all have the same domain $J$, independently of the initial condition $p \in U$.

The following theorem is the main existence theorem of local flows. This is a promoted version of a similar theorem in the classical theory of ODE’s in the case where $M$ is an open subset of $\mathbb{R}^n$. For a full account of this theory, see Lang [114] or Berger and Gostiaux [20].

**Theorem 8.6.** (Existence of a local flow) Let $X$ be a $C^{k-1}$ vector field on a $C^k$-manifold $M$ ($k \geq 2$), and let $p_0$ be a point on $M$. There is an open interval $J \subseteq \mathbb{R}$ containing 0 and an open subset $U \subseteq M$ containing $p_0$, so that there is a unique local flow $\varphi: J \times U \to M$ for $X$ at $p_0$. What this means is that if $\varphi_1: J \times U \to M$ and $\varphi_2: J \times U \to M$ are both local flows with domain $J \times U$, then $\varphi_1 = \varphi_2$. Furthermore, $\varphi$ is $C^{k-1}$.

---

1Recall our convention: if $X$ is a vector field on $M$, then for every point $q \in M$ we identify $X(q) = (q, X_q)$ and $X_q$. 
However, there could be two (or more) integral curves \( \gamma_1: I_1 \to M \) and \( \gamma_2: I_2 \to M \) with initial condition \( p_0 \). This leads to the natural question: How do \( \gamma_1 \) and \( \gamma_2 \) differ on \( I_1 \cap I_2 \)? The next proposition shows they don’t!

**Proposition 8.7.** Let \( X \) be a \( C^{k-1} \) vector field on a \( C^k \)-manifold \( M \) \((k \geq 2)\), and let \( p_0 \) be a point on \( M \). If \( \gamma_1: I_1 \to M \) and \( \gamma_2: I_2 \to M \) are any two integral curves both with initial condition \( p_0 \), then \( \gamma_1 = \gamma_2 \) on \( I_1 \cap I_2 \).

**Proof.** Let \( Q = \{ t \in I_1 \cap I_2 \mid \gamma_1(t) = \gamma_2(t) \} \). Since \( \gamma_1(0) = \gamma_2(0) = p_0 \), the set \( Q \) is nonempty. If we show that \( Q \) is both closed and open in \( I_1 \cap I_2 \), as \( I_1 \cap I_2 \) is connected since it is an open interval of \( \mathbb{R} \), we will be able to conclude that \( Q = I_1 \cap I_2 \).

Since by definition, a manifold is Hausdorff, it is a standard fact in topology that the diagonal \( \Delta = \{ (p, p) \mid p \in M \} \subseteq M \times M \) is closed, and since

\[
Q = I_1 \cap I_2 \cap (\gamma_1, \gamma_2)^{-1}(\Delta)
\]

and \( \gamma_1 \) and \( \gamma_2 \) are continuous, we see that \( Q \) is closed in \( I_1 \cap I_2 \).

Pick any \( u \in Q \) and consider the curves \( \beta_1 \) and \( \beta_2 \) given by

\[
\beta_1(t) = \gamma_1(t + u) \quad \text{and} \quad \beta_2(t) = \gamma_2(t + u),
\]

where \( t \in I_1 - u \) in the first case, and \( t \in I_2 - u \) in the second. (Here, if \( I = (a, b) \), we have \( I - u = (a - u, b - u) \).) Observe that

\[
\dot{\beta}_1(t) = \dot{\gamma}_1(t + u) = X(\gamma_1(t + u)) = X(\beta_1(t)),
\]

and similarly \( \dot{\beta}_2(t) = X(\beta_2(t)) \). We also have

\[
\beta_1(0) = \gamma_1(u) = \gamma_2(u) = \beta_2(0) = q,
\]

since \( u \in Q \) (where \( \gamma_1(u) = \gamma_2(u) \)). Thus, \( \beta_1: (I_1 - u) \to M \) and \( \beta_2: (I_2 - u) \to M \) are two integral curves with the same initial condition \( q \). By Theorem 8.6, the uniqueness of local flow implies that there is some open interval \( \bar{I} \subseteq I_1 \cap I_2 - u \), such that \( \beta_1 = \beta_2 \) on \( \bar{I} \). Consequently, \( \gamma_1 \) and \( \gamma_2 \) agree on \( \bar{I} + u \), an open subset of \( Q \), proving that \( Q \) is indeed open in \( I_1 \cap I_2 \).

Proposition 8.7 implies the important fact that there is a unique maximal integral curve with initial condition \( p \). Indeed, if \( \{ \gamma_j: I_j \to M \}_{j \in K} \) is the family of all integral curves with initial condition \( p \) (for some big index set \( K \)), if we let \( I(p) = \bigcup_{j \in K} I_j \), we can define a curve \( \gamma_p: I(p) \to M \) so that

\[
\gamma_p(t) = \gamma_j(t), \quad \text{if} \quad t \in I_j.
\]

Since \( \gamma_j \) and \( \gamma_l \) agree on \( I_j \cap I_l \) for all \( j, l \in K \), the curve \( \gamma_p \) is indeed well defined, and it is clearly an integral curve with initial condition \( p \) with the largest possible domain (the open interval, \( I(p) \)).
Definition 8.8. The curve \( \gamma_p \) defined above is called the \textit{maximal integral curve with initial condition} \( p \), and it is also denoted by \( \gamma(p,t) \). The domain of \( \gamma_p \) is \( I(p) \).

Note that Proposition 8.7 implies that any two distinct integral curves are disjoint, i.e., do not intersect each other.

Consider the vector field in \( \mathbb{R}^2 \) given by

\[
X = -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y}
\]

shown in Figure 8.6. If we write \( \gamma(t) = (x(t), y(t)) \), the differential equation \( \dot{\gamma}(t) = X(\gamma(t)) \)

![Figure 8.6: A vector field in \( \mathbb{R}^2 \)](image)

is expressed by

\[
\begin{align*}
x'(t) & = -y(t) \\
y'(t) & = x(t),
\end{align*}
\]

or in matrix form,

\[
\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.
\]

If we write \( X = \begin{pmatrix} x \\ y \end{pmatrix} \) and \( A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \), then the above equation is written as

\[
X' = AX.
\]

Now, as

\[
e^{tA} = I + \frac{A}{1!} t + \frac{A^2}{2!} t^2 + \cdots + \frac{A^n}{n!} t^n + \cdots,
\]

we get

\[
\frac{d}{dt}(e^{tA}) = A + \frac{A^2}{1!} t + \frac{A^3}{2!} t^2 + \cdots + \frac{A^n}{(n-1)!} t^{n-1} + \cdots = A e^{tA},
\]
so we see that $e^{tA}p$ is a solution of the ODE $X' = AX$ with initial condition $X = p$, and by uniqueness, $X = e^{tA}p$ is the solution of our ODE starting at $X = p$. Thus, our integral curve $\gamma_p$ through $p = (x_0)$ is the circle given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

Observe that $I(p) = \mathbb{R}$, for every $p \in \mathbb{R}^2$.

If we delete the points $(-1,0)$ and $(1,0)$ on the $x$-axis, then for every point $p_0$ not on the unit circle $S^1$ (given by $x^2 + y^2 = 1$), the maximal integral curve through $p_0$ is the circle of center $O$ through $p_0$, as before. However, for every point $p_0$ on the open upper half unit circle $S^1_+$, the maximal integral curve through $p_0$ is $S^1_+$, and for every point $p_0$ on the open lower half unit circle $S^1_-$, the maximal integral curve through $p_0$ is $S^1_-$. In both cases, the domain of the integral curve is an open interval properly contained in $\mathbb{R}$. This example shows that it may not be possible to extend the domain of an integral curve to the entire real line.

Here is one more example of a vector field on $M = \mathbb{R}$ that has integral curves not defined on the whole of $\mathbb{R}$. Let $X$ be the vector field on $\mathbb{R}$ given by

$$X(x) = (1 + x^2) \frac{\partial}{\partial x}.$$ 

It is easy to see that the maximal integral curve with initial condition $p_0 = 0$ is the curve $\gamma : (-\pi/2, \pi/2) \to \mathbb{R}$ given by

$$\gamma(t) = \tan t.$$

The following interesting question now arises: Given any $p_0 \in M$, if $\gamma_{p_0} : I(p_0) \to M$ is the maximal integral curve with initial condition $p_0$, and for any $t_1 \in I(p_0)$, if $p_1 = \gamma_{p_0}(t_1) \in M$, then there is a maximal integral curve $\gamma_{p_1} : I(p_1) \to M$ with initial condition $p_1$; what is the relationship between $\gamma_{p_0}$ and $\gamma_{p_1}$, if any? The answer is given by

**Proposition 8.8.** Let $X$ be a $C^{k-1}$ vector field on a $C^k$-manifold $M$ ($k \geq 2$), and let $p_0$ be a point on $M$. If $\gamma_{p_0} : I(p_0) \to M$ is the maximal integral curve with initial condition $p_0$, for any $t_1 \in I(p_0)$, if $p_1 = \gamma_{p_0}(t_1) \in M$ and $\gamma_{p_1} : I(p_1) \to M$ is the maximal integral curve with initial condition $p_1$, then

$$I(p_1) = I(p_0) - t_1 \quad \text{and} \quad \gamma_{p_1}(t) = \gamma_{\gamma_{p_0}(t_1)}(t) = \gamma_{p_0}(t + t_1), \quad \text{for all } t \in I(p_0) - t_1.$$

**Proof.** Let $\gamma(t)$ be the curve given by

$$\gamma(t) = \gamma_{p_0}(t + t_1), \quad \text{for all } t \in I(p_0) - t_1.$$ 

Clearly $\gamma$ is defined on $I(p_0) - t_1$, and

$$\dot{\gamma}(t) = \dot{\gamma}_{p_0}(t + t_1) = X(\gamma_{p_0}(t + t_1)) = X(\gamma(t)).$$
and \( \gamma(0) = \gamma_{p_0}(t_1) = p_1 \). Thus, \( \gamma \) is an integral curve defined on \( I(p_0) - t_1 \) with initial condition \( p_1 \). If \( \gamma \) was defined on an interval \( \bar{I} \supset I(p_0) - t_1 \) with \( \bar{I} \neq I(p_0) - t_1 \), then \( \gamma_{p_0} \) would be defined on \( \bar{I} + t_1 \supset I(p_0) \), an interval strictly bigger than \( I(p_0) \), contradicting the maximality of \( I(p_0) \). Therefore, \( I(p_0) - t_1 = I(p_1) \).

Proposition 8.8 says that the traces \( \gamma_{p_0} \) and \( \gamma_{p_1} \) in \( M \) of the maximal integral curves \( \gamma_{p_0} \) and \( \gamma_{p_1} \) are identical; they only differ by a simple reparametrization \((u = t + t_1)\).

It is useful to restate Proposition 8.8 by changing point of view. So far, we have been focusing on integral curves: given any \( p_0 \in M \), we let \( t \) vary in \( I(p_0) \) and get an integral curve \( \gamma_{p_0} \) with domain \( I(p_0) \). Instead of holding \( p_0 \in M \) fixed, we can hold \( t \in \mathbb{R} \) fixed and consider the set

\[
D_t(X) = \{ p \in M \mid t \in I(p) \},
\]

the set of points such that it is possible to “travel for \( t \) units of time from \( p \)” along the maximal integral curve \( \gamma_p \) with initial condition \( p \) (It is possible that \( D_t(X) = \emptyset \)). By definition, if \( D_t(X) \neq \emptyset \), the point \( \gamma_p(t) \) is well defined, and so we obtain a map \( \Phi_t^X : D_t(X) \to M \) with domain \( D_t(X) \), given by

\[
\Phi_t^X(p) = \gamma_p(t).
\]

The above suggests the following definition:

**Definition 8.9.** Let \( X \) be a \( C^{k-1} \) vector field on a \( C^k \)-manifold \( M \) \((k \geq 2)\). For any \( t \in \mathbb{R} \), let

\[
D_t(X) = \{ p \in M \mid t \in I(p) \} \quad \text{and} \quad D(X) = \{ (t, p) \in \mathbb{R} \times M \mid t \in I(p) \},
\]

and let \( \Phi^X : D(X) \to M \) be the map given by

\[
\Phi^X(t, p) = \gamma_p(t).
\]

The map \( \Phi^X \) is called the (global) flow of \( X \), and \( D(X) \) is called its domain of definition. For any \( t \in \mathbb{R} \) such that \( D_t(X) \neq \emptyset \), the map \( p \in D_t(X) \mapsto \Phi^X(t, p) = \gamma_p(t) \) is denoted by \( \Phi^X_t \) \((\text{i.e., } \Phi^X_t(p) = \Phi^X(t, p) = \gamma_p(t))\).

Observe that

\[
D(X) = \bigcup_{p \in M} (I(p) \times \{ p \}).
\]

Also, using the \( \Phi^X_t \) notation, the property of Proposition 8.8 reads

\[
\Phi^X_s \circ \Phi^X_t = \Phi^X_{s+t},
\]

whenever both sides of the equation make sense. Indeed, the above says

\[
\Phi^X_s(\Phi^X_t(p)) = \Phi^X_s(\gamma_p(t)) = \gamma_{p(t)}(s) = \gamma_p(s + t) = \Phi^X_{s+t}(p).
\]
Using the above property, we can easily show that the $\Phi^X_t$ are invertible. In fact, the inverse of $\Phi^X_t$ is $\Phi^{-X}_{-t}$. First, note that

$$D_0(X) = M \quad \text{and} \quad \Phi^X_0 = \text{id},$$

because, by definition, $\Phi^X_0(p) = \gamma_p(0) = p$, for every $p \in M$. Then, (*) implies that

$$\Phi^X_t \circ \Phi^{-X}_{-t} = \Phi^X_{t-t} = \Phi^X_0 = \text{id},$$

which shows that $\Phi^X_t : D_t(X) \to D_{-t}(X)$ and $\Phi^{-X}_{-t} : D_{-t}(X) \to D_t(X)$ are inverse of each other. Moreover, each $\Phi^X_t$ is a $C^{k-1}$-diffeomorphism. We summarize in the following proposition some additional properties of the domains $D(X)$, $D_t(X)$ and the maps $\Phi^X_t$ (for a proof, see Lang [114] or Warner [175]):

**Theorem 8.9.** Let $X$ be a $C^{k-1}$ vector field on a $C^k$-manifold $M$ $(k \geq 2)$. The following properties hold:

(a) For every $t \in \mathbb{R}$, if $D_t(X) \neq \emptyset$, then $D_t(X)$ is open (this is trivially true if $D_t(X) = \emptyset$).

(b) The domain $D(X)$ of the flow $\Phi^X$ is open, and the flow is a $C^{k-1}$ map $\Phi^X : D(X) \to M$.

(c) Each $\Phi^X_t : D_t(X) \to D_{-t}(X)$ is a $C^{k-1}$-diffeomorphism with inverse $\Phi^{-X}_{-t}$.

(d) For all $s, t \in \mathbb{R}$, the domain of definition of $\Phi^X_s \circ \Phi^X_t$ is contained but generally not equal to $D_{s+t}(X)$. However, $\text{dom}(\Phi^X_s \circ \Phi^X_t) = D_{s+t}(X)$ if $s$ and $t$ have the same sign. Moreover, on $\text{dom}(\Phi^X_s \circ \Phi^X_t)$, we have

$$\Phi^X_s \circ \Phi^X_t = \Phi^X_{s+t}.$$

**Remarks:**

(1) We may omit the superscript $X$ and write $\Phi$ instead of $\Phi^X$ if no confusion arises.

(2) The reason for using the terminology flow in referring to the map $\Phi^X$ can be clarified as follows: For any $t$ such that $D_t(X) \neq \emptyset$, every integral curve $\gamma_p$ with initial condition $p \in D_t(X)$ is defined on some open interval containing $[0, t]$, and we can picture these curves as “flow lines” along which the points $p$ flow (travel) for a time interval $t$. Then, $\Phi^X(t, p)$ is the point reached by “flowing” for the amount of time $t$ on the integral curve $\gamma_p$ (through $p$) starting from $p$. Intuitively, we can imagine the flow of a fluid through $M$, and the vector field $X$ is the field of velocities of the flowing particles.

Given a vector field $X$ as above, it may happen that $D_t(X) = M$, for all $t \in \mathbb{R}$. 

**Definition 8.10.** When $D(X) = \mathbb{R} \times M$, we say that the vector field $X$ is *complete*. Then, the $\Phi^X_t$ are diffeomorphisms of $M$, and they form a group. The family $\{\Phi^X_t\}_{t \in \mathbb{R}}$ is called a 1-parameter group of $X$.

If the vector field $X$ is complete, then $\Phi^X$ induces a group homomorphism $(\mathbb{R}, +) \rightarrow \text{Diff}(M)$, from the additive group $\mathbb{R}$ to the group of $C^{k-1}$-diffeomorphisms of $M$.

By abuse of language, even when it is not the case that $D_t(X) = M$ for all $t$, the family $\{\Phi^X_t\}_{t \in \mathbb{R}}$ is called a local 1-parameter group generated by $X$, even though it is not a group, because the composition $\Phi^X_s \circ \Phi^X_t$ may not be defined.

If we go back to the vector field in $\mathbb{R}^2$ given by

$$X = -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y},$$

since the integral curve $\gamma_p(t)$, through $p = (x_0, y_0)$ is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

the global flow associated with $X$ is given by

$$\Phi^X(t, p) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} p,$$

and each diffeomorphism, $\Phi^X_t$ is the rotation

$$\Phi^X_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$ 

The 1-parameter group $\{\Phi^X_t\}_{t \in \mathbb{R}}$ generated by $X$ is the group of rotations in the plane, $\text{SO}(2)$.

More generally, if $B$ is an $n \times n$ invertible matrix that has a real logarithm $A$ (that is, if $e^A = B$), then the matrix $A$ defines a vector field $X$ in $\mathbb{R}^n$, with

$$X = \sum_{i,j=1}^n (a_{ij}x_j) \frac{\partial}{\partial x_i},$$

whose integral curves are of the form

$$\gamma_p(t) = e^{tA}p,$$

and we have

$$\gamma_p(1) = Bp.$$ 

The one-parameter group $\{\Phi^X_t\}_{t \in \mathbb{R}}$ generated by $X$ is given by $\{e^{tA}\}_{t \in \mathbb{R}}$.

When $M$ is compact, it turns out that every vector field is complete, a nice and useful fact.
Proposition 8.10. Let $X$ be a $C^{k-1}$ vector field on a $C^k$-manifold $M$ ($k \geq 2$). If $M$ is compact, then $X$ is complete, which means that $\mathcal{D}(X) = \mathbb{R} \times M$. Moreover, the map $t \mapsto \Phi^X_t$ is a homomorphism from the additive group $\mathbb{R}$ to the group $\text{Diff}(M)$ of $(C^{k-1})$ diffeomorphisms of $M$.

Proof. Pick any $p \in M$. By Theorem 8.6, there is a local flow $\varphi_p: J(p) \times U(p) \to M$, where $J(p) \subseteq \mathbb{R}$ is an open interval containing 0 and $U(p)$ is an open subset of $M$ containing $p$, so that for all $q \in U(p)$, the map $t \mapsto \varphi(t, q)$ is an integral curve with initial condition $q$ (where $t \in J(p)$). Thus, we have $J(p) \times U(p) \subseteq \mathcal{D}(X)$. Now, the $U(p)$'s form an open cover of $M$, and since $M$ is compact, we can extract a finite subcover $\bigcup_{q \in F} U(q) = M$, for some finite subset $F \subseteq M$. But then, we can find $\epsilon > 0$ so that $(-\epsilon, \epsilon) \subseteq J(q)$, for all $q \in F$ and for all $t \in (-\epsilon, \epsilon)$, and for all $p \in M$, if $\gamma_p$ is the maximal integral curve with initial condition $p$, then $(-\epsilon, \epsilon) \subseteq I(p)$.

For any $t \in (-\epsilon, \epsilon)$, consider the integral curve $\gamma_{\gamma_p(t)}$, with initial condition $\gamma_p(t)$. This curve is well defined for all $t \in (-\epsilon, \epsilon)$, and we have

$$
\gamma_{\gamma_p(t)}(t) = \gamma_p(t + t) = \gamma_p(2t),
$$

which shows that $\gamma_p$ is in fact defined for all $t \in (-2\epsilon, +2\epsilon)$. By induction we see that

$$
(-2^n \epsilon, +2^n \epsilon) \subseteq I(p),
$$

for all $n \geq 0$, which proves that $I(p) = \mathbb{R}$. As this holds for all $p \in M$, we conclude that $\mathcal{D}(X) = \mathbb{R} \times M$. \qed

Remarks:

1. The proof of Proposition 8.10 also applies when $X$ is a vector field with compact support (this means that the closure of the set $\{p \in M \mid X(p) \neq 0\}$ is compact).

2. If $h: M \to N$ is a diffeomorphism and $X$ is a vector field on $M$, then it can be shown that the local 1-parameter group associated with the vector field $h_* X$ is

$$
\{h \circ \Phi^X_t \circ h^{-1}\}_{t \in \mathbb{R}}.
$$

A point $p \in M$ where a vector field vanishes (i.e., $X(p) = 0$) is called a critical point of $X$. Critical points play a major role in the study of vector fields, in differential topology (e.g., the celebrated Poincaré–Hopf index theorem), and especially in Morse theory, but we won’t go into this here (curious readers should consult Milnor [125], Guillemin and Pollack [83] or DoCarmo [59], which contains an informal but very clear presentation of the Poincaré–Hopf index theorem). Another famous theorem about vector fields says that every smooth
In the case of a vector field on a sphere of even dimension \((S^{2n})\) must vanish in at least one point (the so-called “hairy-ball theorem.”) On \(S^2\), it says that you can’t comb your hair without having a singularity somewhere. Try it, it’s true!

Let us just observe that if an integral curve \(\gamma\) passes through a critical point \(p\), then \(\gamma\) is reduced to the point \(p\); that is, \(\gamma(t) = p\), for all \(t\). Indeed, such a curve is an integral curve with initial condition \(p\). By the uniqueness property, it is the only one. Then, we see that if a maximal integral curve is defined on the whole of \(\mathbb{R}\), either it is injective (it has no self-intersection), or it is simply periodic (i.e., there is some \(T > 0\) so that \(\gamma(t + T) = \gamma(t)\), for all \(t \in \mathbb{R}\) and \(\gamma\) is injective on \([0, T]\)), or it is reduced to a single point.

We conclude this section with the definition of the Lie derivative of a vector field with respect to another vector field.

Say we have two vector fields \(X\) and \(Y\) on \(M\). For any \(p \in M\), we can flow along the integral curve of \(X\) with initial condition \(p\) to \(\Phi_t(p)\) (for \(t\) small enough) and then evaluate \(Y\) there, getting \(Y(\Phi_t(p))\). Now, this vector belongs to the tangent space \(T_{\Phi_t(p)}(M)\), but \(Y(p)\) belongs to \(T_p(M)\). So, to “compare” \(Y(\Phi_t(p))\) and \(Y(p)\), we bring back \(Y(\Phi_t(p))\) to \(T_p(M)\) by applying the tangent map \(d\Phi_{-t}\) at \(\Phi_t(p)\) to \(Y(\Phi_t(p))\) (Note that to alleviate the notation, we use the slight abuse of notation \(d\Phi_{-t}\) instead of \(d(\Phi_{-t})_{\Phi_t(p)}\).) Then, we can form the difference \(d\Phi_{-t}(Y(\Phi_t(p))) - Y(p)\), divide by \(t\), and consider the limit as \(t\) goes to 0.

**Definition 8.11.** Let \(M\) be a \(C^{k+1}\) manifold. Given any two \(C^k\) vector fields \(X\) and \(Y\) on \(M\), for every \(p \in M\), the Lie derivative of \(Y\) with respect to \(X\) at \(p\) denoted \((L_X Y)_p\), is given by

\[
(L_X Y)_p = \lim_{t \to 0} \frac{d\Phi_{-t}(Y(\Phi_t(p))) - Y(p)}{t} = \frac{d}{dt}(d\Phi_{-t}(Y(\Phi_t(p))))\bigg|_{t=0}.
\]

It can be shown that \((L_X Y)_p\) is our old friend the Lie bracket; that is,

\[(L_X Y)_p = [X, Y]_p.\]

For a proof, see Warner [175] (Chapter 2, Proposition 2.25) or O’Neill [138] (Chapter 1, Proposition 58).

In terms of Definition 8.4, observe that

\[
(L_X Y)_p = \lim_{t \to 0} \frac{((\Phi_{-t})_* Y)(p) - Y(p)}{t} = \lim_{t \to 0} \frac{(\Phi_t^* Y)(p) - Y(p)}{t} = \frac{d}{dt}(\Phi_t^* Y)(p)\bigg|_{t=0},
\]

since \((\Phi_{-t})^{-1} = \Phi_t\).

Next, we discuss the application of vector fields and integral curves to the blending of locally affine transformations, known as Log-Euclidean polyaffine transformations, as presented in Arsigny, Commowick, Pennec and Ayache [7].
8.4 Log-Euclidean Polyaffine Transformations

The registration of medical images is an important and difficult problem. The work described in Arsigny, Commowick, Pennec and Ayache [7] (and Arsigny’s thesis [6]) makes an original
and valuable contribution to this problem by describing a method for parametrizing a class
of non-rigid deformations with a small number of degrees of freedom. After a global affine
alignment, this sort of parametrization allows a finer local registration with very smooth
transformations. This type of parametrization is particularly well adapted to the registration
of histological slices, see Arsigny, Pennec and Ayache [9].

The goal is to fuse some affine or rigid transformations in such a way that the resulting
transformation is invertible and smooth. The direct approach which consists in blending
$N$ global affine or rigid transformations $T_1, \ldots, T_N$ using weights $w_1, \ldots, w_N$ does not work,
because the resulting transformation

$$T = \sum_{i=1}^{N} w_i T_i$$

is not necessarily invertible. The purpose of the weights is to define the domain of influence
in space of each $T_i$.

The key idea is to associate to each rigid (or affine) transformation $T$ of $\mathbb{R}^n$ a vector field
$V$, and to view $T$ as the diffeomorphism $\Phi_V^1$ corresponding to the time $t = 1$, where $\Phi_V^t$ is
the global flow associated with $V$. In other words, $T$ is the result of integrating an ODE

$$X' = V(X, t),$$

starting with some initial condition $X_0$, and $T = X(1)$.

Now, it would be highly desirable if the vector field $V$ did not depend on the time
parameter, and this is indeed possible for a large class of affine transformations, which is one
of the nice contributions of the work of Arsigny, Commowick, Pennec and Ayache [7]. Recall
that an affine transformation $X \mapsto LX + v$ (where $L$ is an $n \times n$ matrix and $X, v \in \mathbb{R}^n$) can
be conveniently represented as a linear transformation from $\mathbb{R}^{n+1}$ to itself if we write

$$\begin{pmatrix} X \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} L & v \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}.$$ 

Then the ODE with constant coefficients

$$X' = LX + v$$

can be written

$$\begin{pmatrix} X' \\ 0 \end{pmatrix} = \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix},$$
and for every initial condition \( X = X_0 \), its unique solution is given by

\[
\begin{pmatrix}
X(t) \\
1
\end{pmatrix} = \exp \left( t \begin{pmatrix}
L & v \\
0 & 0
\end{pmatrix} \right) \begin{pmatrix}
X_0 \\
1
\end{pmatrix}.
\]

Therefore, if we can find reasonable conditions on matrices \( T = \begin{pmatrix}
M & u \\
0 & 1
\end{pmatrix} \) to ensure that they have a unique real logarithm

\[
\log(T) = \begin{pmatrix}
L & v \\
0 & 0
\end{pmatrix},
\]

then we will be able to associate a vector field \( V(X) = LX + v \) to \( T \), in such a way that \( T \) is recovered by integrating the ODE \( X' = LX + v \). Furthermore, given \( N \) transformations \( T_1, \ldots, T_N \) such that \( \log(T_1), \ldots, \log(T_N) \) are uniquely defined, we can fuse \( T_1, \ldots, T_N \) at the \textit{infinitesimal level} by defining the ODE obtained by blending the vector fields \( V_1, \ldots, V_N \) associated with \( T_1, \ldots, T_N \) (with \( V_i(X) = L_iX + v_i \)), namely

\[
V(X) = \sum_{i=1}^{N} w_i(X)(L_iX + v_i).
\]

Then, it is easy to see that the ODE

\[
X' = V(X)
\]

has a unique solution for every \( X = X_0 \) defined for all \( t \), and the fused transformation is just \( T = X(1) \). Thus, the fused vector field

\[
V(X) = \sum_{i=1}^{N} w_i(X)(L_iX + v_i)
\]

yields a one-parameter group of diffeomorphisms \( \Phi_t \). Each transformation \( \Phi_t \) is smooth and invertible, and is called a \textit{Log-Euclidean polyaffine tranformation}, for short, \textit{LEPT}. Of course, we have the equation

\[
\Phi_{s+t} = \Phi_s \circ \Phi_t,
\]

for all \( s, t \in \mathbb{R} \), so in particular, the inverse of \( \Phi_t \) is \( \Phi_{-t} \). We can also interpret \( \Phi_s \) as \((\Phi_1)^s\), which will yield a fast method for computing \( \Phi_s \). Observe that when the weight are scalars, the one-parameter group is given by

\[
\begin{pmatrix}
\Phi_t(X) \\
1
\end{pmatrix} = \exp \left( t \sum_{i=1}^{N} w_i \begin{pmatrix}
L_i & v_i \\
0 & 0
\end{pmatrix} \right) \begin{pmatrix}
X \\
1
\end{pmatrix},
\]

which is the Log-Euclidean mean of the affine transformations \( T_i \)'s (w.r.t. the weights \( w_i \)).
Fortunately, there is a sufficient condition for a real matrix to have a unique real logarithm and this condition is not too restrictive in practice.

Recall that $\mathcal{S}(n)$ denotes the set of all real matrices whose eigenvalues $\lambda + i\mu$ lie in the horizontal strip determined by the condition $-\pi < \mu < \pi$. We have the following version of Theorem 2.27:

**Theorem 8.11.** The image $\exp(\mathcal{S}(n))$ of $\mathcal{S}(n)$ by the exponential map is the set of real invertible matrices with no negative eigenvalues and $\exp: \mathcal{S}(n) \rightarrow \exp(\mathcal{S}(n))$ is a bijection.

Theorem 8.11 is stated in Kenney and Laub [100] without proof. Instead, Kenney and Laub cite DePrima and Johnson [49] for a proof, but this latter paper deals with complex matrices and does not contain a proof of our result either. The injectivity part of Theorem 8.11 can be found in Mneimné and Testard [130], Chapter 3, Theorem 3.8.4.

In fact, $\exp: \mathcal{S}(n) \rightarrow \exp(\mathcal{S}(n))$ is a diffeomorphism, a result proved in Bourbaki [28]; see Chapter III, Section 6.9, Proposition 17 and Theorem 6. Curious readers should read Gallier [71] for the full story.

For any matrix $A \in \exp(\mathcal{S}(n))$, we refer to the unique matrix $X \in \mathcal{S}(n)$ such that $e^X = A$ as the principal logarithm of $A$, and we denote it as $\log A$.

Observe that if $T$ is an affine transformation given in matrix form by

$$T = \begin{pmatrix} M & t \\ 0 & 1 \end{pmatrix},$$

since the eigenvalues of $T$ are those of $M$ plus the eigenvalue $1$, the matrix $T$ has no negative eigenvalues iff $M$ has no negative eigenvalues, and thus the principal logarithm of $T$ exists iff the principal logarithm of $M$ exists.

It is proved in Arsigny, Commowick, Pennec and Ayache that LEPT’s are affine invariant; see [7], Section 2.3. This shows that LEPT’s are produced by a truly geometric kind of blending, since the result does not depend at all on the choice of the coordinate system.

In the next section, we describe a fast method for computing due to Arsigny, Commowick, Pennec and Ayache [7].

### 8.5 Fast Polyaffine Transforms

Recall that since LEPT’s are members of the one-parameter group $(\Phi_t)_{t \in \mathbb{R}}$, we have

$$\Phi_{2t} = \Phi_{t+t} = \Phi_t^2,$$

and thus,

$$\Phi_1 = (\Phi_{1/2^N})^{2^N}.$$
Observe the formal analogy of the above formula with the formula

$$\exp(M) = \exp \left( \frac{M}{2^N} \right)^{2^N}$$

for computing the exponential of a matrix $M$ by the scaling and squaring method.

It turns out that the “scaling and squaring method” is one of the most efficient methods for computing the exponential of a matrix; see Kenney and Laub [100] and Higham [89]. The key idea is that $\exp(M)$ is easy to compute if $M$ is close zero, since in this case, one can use a few terms of the exponential series, or better a Padé approximant (see Higham [89]). The scaling and squaring method for computing the exponential of a matrix $M$ can be sketched as follows:

1. **Scaling Step**: Divide $M$ by a factor $2^N$, so that $\frac{M}{2^N}$ is close enough to zero.

2. **Exponentiation step**: Compute $\exp \left( \frac{M}{2^N} \right)$ with high precision, for example using a Padé approximant.

3. **Squaring Step**: Square $\exp \left( \frac{M}{2^N} \right)$ repeatedly $N$ times to obtain $\exp \left( \frac{M}{2^N} \right)^{2^N}$, a very accurate approximation of $e^M$.

There is also a so-called inverse scaling and squaring method to compute efficiently the principal logarithm of a real matrix; see Cheng, Higham, Kenney and Laub [39].

Arsigny, Commowick, Pennec and Ayache made the very astute observation that the scaling and squaring method can be adapted to compute LEPT’s very efficiently [7]. This method called fast polyaffine transform computes the values of a Log-Euclidean polyaffine transformation $T = \Phi_1$ at the vertices of a regular $n$-dimensional grid (in practice, for $n = 2$ or $n = 3$). Recall that $T$ is obtained by integrating an ODE $X' = V(X)$, where the vector field $V$ is obtained by blending the vector fields associated with some affine transformations $T_1, \ldots, T_n$, having a principal logarithm.

Here are the three steps of the fast polyaffine transform:

1. **Scaling Step**: Divide the vector field $V$ by a factor $2^N$, so that $\frac{V}{2^N}$ is close enough to zero.

2. **Exponentiation step**: Compute $\Phi_{1/2^N}$, using some adequate numerical integration method.

3. **Squaring Step**:Compose $\Phi_{1/2^N}$ with itself recursively $N$ times to obtain an accurate approximation of $T = \Phi_1$. 
Of course, one has to provide practical methods to achieve step 2 and step 3. Several methods to achieve step 2 and step 3 are proposed in Arsigny, Commowick, Pennec and Ayache [7]. One also has to worry about boundary effects, but this problem can be alleviated too, using bounding boxes. At this point, the reader is urged to read the full paper [7] for complete details and beautiful pictures illustrating the use of LEPT’s in medical imaging.

For more details regarding the LEPT, including the Log-Euclidean framework for locally rigid or affine deformation, the reader should read Arsigny, Commowick, Pennec and Ayache [7].
Chapter 9

Partitions of Unity, Covering Maps

This chapter contains a selection of technical tools. It is preparatory for best understanding certain proofs which occur in the remaining chapters.

9.1 Partitions of Unity

To study manifolds, it is often necessary to construct various objects such as functions, vector fields, Riemannian metrics, volume forms, etc., by gluing together items constructed on the domains of charts. Partitions of unity are a crucial technical tool in this gluing process.

The first step is to define “bump functions” (also called plateau functions). For any $r > 0$, we denote by $B(r)$ the open ball

$$B(r) = \{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 < r \},$$

and by $\overline{B(r)} = \{ (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 \leq r \}$ its closure.

**Proposition 9.1.** There is a smooth function $b: \mathbb{R}^n \to \mathbb{R}$, so that

$$b(x) = \begin{cases} 
1 & \text{if } x \in \overline{B(1)} \\
0 & \text{if } x \in \mathbb{R}^n - B(2).
\end{cases}$$

*See Figures 9.1 and 9.2.*
Proof. There are many ways to construct such a function. We can proceed as follows: Consider the function $h: \mathbb{R} \to \mathbb{R}$ given by

$$h(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

It is easy to show that $h$ is $C^\infty$ (but not analytic!). Then, define $b: \mathbb{R}^n \to \mathbb{R}$, by

$$b(x_1, \ldots, x_n) = \frac{h((4 - x_1^2 - \cdots - x_n^2)/3)}{h((4 - x_1^2 - \cdots - x_n^2)/3) + h((x_1^2 + \cdots + x_n^2 - 1)/3)},$$
It is immediately verified that $b$ satisfies the required conditions.

**Remark:** The function obtained by omitting the factor $1/3$ also yields a smooth bump function, but it looks a little different; its cross-section by a plane through the $x_{n+1}$-axis has four inflection points instead of two. See Figures 9.4 and 9.5.

Given a topological space $X$, for any function $f: X \to \mathbb{R}$, the *support* of $f$, denoted $\text{supp} \ f$, is the closed set

$$\text{supp} \ f = \{x \in X \mid f(x) \neq 0\}.$$ 

Proposition 9.1 yields the following useful technical result:

**Proposition 9.2.** Let $M$ be a smooth manifold. For any open subset $U \subseteq M$, any $p \in U$ and any smooth function $f: U \to \mathbb{R}$, there exist an open subset $V$ with $p \in V$ and a smooth function $\tilde{f}: M \to \mathbb{R}$ defined on the whole of $M$, so that $\bar{V}$ is compact,

$$\bar{V} \subseteq U, \quad \text{supp} \ \tilde{f} \subseteq U,$$

and

$$\tilde{f}(q) = f(q), \quad \text{for all} \quad q \in \bar{V}. $$
Proof. Using a scaling function, it is easy to find a chart \((W, \varphi)\) at \(p\) so that \(W \subseteq U\), \(B(3) \subseteq \varphi(W)\), and \(\varphi(p) = 0\). Let \(\tilde{b} = b \circ \varphi\), where \(b\) is the function given by Proposition 9.1. Then, \(\tilde{b}\) is a smooth function on \(W\) with support \(\varphi^{-1}(B(2)) \subseteq W\). We can extend \(\tilde{b}\) outside \(W\), by setting it to be 0, and we get a smooth function on the whole \(M\). If we let \(V = \varphi^{-1}(B(1))\), then \(V\) is an open subset around \(p\), \(\overline{V} = \varphi^{-1}(\overline{B(1)}) \subseteq W\) is compact, and clearly, \(\tilde{b} = 1\) on \(\overline{V}\). Therefore, if we set

\[
\tilde{f}(q) = \begin{cases} 
\tilde{b}(q)f(q) & \text{if } q \in W \\
0 & \text{if } q \in M - W,
\end{cases}
\]

we see that \(\tilde{f}\) satisfies the required properties. \(\square\)

**Definition 9.1.** If \(X\) is a (Hausdorff) topological space, a family \(\{U_\alpha\}_{\alpha \in I}\) of subsets \(U_\alpha\) of \(X\) is a cover (or covering) of \(X\) iff \(X = \bigcup_{\alpha \in I} U_\alpha\). A cover \(\{U_\alpha\}_{\alpha \in I}\) such that each \(U_\alpha\) is open is an open cover. If \(\{U_\alpha\}_{\alpha \in I}\) is a cover of \(X\), for any subset \(J \subseteq I\), the subfamily \(\{U_\alpha\}_{\alpha \in J}\) is a subcover of \(\{U_\alpha\}_{\alpha \in I}\) if \(X = \bigcup_{\alpha \in J} U_\alpha\), i.e., \(\{U_\alpha\}_{\alpha \in J}\) is still a cover of \(X\). Given a cover \(\{V_\beta\}_{\beta \in J}\), we say that a family \(\{U_\alpha\}_{\alpha \in I}\) is a refinement of \(\{V_\beta\}_{\beta \in J}\) if it is a cover and if there is a function \(h : I \to J\) so that \(U_\alpha \subseteq V_{h(\alpha)}\), for all \(\alpha \in I\).

**Definition 9.2.** A family \(\{U_\alpha\}_{\alpha \in I}\) of subsets of \(X\) is locally finite iff for every point \(p \in X\), there is some open subset \(U\) with \(p \in U\), so that \(U \cap U_\alpha \neq \emptyset\) for only finitely many \(\alpha \in I\). A space \(X\) is paracompact iff every open cover has an open locally finite refinement.
Remark: Recall that a space $X$ is compact iff it is Hausdorff and if every open cover has a finite subcover. Thus, the notion of paracompactness (due to Jean Dieudonné) is a generalization of the notion of compactness.

Definition 9.3. A topological space $X$ is second-countable if it has a countable basis; that is, if there is a countable family of open subsets $\{U_i\}_{i \geq 1}$, so that every open subset of $X$ is the union of some of the $U_i$’s. A topological space $X$ if locally compact iff it is Hausdorff, and for every $a \in X$, there is some compact subset $K$ and some open subset $U$, with $a \in U$ and $U \subseteq K$.

As we will see shortly, every locally compact and second-countable topological space is paracompact.

It is important to observe that every manifold (even not second-countable) is locally compact. Indeed, for every $p \in M$, if we pick a chart $(U, \varphi)$ around $p$, then $\varphi(U) = \Omega$ for some open $\Omega \subseteq \mathbb{R}^n$ ($n = \dim M$). So, we can pick a small closed ball $\overline{B(q, \epsilon)} \subseteq \Omega$ of center $q = \varphi(p)$ and radius $\epsilon$, and as $\varphi$ is a homeomorphism, we see that 

$$p \in \varphi^{-1}(B(q, \epsilon/2)) \subseteq \varphi^{-1}(\overline{B(q, \epsilon)}),$$

where $\varphi^{-1}(\overline{B(q, \epsilon)})$ is compact and $\varphi^{-1}(B(q, \epsilon/2))$ is open.

Finally we define partitions of unity.

Definition 9.4. Let $M$ be a (smooth) manifold. A partition of unity on $M$ is a family $\{f_i\}_{i \in I}$ of smooth functions on $M$ (the index set $I$ may be uncountable), such that:

(a) The family of supports $\{\text{supp } f_i\}_{i \in I}$ is locally finite.

(b) For all $i \in I$ and all $p \in M$, we have $0 \leq f_i(p) \leq 1$, and 

$$\sum_{i \in I} f_i(p) = 1, \text{ for every } p \in M.$$ 

Note that condition (b) implies that for every $p \in M$, there must be some $i \in I$ such that $f_i(p) > 0$. Thus, $\{\text{supp } f_i\}_{i \in I}$ is a cover of $M$. If $\{U_a\}_{a \in J}$ is a cover of $M$, we say that the partition of unity $\{f_i\}_{i \in I}$ is subordinate to the cover $\{U_a\}_{a \in J}$ if $\{\text{supp } f_i\}_{i \in I}$ is a refinement of $\{U_a\}_{a \in J}$. When $I = J$ and $\text{supp } f_i \subseteq U_i$, we say that $\{f_i\}_{i \in I}$ is subordinate to $\{U_a\}_{a \in I}$ with the same index set as the partition of unity.

In Definition 9.4, by (a), for every $p \in M$, there is some open set $U$ with $p \in U$, and $U$ meets only finitely many of the supports $\text{supp } f_i$. So, $f_i(p) \neq 0$ for only finitely many $i \in I$, and the infinite sum $\sum_{i \in I} f_i(p)$ is well defined.

Proposition 9.3. Let $X$ be a topological space which is second-countable and locally compact (thus, also Hausdorff). Then, $X$ is paracompact. Moreover, every open cover has a countable, locally finite refinement consisting of open sets with compact closures.
Proof. The proof is quite technical, but since this is an important result, we reproduce Warner’s proof for the reader’s convenience (Warner [175], Lemma 1.9).

The first step is to construct a sequence of open sets $G_i$, such that

1. $G_i$ is compact,
2. $G_i \subseteq G_{i+1}$,
3. $X = \bigcup_{i=1}^{\infty} G_i$.

As $M$ is second-countable, there is a countable basis of open sets $\{U_i\}_{i \geq 1}$ for $M$. Since $M$ is locally compact, we can find a subfamily of $\{U_i\}_{i \geq 1}$ consisting of open sets with compact closures such that this subfamily is also a basis of $M$. Therefore, we may assume that we start with a countable basis $\{U_i\}_{i \geq 1}$ of open sets with compact closures. Set $G_1 = U_1$, and assume inductively that

$$G_k = U_1 \cup \cdots \cup U_{j_k}.$$ 

Since $G_k$ is compact, it is covered by finitely many of the $U_j$’s. So, let $j_{k+1}$ be the smallest integer greater than $j_k$ so that

$$G_k \subseteq U_1 \cup \cdots \cup U_{j_{k+1}},$$ 

and set

$$G_{k+1} = U_1 \cup \cdots \cup U_{j_{k+1}}.$$ 

Obviously, the family $\{G_i\}_{i \geq 1}$ satisfies (1)–(3).

Now, let $\{U_\alpha\}_{\alpha \in I}$ be an arbitrary open cover of $M$. For any $i \geq 3$, the set $G_{i+1} - G_{i-1}$ is compact and contained in the open $G_{i+1} - G_{i-2}$. For each $i \geq 3$, choose a finite subcover of the open cover $\{U_\alpha \cap (G_{i+1} - G_{i-2})\}_{\alpha \in I}$ of $G_{i+1} - G_{i-1}$, and choose a finite subcover of the open cover $\{U_\alpha \cap G_3\}_{\alpha \in I}$ of the compact set $G_2$. We leave it to the reader to check that this family of open sets is indeed a countable, locally finite refinement of the original open cover $\{U_\alpha\}_{\alpha \in I}$ and consists of open sets with compact closures.

Remarks:

1. Proposition 9.3 implies that a second-countable, locally compact (Hausdorff) topological space is the union of countably many compact subsets. Thus, $X$ is countable at infinity, a notion that we already encountered in Proposition 5.11 and Theorem 5.14. The reason for this odd terminology is that in the Alexandroff one-point compactification of $X$, the family of open subsets containing the point at infinity (ω) has a countable basis of open sets. (The open subsets containing $\omega$ are of the form $(M - K) \cup \{\omega\}$, where $K$ is compact.)
2. A manifold that is countable at infinity has a countable open cover by domains of charts. This is because, if \( M = \bigcup_{i \geq 1} K_i \), where the \( K_i \subseteq M \) are compact, then for any open cover of \( M \) by domains of charts, for every \( K_i \), we can extract a finite subcover, and the union of these finite subcovers is a countable open cover of \( M \) by domains of charts. But then, since for every chart \((U_i, \varphi_i)\), the map \( \varphi_i \) is a homeomorphism onto some open subset of \( \mathbb{R}^n \), which is second-countable, so we deduce easily that \( M \) is second-countable. Thus, for manifolds, second-countable is equivalent to countable at infinity.

We can now prove the main theorem stating the existence of partitions of unity. Recall that we are assuming that our manifolds are Hausdorff and second-countable.

**Theorem 9.4.** Let \( M \) be a smooth manifold and let \( \{U_\alpha\}_{\alpha \in I} \) be an open cover for \( M \). Then, there is a countable partition of unity \( \{f_i\}_{i \geq 1} \) subordinate to the cover \( \{U_\alpha\}_{\alpha \in I} \), and the support \( \text{supp } f_i \) of each \( f_i \) is compact. If one does not require compact supports, then there is a partition of unity \( \{f_\alpha\}_{\alpha \in I} \) subordinate to the cover \( \{U_\alpha\}_{\alpha \in I} \) with at most countably many of the \( f_\alpha \) not identically zero. (In the second case, \( \text{supp } f_\alpha \subseteq U_\alpha \).)

**Proof.** Again, we reproduce Warner’s proof (Warner [175], Theorem 1.11). As our manifolds are second-countable, Hausdorff and locally compact, from the proof of Proposition 9.3, we have the sequence of open subsets \( \{G_i\}_{i \geq 1} \), and we set \( G_0 = \emptyset \). For any \( p \in M \), let \( i_p \) be the largest integer such that \( p \in M - \overline{G_{i_p}} \). Choose an \( \alpha_p \) such that \( p \in U_{\alpha_p} \); we can find a chart \((U, \varphi)\) centered at \( p \) such that \( U \subseteq U_{\alpha_p} \cap (G_{i_p+2} - \overline{G_{i_p}}) \) and such that \( B(2) \subseteq \varphi(U) \). Define

\[
\psi_p = \begin{cases} 
  b \circ \varphi & \text{on } U \\
  0 & \text{on } M - U,
\end{cases}
\]

where \( b \) is the bump function defined just before Proposition 9.1. Then, \( \psi_p \) is a smooth function on \( M \) which has value 1 on some open subset \( W_p \) containing \( p \) and has compact support lying in \( U \subseteq U_{\alpha_p} \cap (G_{i_p+2} - \overline{G_{i_p}}) \). For each \( i \geq 1 \), choose a finite set of points \( p \in M \), whose corresponding open \( W_p \) cover \( \overline{G_i} - G_{i-1} \). Order the corresponding \( \psi_p \) functions in a sequence \( \psi_j, j = 1, 2, \ldots \). The supports of the \( \psi_j \) form a locally finite family of subsets of \( M \). Thus, the function

\[
\psi = \sum_{j=1}^{\infty} \psi_j
\]

is well-defined on \( M \) and smooth. Moreover, \( \psi(p) > 0 \) for each \( p \in M \). For each \( i \geq 1 \), set

\[
f_i = \frac{\psi_i}{\psi}
\]

Then, the family \( \{f_i\}_{i \geq 1} \) is a partition of unity subordinate to the cover \( \{U_\alpha\}_{\alpha \in I} \), and \( \text{supp } f_i \) is compact for all \( i \geq 1 \). Now, when we don’t require compact support, if we let \( f_\alpha \) be identically zero if no \( f_i \) has support in \( U_\alpha \) and otherwise let \( f_\alpha \) be the sum of the \( f_i \).
with support in $U_\alpha$, then we obtain a partition of unity subordinate to $\{U_\alpha\}_{\alpha \in I}$ with at most countably many of the $f_\alpha$ not identically zero. We must have $\text{supp} f_\alpha \subseteq U_\alpha$, because for any locally finite family of closed sets $\{F_\beta\}_{\beta \in J}$, we have $\bigcup_{\beta \in J} F_\beta = \bigcup_{\beta \in J} F_\beta$.

We close this section by stating a famous theorem of Whitney whose proof uses partitions of unity.

**Theorem 9.5.** (Whitney, 1935) Any smooth manifold (Hausdorff and second-countable) $M$ of dimension $n$ is diffeomorphic to a closed submanifold of $\mathbb{R}^{2n+1}$.

For a proof, see Hirsch [91], Chapter 2, Section 2, Theorem 2.14.

## 9.2 Covering Maps and Universal Covering Manifolds

Covering maps are an important technical tool in algebraic topology, and more generally in geometry. This brief section only gives some basic definitions and states a few major facts. Appendix A of O’Neill [138] gives a review of definitions and main results about covering manifolds. Expositions including full details can be found in Hatcher [85], Greenberg [79], Munkres [134], Fulton [69], and Massey [122, 123] (the most extensive).

We begin with covering maps.

**Definition 9.5.** A map $\pi: M \to N$ between two smooth manifolds is a covering map (or cover) iff

1. The map $\pi$ is smooth and surjective.

2. For any $q \in N$, there is some open subset $V \subseteq N$ so that $q \in V$ and

   $$\pi^{-1}(V) = \bigcup_{i \in I} U_i,$$

   where the $U_i$ are pairwise disjoint open subsets $U_i \subseteq M$, and $\pi: U_i \to V$ is a diffeomorphism for every $i \in I$. We say that $V$ is evenly covered.

The manifold $M$ is called a covering manifold of $N$. See Figure 9.6.

A homomorphism of coverings $\pi_1: M_1 \to N$ and $\pi_2: M_2 \to N$ is a smooth map $\phi: M_1 \to M_2$, so that

$$\pi_1 = \pi_2 \circ \phi;$$

that is, the following diagram commutes:

$$\begin{array}{ccc}
M_1 & \xrightarrow{\phi} & M_2 \\
\downarrow{\pi_1} & & \downarrow{\pi_2} \\
N & & 
\end{array}$$

We say that the coverings $\pi_1: M_1 \to N$ and $\pi_2: M_2 \to N$ are equivalent iff there is a homomorphism $\phi: M_1 \to M_2$ between the two coverings, and $\phi$ is a diffeomorphism.
Figure 9.6: Two examples of a covering map. The left illustration is $\pi: \mathbb{R} \to S^1$ with $\pi(t) = (\cos(2\pi t), \sin(2\pi t))$, while the right illustration is the 2-fold antipodal covering of $\mathbb{R}P^2$ by $S^2$.

As usual, the inverse image $\pi^{-1}(q)$ of any element $q \in N$ is called the fibre over $q$, the space $N$ is called the base, and $M$ is called the covering space. As $\pi$ is a covering map, each fibre is a discrete space. Note that a homomorphism maps each fibre $\pi^{-1}_1(q)$ in $M_1$ to the fibre $\pi^{-1}_2(\phi(q))$ in $M_2$, for every $q \in M_1$.

**Proposition 9.6.** Let $\pi: M \to N$ be a covering map. If $N$ is connected, then all fibres $\pi^{-1}(q)$ have the same cardinality for all $q \in N$. Furthermore, if $\pi^{-1}(q)$ is not finite then it is countably infinite.

**Proof.** Pick any point, $p \in N$. We claim that the set

$$S = \{ q \in N \mid |\pi^{-1}(q)| = |\pi^{-1}(p)| \}$$

is open and closed.

If $q \in S$, then there is some open subset $V$ with $q \in V$, so that $\pi^{-1}(V)$ is evenly covered by some family $\{U_i\}_{i \in I}$ of disjoint open subsets $U_i$, each diffeomorphic to $V$ under $\pi$. Then, every $s \in V$ must have a unique preimage in each $U_i$, so

$$|I| = |\pi^{-1}(s)|, \quad \text{for all } s \in V.$$

However, as $q \in S$, $|\pi^{-1}(q)| = |\pi^{-1}(p)|$, so

$$|I| = |\pi^{-1}(p)| = |\pi^{-1}(s)|, \quad \text{for all } s \in V,$$
and thus, $V \subseteq S$. Therefore, $S$ is open. Similarly the complement of $S$ is open. As $N$ is connected, $S = N$.

Since $M$ is a manifold, it is second-countable, that is every open subset can be written as some countable union of open subsets. But then, every family $\{U_i\}_{i \in I}$ of pairwise disjoint open subsets forming an even cover must be countable, and since $|I|$ is the common cardinality of all the fibres, every fibre is countable.

When the common cardinality of fibres is finite, it is called the multiplicity of the covering (or the number of sheets).

For any integer, $n > 0$, the map $z \mapsto z^n$ from the unit circle $S^1 = U(1)$ to itself is a covering with $n$ sheets. The map,

$$t: \mapsto (\cos(2\pi t), \sin(2\pi t)),$$

is a covering $\mathbb{R} \rightarrow S^1$, with infinitely many sheets.

It is also useful to note that a covering map $\pi: M \rightarrow N$ is a local diffeomorphism (which means that $d\pi_p: T_pM \rightarrow T_{\pi(p)}N$ is a bijective linear map for every $p \in M$). Indeed, given any $p \in M$, if $q = \pi(p)$, then there is some open subset $V \subseteq N$ containing $q$ so that $V$ is evenly covered by a family of disjoint open subsets $\{U_i\}_{i \in I}$, with each $U_i \subseteq M$ diffeomorphic to $V$ under $\pi$. As $p \in U_i$ for some $i$, we have a diffeomorphism $\pi \upharpoonright U_i: U_i \longrightarrow V$, as required.

The crucial property of covering manifolds is that curves in $N$ can be lifted to $M$, in a unique way. For any map $\phi: P \rightarrow N$, a lift of $\phi$ through $\pi$ is a map $\tilde{\phi}: P \rightarrow M$ so that

$$\phi = \pi \circ \tilde{\phi},$$

as in the following commutative diagram:

$$\begin{array}{ccc}
\tilde{\phi} & \rightarrow & M \\
\downarrow & & \downarrow \pi \\
P & \phi \rightarrow & N
\end{array}$$

We would like to state three propositions regarding covering spaces. However, two of these propositions use the notion of a simply connected manifold. Intuitively, a manifold is simply connected if it has no “holes.” More precisely, a manifold is simply connected if it has a trivial fundamental group. Those readers familiar with the fundamental group may proceed directly to Proposition 9.11 as we now provide a brief review of the fundamental group construction based on Sections 5.1 and 5.2 of Armstrong [5].

A fundamental group is a homotopic loop group. Therefore, given topological spaces $X$ and $Y$, we need to define a homotopy between two continuous functions $f: X \rightarrow Y$ and $g: X \rightarrow Y$. 
Definition 9.6. Let $X$ and $Y$ be topological spaces, $f : X \to Y$ and $g : X \to Y$ be two continuous functions, and let $I = [0, 1]$. We say that $f$ is homotopic to $g$ if there exists a continuous function $F : X \times I \to Y$ (where $X \times I$ is given the product topology) such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$ for all $x \in X$. The map $F$ is a homotopy from $f$ to $g$, and this is denoted $f \sim_F g$. If $f$ and $g$ agree on $A \subseteq X$, i.e. $f(a) = g(a)$ whenever $a \in A$, we say $f$ is homotopic to $g$ relative to $A$, and this is denoted $f \sim_F g$ rel $A$.

A homotopy provides a means of continuously deforming $f$ into $g$ through a family $\{f_t\}$ of continuous functions $f_t : X \to Y$ where $t \in [0, 1]$ and $f_0(x) = f(x)$ and $f_1(x) = g(x)$ for all $x \in X$. For example, let $D$ be the unit disk in $\mathbb{R}^2$ and consider two continuous functions $f : I \to D$ and $g : I \to D$. Then $f \sim_F g$ via the straight line homotopy $F : I \times I \to D$, where $F(x, t) = (1 - t)f(x) + tg(x)$.

Proposition 9.7. Let $X$ and $Y$ be topological spaces and let $A \subseteq X$. Homotopy (or homotopy rel $A$) is an equivalence relation on the set of all continuous functions from $X$ to $Y$.

The next two propositions show that homotopy behaves well with respect to composition.

Proposition 9.8. Let $X$, $Y$, and $Z$ be topological spaces and let $A \subseteq X$. For any continuous functions $f : X \to Y$, $g : X \to Y$, and $h : Y \to Z$, if $f \sim_F g$ rel $A$, then $h \circ f \sim_{hF} h \circ g$ rel $A$ as maps from $X$ to $Z$.

\[
\begin{array}{ccc}
X & \xrightarrow{f} & Y \\
\downarrow{g} & & \downarrow{h} \\
\end{array}
\xrightarrow{h \circ f \sim_{hF} h \circ g}
\begin{array}{c}
Z
\end{array}
\]

Proposition 9.9. Let $X$, $Y$, and $Z$ be topological spaces and let $B \subseteq Y$. For any continuous functions $f : X \to Y$, $g : Y \to Z$, and $h : Y \to Z$, if $g \sim_G h$ rel $B$, then $g \circ f \sim_F h \circ f$ rel $f^{-1}B$, where $F(x, t) = G(f(x), t)$.

\[
\begin{array}{ccc}
X & \xrightarrow{f} & Y \\
\downarrow{g} & & \downarrow{h} \\
\end{array}
\xrightarrow{g \circ f \sim_F h \circ f}
\begin{array}{c}
Z
\end{array}
\]

In order to define the fundamental group of a topological space $X$, we recall the definition of a loop.

Definition 9.7. Let $X$ be a topological space, $p$ be a point in $X$, and let $I = [0, 1]$. We say $\alpha$ is a loop based at $p = \alpha(0)$ if $\alpha$ is a continuous map $\alpha : I \to X$ with $\alpha(0) = \alpha(1)$.

Given a topological space $X$, choose a point $p \in X$ and form $S$, the set of all loops in $X$ based at $p$. By applying Proposition 9.7, we know that the relation of homotopy relative to $\{0, 1\}$ is an equivalence relation on $S$. This leads to the following definition.

Definition 9.8. Let $X$ be a topological space, $p$ be a point in $X$, and let $\alpha$ be a loop in $X$ based at $p$. The set of all loops homotopic to $\alpha$ relative to $\{0, 1\}$ is the homotopy class of $\alpha$ and is denoted $\langle \alpha \rangle$. 

Given two loops $\alpha$ and $\beta$ in $X$ based at $p$, the *product* $\alpha \cdot \beta$ is a loop in $X$ based at $p$ defined by
\[
\alpha \cdot \beta(t) = \begin{cases} 
\alpha(2t), & 0 \leq t \leq \frac{1}{2} \\
\beta(2t - 1), & \frac{1}{2} < t \leq 1.
\end{cases}
\]
The product of loops gives rise to the product of homotopy classes where
\[
\langle \alpha \rangle \cdot \langle \beta \rangle = \langle \alpha \cdot \beta \rangle.
\]
We leave it the reader to check that the multiplication of homotopy classes is well defined and associative, namely $\langle \alpha \cdot \beta \rangle \cdot \langle \gamma \rangle = \langle \alpha \rangle \cdot \langle \beta \cdot \gamma \rangle$ whenever $\alpha$, $\beta$, and $\gamma$ are loops in $X$ based at $p$.

Let $\langle e \rangle$ be the homotopy class of the constant loop in $X$ based at $p$, and define the inverse of $\langle \alpha \rangle$ as $\langle \alpha \rangle^{-1} = \langle \alpha^{-1} \rangle$, where $\alpha^{-1}(t) = \alpha(1 - t)$. With these conventions, the product operation between homotopy classes gives rise to a group. In particular,

**Proposition 9.10.** Let $X$ be a topological space and let $p$ be a point in $X$. The set of homotopy classes of loops in $X$ based at $p$ is a group with multiplication given by $\langle \alpha \rangle \cdot \langle \beta \rangle = \langle \alpha \cdot \beta \rangle$.

**Definition 9.9.** Let $X$ be a topological space and $p$ a point in $X$. The group of homotopy classes of loops in $X$ based at $p$ is the *fundamental group* of $X$ based at $p$, and is denoted by $\pi_1(X, p)$.

If we assume $X$ is path connected, we can show that $\pi_1(X, p) \cong \pi_1(X, q)$ for any points $p$ and $q$ in $X$. Therefore, when $X$ is path connected, we simply write $\pi_1(X)$. If $X$ is path connected and $\pi_1(X) = \langle e \rangle$, (which is also denoted as $\pi_1(X) = (0)$), we say $X$ is *simply connected*. In other words, every loop in $X$ can be shrunk in a continuous manner within $X$ to its basepoint. Examples of simply connected spaces include $\mathbb{R}^n$ and $S^n$ whenever $n \geq 2$. On the other hand, the torus and the circle are not simply connected. See Figures 9.7 and 9.8.

![Figure 9.7: The torus is not simply connected. The loop at $p$ is homotopic to a point, but the loop at $q$ is not.](image)
Figure 9.8: The unit sphere $S^2$ is simply connected since every loop can be continuously deformed to a point. This deformation is represented by the map $F: I \times I \to S^2$ where $F(x, 0) = \alpha$ and $F(x, 1) = p$.

We now state without proof the following results regarding covering spaces:

**Proposition 9.11.** If $\pi: M \to N$ is a covering map, then for every smooth curve $\alpha: I \to N$ in $N$ (with $0 \in I$) and for any point $p \in M$ such that $\pi(p) = \alpha(0)$, there is a unique smooth curve $\tilde{\alpha}: I \to M$ lifting $\alpha$ through $\pi$ such that $\tilde{\alpha}(0) = q$. See Figure 9.9.

**Proposition 9.12.** Let $\pi: M \to N$ be a covering map and let $\phi: P \to N$ be a smooth map. For any $p_0 \in P$, any $q_0 \in M$ and any $r_0 \in N$ with $\pi(q_0) = \phi(p_0) = r_0$, the following properties hold:

1. If $P$ is connected then there is at most one lift $\tilde{\phi}: P \to M$ of $\phi$ through $\pi$ such that $\tilde{\phi}(p_0) = q_0$.

2. If $P$ is simply connected, then such a lift exists.

**Theorem 9.13.** Every connected manifold $M$ possesses a simply connected covering map $\pi: \tilde{M} \to M$; that is, with $\tilde{M}$ simply connected. Any two simply connected coverings of $N$ are equivalent.
In view of Theorem 9.13, it is legitimate to speak of the simply connected cover $\tilde{M}$ of $M$, also called universal covering (or cover) of $M$.

Given any point $p \in M$, let $\pi_1(M, p)$ denote the fundamental group of $M$ with basepoint $p$. See Definition 9.9. If $\phi: M \to N$ is a smooth map, for any $p \in M$, if we write $q = \phi(p)$, then we have an induced group homomorphism

$$\phi_*: \pi_1(M, p) \to \pi_1(N, q).$$

**Proposition 9.14.** If $\pi: M \to N$ is a covering map, for every $p \in M$, if $q = \pi(p)$, then the induced homomorphism $\pi_*: \pi_1(M, p) \to \pi_1(N, q)$ is injective.

The next proposition is a stronger version of part (2) of Proposition 9.12:

**Proposition 9.15.** Let $\pi: M \to N$ be a covering map and let $\phi: P \to N$ be a smooth map. For any $p_0 \in P$, any $q_0 \in M$ and any $r_0 \in N$ with $\pi(q_0) = \phi(p_0) = r_0$, if $P$ is connected, then a lift $\tilde{\phi}: P \to M$ of $\phi$ such that $\tilde{\phi}(p_0) = q_0$ exists iff

$$\phi_* (\pi_1(P, p_0)) \subseteq \pi_* (\pi_1(M, q_0)),$$

as illustrated in the diagram below.

![Diagram](image-url)
Basic Assumption: For any covering $\pi : M \to N$, if $N$ is connected then we also assume that $M$ is connected.

Using Proposition 9.14, we get

**Proposition 9.16.** If $\pi : M \to N$ is a covering map and $N$ is simply connected, then $\pi$ is a diffeomorphism (recall that $M$ is connected); thus, $M$ is diffeomorphic to the universal cover $\tilde{N}$, of $N$.

**Proof.** Pick any $p \in M$ and let $q = \pi(p)$. As $N$ is simply connected, $\pi_1(N,q) = (0)$. By Proposition 9.14, since $\pi_* : \pi_1(M,p) \to \pi_1(N,q)$ is injective, $\pi_1(M,p) = (0)$, so $M$ is simply connected (by hypothesis, $M$ is connected). But then, by Theorem 9.13, $M$ and $N$ are diffeomorphic. \qed

The following proposition shows that the universal covering of a space covers every other covering of that space. This justifies the terminology “universal covering.”

**Proposition 9.17.** Say $\pi_1 : M_1 \to N$ and $\pi_2 : M_2 \to N$ are two coverings of $N$, with $N$ connected. Every homomorphism $\phi : M_1 \to M_2$ between these two coverings is a covering map.

\[ \begin{array}{ccc}
M_1 & \xrightarrow{\phi} & M_2 \\
\downarrow{\pi_1} & & \downarrow{\pi_2} \\
N & & N
\end{array} \]

As a consequence, if $\pi : \tilde{N} \to N$ is a universal covering of $N$, then for every covering $\pi' : M \to N$ of $N$, there is a covering $\phi : \tilde{N} \to M$ of $M$.

The notion of deck-transformation group of a covering is also useful because it yields a way to compute the fundamental group of the base space.

**Definition 9.10.** If $\pi : M \to N$ is a covering map, a deck-transformation is any diffeomorphism $\phi : M \to M$ such that $\pi = \pi \circ \phi$; that is, the following diagram commutes:

\[ \begin{array}{ccc}
M & \xrightarrow{\phi} & M \\
\downarrow{\pi} & & \downarrow{\pi} \\
N & & N
\end{array} \]

Note that deck-transformations are just automorphisms of the covering map. The commutative diagram of Definition 9.10 means that a deck transformation permutes every fibre. It is immediately verified that the set of deck transformations of a covering map is a group under composition denoted $\Gamma_{\pi}$ (or simply $\Gamma$), called the deck-transformation group of the covering.
Observe that any deck transformation $\phi$ is a lift of $\pi$ through $\pi$. Consequently, if $M$ is connected, by Proposition 9.12 (1), every deck-transformation is determined by its value at a single point. So, the deck-transformations are determined by their action on each point of any fixed fibre $\pi^{-1}(q)$, with $q \in N$. Since the fibre $\pi^{-1}(q)$ is countable, $\Gamma$ is also countable, that is, a discrete Lie group. Moreover, if $M$ is compact, as each fibre $\pi^{-1}(q)$ is compact and discrete, it must be finite and so, the deck-transformation group is also finite.

The following proposition gives a useful method for determining the fundamental group of a manifold.

**Proposition 9.18.** If $\pi: \tilde{M} \to M$ is the universal covering of a connected manifold $M$, then the deck-transformation group $\tilde{\Gamma}$ is isomorphic to the fundamental group $\pi_1(M)$ of $M$.

**Remark:** When $\pi: \tilde{M} \to M$ is the universal covering of $M$, it can be shown that the group $\tilde{\Gamma}$ acts simply and transitively on every fibre $\pi^{-1}(q)$. This means that for any two elements $x, y \in \pi^{-1}(q)$, there is a unique deck-transformation $\phi \in \tilde{\Gamma}$ such that $\phi(x) = y$. So, there is a bijection between $\pi_1(M) \cong \tilde{\Gamma}$ and the fibre $\pi^{-1}(q)$.

Proposition 9.13 together with previous observations implies that if the universal cover of a connected (compact) manifold is compact, then $M$ has a finite fundamental group. We will use this fact later, in particular in the proof of Myers’ Theorem.
Chapter 10

Riemannian Metrics, Riemannian Manifolds

Fortunately, the rich theory of vector spaces endowed with a Euclidean inner product can, to a great extent, be lifted to the tangent bundle of a manifold. The idea is to equip the tangent space $T_pM$ at $p$ to the manifold $M$ with an inner product $\langle -, - \rangle_p$, in such a way that these inner products vary smoothly as $p$ varies on $M$. It is then possible to define the length of a curve segment on a $M$ and to define the distance between two points on $M$.

In Section 10.1, we define the notion of local (and global) frame. Using frames, we obtain a criterion for the tangent bundle $TM$ of a smooth manifold $M$ to be trivial (that is, isomorphic to $M \times \mathbb{R}^n$).

Riemannian metrics and Riemannian manifolds are defined in Section 10.2, where several examples are given. The generalization of the notion of the gradient of a function defined on a smooth manifold requires a metric. We define the gradient of a function on a Riemannian manifold. We conclude by defining local isometries, isometries, and the isometry group $\text{Isom}(M,g)$ of a Riemannian manifold $(M,g)$.

10.1 Frames

Definition 10.1. Let $M$ be an $n$-dimensional smooth manifold. For any open subset $U \subseteq M$, an $n$-tuple of vector fields $(X_1, \ldots, X_n)$ over $U$ is called a frame over $U$ iff $(X_1(p), \ldots, X_n(p))$ is a basis of the tangent space $T_pM$, for every $p \in U$. If $U = M$, then the $X_i$ are global sections and $(X_1, \ldots, X_n)$ is called a frame (of $M$).

The notion of a frame is due to Élie Cartan who (after Darboux) made extensive use of them under the name of moving frame (and the moving frame method). Cartan’s terminology is intuitively clear: As a point $p$ moves in $U$, the frame $(X_1(p), \ldots, X_n(p))$ moves from fibre to fibre. Physicists refer to a frame as a choice of local gauge.
If \( \dim(M) = n \), then for every chart \((U, \varphi)\), since \( d\varphi^{-1}_{\varphi(p)} : \mathbb{R}^n \to T_p M \) is a bijection for every \( p \in U \), the \( n \)-tuple of vector fields \((X_1, \ldots, X_n)\), with \( X_i(p) = d\varphi^{-1}_{\varphi(p)}(e_i) \), is a frame of \( TM \) over \( U \), where \((e_1, \ldots, e_n)\) is the canonical basis of \( \mathbb{R}^n \). See Figure 10.1.

Figure 10.1: A frame on \( S^2 \)

The following proposition tells us when the tangent bundle is trivial (that is, isomorphic to the product \( M \times \mathbb{R}^n \)).

**Proposition 10.1.** The tangent bundle \( TM \) of a smooth \( n \)-dimensional manifold \( M \) is trivial iff it possesses a frame of global sections (vector fields defined on \( M \)).

Proposition 10.1 is a special case of Proposition 28.8 which holds for vector bundles and is proved in Chapter 28.

As an illustration of Proposition 10.1 we can prove that the tangent bundle \( TS^1 \) of the circle is trivial. Indeed, we can find a section that is everywhere nonzero, i.e. a non-vanishing vector field, namely

\[ X(\cos \theta, \sin \theta) = (-\sin \theta, \cos \theta). \]
The reader should try proving that $TS^3$ is also trivial (use the quaternions).

However, $TS^2$ is nontrivial, although this not so easy to prove. More generally, it can be shown that $TS^n$ is nontrivial for all even $n \geq 2$. It can even be shown that $S^1$, $S^3$ and $S^7$ are the only spheres whose tangent bundle is trivial. This is a deep theorem and its proof is hard.

**Remark:** A manifold $M$ such that its tangent bundle $TM$ is trivial is called parallelizable.

We now define Riemannian metrics and Riemannian manifolds.

### 10.2 Riemannian Metrics

**Definition 10.2.** Given a smooth $n$-dimensional manifold $M$, a Riemannian metric on $M$ (or $TM$) is a family $(\langle - , - \rangle_p)_{p \in M}$ of inner products on each tangent space $T_pM$, such that $\langle - , - \rangle_p$ depends smoothly on $p$, which means that for every chart $\varphi_\alpha : U_\alpha \to \mathbb{R}^n$, for every frame $(X_1, \ldots, X_n)$ on $U_\alpha$, the maps

$$ p \mapsto \langle X_i(p), X_j(p) \rangle_p, \quad p \in U_\alpha, \; 1 \leq i, j \leq n, $$

are smooth. A smooth manifold $M$ with a Riemannian metric is called a Riemannian manifold.

If $\dim(M) = n$, then for every chart $(U, \varphi)$, we have the frame $(X_1, \ldots, X_n)$ over $U$, with $X_i(p) = d\varphi^{-1}_\varphi(p)(e_i)$, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. Since every vector field over $U$ is a linear combination $\sum_{i=1}^n f_i X_i$, for some smooth functions $f_i : U \to \mathbb{R}$, the condition of Definition 10.2 is equivalent to the fact that the maps

$$ p \mapsto \langle d\varphi^{-1}_\varphi(p)(e_i), d\varphi^{-1}_\varphi(p)(e_j) \rangle_p, \quad p \in U, \; 1 \leq i, j \leq n, $$

are smooth. If we let $x = \varphi(p)$, the above condition says that the maps

$$ x \mapsto \langle d\varphi^{-1}_x(e_i), d\varphi^{-1}_x(e_j) \rangle_{\varphi^{-1}(x)}, \quad x \in \varphi(U), \; 1 \leq i, j \leq n, $$

are smooth.

If $M$ is a Riemannian manifold, the metric on $TM$ is often denoted $g = (g_p)_{p \in M}$. In a chart, using local coordinates, we often use the notation $g = \sum_{ij} g_{ij} dx_i \otimes dx_j$, or simply $g = \sum_{ij} g_{ij} dx_i dx_j$, where

$$ g_{ij}(p) = \left\langle \left( \frac{\partial}{\partial x_i} \right)_p, \left( \frac{\partial}{\partial x_j} \right)_p \right\rangle_\varphi. $$

For every $p \in U$, the matrix $(g_{ij}(p))$ is symmetric, positive definite.

The standard Euclidean metric on $\mathbb{R}^n$, namely

$$ g = dx_1^2 + \cdots + dx_n^2, $$

is a Riemannian metric.
makes $\mathbb{R}^n$ into a Riemannian manifold. Then, every submanifold $M$ of $\mathbb{R}^n$ inherits a metric by restricting the Euclidean metric to $M$.

For example, the sphere $S^{n-1}$ inherits a metric that makes $S^{n-1}$ into a Riemannian manifold. It is instructive to find the local expression of this metric for $S^2$ in spherical coordinates. We can parametrize the sphere $S^2$ in terms of two angles $\theta$ (the colatitude) and $\varphi$ (the longitude) as follows:

$$
x = \sin \theta \cos \varphi \\
y = \sin \theta \sin \varphi \\
z = \cos \theta.
$$

See Figure 10.2.

![Figure 10.2: The spherical coordinates of $S^2$](image)

In order for the above to be a parametrization, we need to restrict its domain to $V = \{ (\theta, \varphi) \mid 0 < \theta < \pi, 0 < \varphi < 2\pi \}$. Then, the semicircle from the north pole to the south pole lying in the $xz$-plane is omitted from the sphere. In order to cover the whole sphere, we need another parametrization obtained by choosing the axes in a suitable fashion; for example, to omit the semicircle in the $xy$-plane from $(0, 1, 0)$ to $(0, -1, 0)$ and with $x \leq 0$.

To compute the matrix giving the Riemannian metric in this chart, we need to compute a basis $(u(\theta, \varphi), v(\theta, \varphi))$ of the tangent plane $T_pS^2$ at $p = (\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta)$. We can use

$$
u(\theta, \varphi) = \frac{\partial p}{\partial \theta} = (\cos \theta \cos \varphi, \cos \theta \sin \varphi, -\sin \theta)$$
$$
u(\theta, \varphi) = \frac{\partial p}{\partial \varphi} = (-\sin \theta \sin \varphi, \sin \theta \cos \varphi, 0),$$

...
and we find that
\[
\langle u(\theta, \varphi), u(\theta, \varphi) \rangle = 1 \\
\langle u(\theta, \varphi), v(\theta, \varphi) \rangle = 0 \\
\langle v(\theta, \varphi), v(\theta, \varphi) \rangle = \sin^2 \theta,
\]
so the metric on \( T_p S^2 \) w.r.t. the basis \((u(\theta, \varphi), v(\theta, \varphi))\) is given by the matrix
\[
g_p = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2 \theta \end{pmatrix}.
\]
Thus, for any tangent vector
\[
w = au(\theta, \varphi) + bv(\theta, \varphi), \quad a, b \in \mathbb{R},
\]
we have
\[
g_p(w, w) = a^2 + \sin^2 \theta b^2.
\]

A nontrivial example of a Riemannian manifold is the Poincaré upper half-space, namely, the set \( H = \{(x, y) \in \mathbb{R}^2 \mid y > 0\} \) equipped with the metric
\[
g = \frac{dx^2 + dy^2}{y^2}.
\]

Consider the Lie group \( \text{SO}(n) \). We know from Section 7.2 that its tangent space at the identity \( T_I \text{SO}(n) \) is the vector space \( \mathfrak{so}(n) \) of \( n \times n \) skew symmetric matrices, and that the tangent space \( T_Q \text{SO}(n) \) to \( \text{SO}(n) \) at \( Q \) is isomorphic to
\[
Q\mathfrak{so}(n) = \{QB \mid B \in \mathfrak{so}(n)\}.
\]
(It is also isomorphic to \( \mathfrak{so}(n)Q = \{BQ \mid B \in \mathfrak{so}(n)\} \).) If we give \( \mathfrak{so}(n) \) the inner product
\[
\langle B_1, B_2 \rangle = \text{tr}(B_1^\top B_2) = -\text{tr}(B_1 B_2),
\]
the inner product on \( T_Q \text{SO}(n) \) is given by
\[
\langle QB_1, QB_2 \rangle = \text{tr}((QB_1)^\top QB_2) = \text{tr}(B_1^\top Q^\top QB_2) = \text{tr}(B_1^\top B_2).\]
We will see in Chapter 12 that the length \( L(\gamma) \) of the curve segment \( \gamma \) from \( I \) to \( e^B \) given by \( t \mapsto e^{tB} \) (with \( B \in \mathfrak{so}(n) \)) is given by
\[
L(\gamma) = \left( \text{tr}(-B^2) \right)^{\frac{1}{2}}.
\]

More generally, given any Lie group \( G \), any inner product \( \langle -, - \rangle \) on its Lie algebra \( \mathfrak{g} \) induces by left translation an inner product \( \langle -, - \rangle_g \) on \( T_g G \) for every \( g \in G \), and this yields a Riemannian metric on \( G \) (which happens to be left-invariant; see Chapter 17).
Going back to the second example of Section 7.5, where we computed the differential $df_R$ of the function $f : \text{SO}(3) \to \mathbb{R}$ given by

$$f(R) = (u^\top Rv)^2, \quad u, v \in \mathbb{R}^3$$

we found that

$$df_R(X) = 2u^\top Xvu^\top Rv, \quad X \in R\text{so}(3).$$

Since each tangent space $T_R\text{SO}(3)$ is a Euclidean space under the inner product defined above, by duality (see Proposition 21.1 applied to the pairing $\langle -, - \rangle$), there is a unique vector $Y \in T_R\text{SO}(3)$ defining the linear form $df_R$; that is,

$$\langle Y, X \rangle = df_R(X), \quad \text{for all } X \in T_R\text{SO}(3).$$

By definition, the vector $Y$ is the gradient of $f$ at $R$, denoted $(\text{grad}(f))_R$. The gradient of $f$ at $R$ is given by

$$(\text{grad}(f))_R = u^\top Rv(R^\top uv^\top - vu^\top R)$$

since

$$\langle (\text{grad}(f))_R, X \rangle = \text{tr}((\text{grad}(f))_R^\top X)$$

$$= u^\top Rv \text{tr}((R^\top uv^\top - uu^\top R)^\top R^\top X)$$

$$= u^\top Rv \text{tr}((vu^\top R - R^\top uv^\top)^\top R^\top X)$$

$$= u^\top Rv(\text{tr}(vu^\top X) - \text{tr}(R^\top uv^\top R^\top X)), \quad \text{since } RR^\top = I$$

$$= u^\top Rv(\text{tr}(u^\top Xv) - \text{tr}(R^\top uv^\top R^\top X))$$

$$= u^\top Rv(\text{tr}(u^\top Xv) - \text{tr}(R^\top uv^\top RB)), \quad X = RB \text{ with } B^\top = -B$$

$$= u^\top Rv(\text{tr}(u^\top Xv) - \text{tr}(R^\top uv^\top B))$$

$$= u^\top Rv(\text{tr}(u^\top Xv) - \text{tr}((R^\top uv^\top B)^\top))$$

$$= u^\top Rv(\text{tr}(u^\top Xv) + \text{tr}(Bvu^\top R))$$

$$= u^\top Rv(\text{tr}(u^\top Xv) + \text{tr}(vu^\top RB))$$

$$= u^\top Rv(\text{tr}(u^\top Xv) + \text{tr}(vu^\top X))$$

$$= u^\top Rv(\text{tr}(u^\top Xv) + \text{tr}(u^\top X)v)$$

$$= 2u^\top Xvu^\top Rv, \quad \text{since } u^\top Xv \in \mathbb{R}$$

$$= df_R(X).$$

More generally, if $(M, \langle -, - \rangle)$ is a smooth manifold with a Riemannian metric and if $f : M \to \mathbb{R}$ is a smooth function on $M$, the unique smooth vector field $(\text{grad}(f))$ defined such that

$$\langle (\text{grad}(f))_p, u \rangle_p = df_p(u), \quad \text{for all } p \in M \text{ and all } u \in T_pM$$

is called the gradient of $f$. It is usually complicated to find the gradient of a function.

A way to obtain a metric on a manifold $N$, is to pull-back the metric $g$ on another manifold $M$ along a local diffeomorphism $\varphi : N \to M$. 
**Definition 10.3.** Recall that \( \varphi \) is a local diffeomorphism iff

\[
d\varphi_p : T_p N \to T_{\varphi(p)} M
\]

is a bijective linear map for every \( p \in N \). Given any metric \( g \) on \( M \), if \( \varphi \) is a local diffeomorphism, we define the pull-back metric \( \varphi^* g \) on \( N \) induced by \( g \) as follows: For all \( p \in N \), for all \( u, v \in T_p N \),

\[
(\varphi^* g)_p(u, v) = g_{\varphi(p)}(d\varphi_p(u), d\varphi_p(v)).
\]

We need to check that \( (\varphi^* g)_p \) is an inner product, which is very easy since \( d\varphi_p \) is a linear isomorphism.

The local diffeomorphism \( \varphi \) between the two Riemannian manifolds \( (N, \varphi^* g) \) and \( (M, g) \) has the special property that it is metric-preserving. Such maps are called local isometries, as defined below.

**Definition 10.4.** Given two Riemannian manifolds \( (M_1, g_1) \) and \( (M_2, g_2) \), a local isometry is a smooth map \( \varphi : M_1 \to M_2 \), such that \( d\varphi_p : T_p M_1 \to T_{\varphi(p)} M_2 \) is an isometry between the Euclidean spaces \( (T_p M_1, (g_1)_p) \) and \( (T_{\varphi(p)} M_2, (g_2)_{\varphi(p)}) \), for every \( p \in M_1 \); that is,

\[
(g_1)_p(u, v) = (g_2)_{\varphi(p)}(d\varphi_p(u), d\varphi_p(v)),
\]

for all \( u, v \in T_p M_1 \), or equivalently, \( \varphi^* g_2 = g_1 \). Moreover, \( \varphi \) is an isometry iff it is a local isometry and a diffeomorphism.

The isometries of a Riemannian manifold \( (M, g) \) form a group \( \text{Isom}(M, g) \), called the isometry group of \( (M, g) \). An important theorem of Myers and Steenrod asserts that the isometry group \( \text{Isom}(M, g) \) is a Lie group.

Given a map \( \varphi : M_1 \to M_2 \) and a metric \( g_1 \) on \( M_1 \), in general, \( \varphi \) does not induce any metric on \( M_2 \). However, if \( \varphi \) has some extra properties, it does induce a metric on \( M_2 \). This is the case when \( M_2 \) arises from \( M_1 \) as a quotient induced by some group of isometries of \( M_1 \). For more on this, see Gallot, Hulin and Lafontaine [73] (Chapter 2, Section 2.A), and Chapter 19.

Now, because a manifold is paracompact (see Section 9.1), a Riemannian metric always exists on \( M \). This is a consequence of the existence of partitions of unity (see Theorem 9.4).

**Theorem 10.2.** Every smooth manifold admits a Riemannian metric.

Theorem 10.2 is a special case of Theorem 28.9 which holds for vector bundles and is proved in Chapter 28.
Chapter 11
Connections on Manifolds

Given a manifold \( M \), in general, for any two points \( p, q \in M \), there is no “natural” isomorphism between the tangent spaces \( T_p M \) and \( T_q M \). Given a curve \( c: [0,1] \to M \) on \( M \), as \( c(t) \) moves on \( M \), how does the tangent space \( T_{c(t)} M \) change as \( c(t) \) moves?

If \( M = \mathbb{R}^n \), then the spaces \( T_{c(t)} \mathbb{R}^n \) are canonically isomorphic to \( \mathbb{R}^n \), and any vector \( v \in T_{c(0)} \mathbb{R}^n \cong \mathbb{R}^n \) is simply moved along \( c \) by parallel transport; that is, at \( c(t) \), the tangent vector \( v \) also belongs to \( T_{c(t)} \mathbb{R}^n \). However, if \( M \) is curved, for example a sphere, then it is not obvious how to “parallel transport” a tangent vector at \( c(0) \) along a curve \( c \). A way to achieve this is to define the notion of parallel vector field along a curve, and this can be defined in terms of the notion of covariant derivative of a vector field.

In Section 11.1, we define the general notion of a connection on a manifold \( M \) as a function \( \nabla: \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M) \) defined on vector fields and satisfying some properties that make it a generalization of the notion of covariant derivative on a surface. We show that \((\nabla_X Y)(p)\) only depends on the value of \( X \) at \( p \) and on the value of \( Y \) in a neighborhood of \( p \).

In Section 11.2, we show that the notion of covariant derivative is well-defined for vector fields along a curve. Given a vector field \( X \) along a curve \( \gamma \), this covariant derivative is denoted by \( DX/dt \). Then, we define the crucial notion of a vector field parallel along a curve \( \gamma \), which means that \( DX/dt(s) = 0 \) for all \( s \) (in the domain of \( \gamma \)). As a consequence, we can define the notion of parallel transport of a vector along a curve.

The notion of a connection on a manifold does not assume that the manifold is equipped with a Riemannian metric. In Section 11.3, we consider connections having additional properties, such as being compatible with a Riemannian metric or being torsion-free. Then, we have a phenomenon called by some people the “miracle” of Riemannian geometry, namely that for every Riemannian manifold, there is a unique connection which is torsion-free and compatible with the metric. Furthermore, this connection is determined by an implicit formula known as the Koszul formula. Such a connection is called the Levi-Civita connection. We conclude this section with some properties of connections compatible with a metric, in particular about parallel vectors fields along a curve.
11.1 Connections on Manifolds

Assume for simplicity that $M$ is a surface in $\mathbb{R}^3$. Given any two vector fields $X$ and $Y$ defined on some open subset $U \subseteq \mathbb{R}^3$, for every $p \in U$, the directional derivative $D_X Y(p)$ of $Y$ with respect to $X$ is defined by

$$D_X Y(p) = \lim_{t \to 0} \frac{Y(p + tX(p)) - Y(p)}{t}.$$ 

See Figure 11.1.

![Figure 11.1](image)

Figure 11.1: The directional derivative of the blue vector field $Y(p)$ in the direction of $X$.

If $f : U \to \mathbb{R}$ is a differentiable function on $U$, for every $p \in U$, the directional derivative $X[f](p)$ (or $X(f)(p)$) of $f$ with respect to $X$ is defined by

$$X[f](p) = \lim_{t \to 0} \frac{f(p + tX(p)) - f(p)}{t}.$$ 

We know that $X[f](p) = df_p(X(p))$.

It is easily shown that $D_X Y(p)$ is $\mathbb{R}$-bilinear in $X$ and $Y$, is $C^\infty(U)$-linear in $X$, and satisfies the Leibniz derivation rule with respect to $Y$; that is:

**Proposition 11.1.** If $X$ and $Y$ are vector fields from $U$ to $\mathbb{R}^3$ that are differentiable on some open subset $U$ of $\mathbb{R}^3$, then their directional derivatives satisfy the following properties:

$$D_{X_1 + X_2} Y(p) = D_{X_1} Y(p) + D_{X_2} Y(p)$$
$$D_{fX} Y(p) = f D_X Y(p)$$
$$D_X (Y_1 + Y_2)(p) = D_X Y_1(p) + D_X Y_2(p)$$
$$D_X (fY)(p) = X[f](p) Y(p) + f(p) D_X Y(p),$$
for all $X, X_1, X_2, Y, Y_1, Y_2 \in \mathfrak{X}(U)$ and all $f \in C^\infty(U)$.

**Proof.** By definition we have

$$D_X(Y_1 + Y_2)(p) = \lim_{t \to 0} \frac{(Y_1 + Y_2)(p + tX(p)) - (Y_1 + Y_2)(p)}{t}$$

$$= \lim_{t \to 0} \frac{Y_1(p + tX(p)) - Y_1(p)}{t} + \lim_{t \to 0} \frac{Y_2(p + tX(p)) - Y_2(p)}{t}$$

$$= D_X Y_1(p) + D_X Y_2(p).$$

Since $Y$ is assumed to be differentiable, $D_X Y(p) = dY_p(X(p))$, so by linearity of $dY_p$, we have

$$D_{X_1 + X_2} Y(p) = dY_p(X_1(p) + X_2(p)) = dY_p(X_1(p)) + dY_p(X_2(p)) = D_{X_1} Y(p) + D_{X_2} Y(p).$$

The definition also implies

$$D_X(fY)(p) = \lim_{t \to 0} \frac{fY(p + tX(p)) - fY(p)}{t}$$

$$= \lim_{t \to 0} \frac{f(p + tX(p))Y(p + tX(p)) - f(p)Y(p)}{t}$$

$$= \lim_{t \to 0} \frac{f(p + tX(p))Y(p + tX(p)) - f(p)Y(p + tX(p))}{t}$$

$$+ \lim_{t \to 0} \frac{f(p)Y(p + tX(p)) - f(p)Y(p)}{t}$$

$$= X[f](Y(p)) + f(p)D_X Y(p).$$

It remains to prove $D_{fX} Y(p) = fD_X Y(p)$. If $f(p) = 0$, this trivially true. So assume $f(p) \neq 0$. Then

$$D_{fX} Y(p) = f(p) \lim_{t \to 0} \frac{Y(p + tfX(p)) - Y(p)}{tf(p)} = f(p) \lim_{t \to 0} \frac{Y(p + tf(p)X(p)) - Y(p)}{tf(p)}$$

$$= f(p) \lim_{u \to 0} \frac{Y(p + uX(p)) - Y(p)}{u} = f(p)D_X Y(p).$$

\[\square\]

If $X$ and $Y$ are two vector fields defined on some open subset $U \subseteq \mathbb{R}^3$, and if there is some open subset $W \subseteq M$ of the surface $M$ such that $X(p), Y(p) \in T_p M$ for all $p \in W$, for every $p \in W$, the directional derivative $D_X Y(p)$ makes sense and it has an orthogonal decomposition

$$D_X Y(p) = \nabla_X Y(p) + (D_n) X Y(p),$$

where its *horizontal (or tangential) component* is $\nabla_X Y(p) \in T_p M$, and its normal component is $(D_n) X Y(p)$.  

The component $\nabla_X Y(p)$ is the covariant derivative of $Y$ with respect to $X \in T_p M$, and it allows us to define the covariant derivative of a vector field $Y \in \mathfrak{X}(U)$ with respect to a vector field $X \in \mathfrak{X}(M)$ on $M$. We easily check that $\nabla_X Y$ satisfies the four equations of Proposition 11.1.

In particular, $Y$ may be a vector field associated with a curve $c : [0, 1] \to M$. A vector field along a curve $c$ is a vector field $Y$ such that $Y(c(t)) \in T_{c(t)} M$, for all $t \in [0, 1]$. We also write $Y(t)$ for $Y(c(t))$. Then, we say that $Y$ is parallel along $c$ iff $\nabla_{c'(t)} Y = 0$ along $c$.

The notion of parallel transport on a surface can be defined using parallel vector fields along curves. Let $p, q$ be any two points on the surface $M$, and assume there is a curve $c : [0, 1] \to M$ joining $p = c(0)$ to $q = c(1)$. Then, using the uniqueness and existence theorem for ordinary differential equations, it can be shown that for any initial tangent vector $Y_0 \in T_p M$, there is a unique parallel vector field $Y$ along $c$, with $Y(0) = Y_0$. If we set $Y_1 = Y(1)$, we obtain a linear map $Y_0 \mapsto Y_1$ from $T_p M$ to $T_q M$ which is also an isometry.

As a summary, given a surface $M$, if we can define a notion of covariant derivative $\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ satisfying the properties of Proposition 11.1, then we can define the notion of parallel vector field along a curve, and the notion of parallel transport, which yields a natural way of relating two tangent spaces $T_p M$ and $T_q M$, using curves joining $p$ and $q$.

This can be generalized to manifolds using the notion of connection. We will see that the notion of connection induces the notion of curvature. Moreover, if $M$ has a Riemannian metric, we will see that this metric induces a unique connection with two extra properties (the Levi-Civita connection).

**Definition 11.1.** Let $M$ be a smooth manifold. A connection on $M$ is a $\mathbb{R}$-bilinear map

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M),$$

where we write $\nabla_X Y$ for $\nabla(X, Y)$, such that the following two conditions hold:

$$\nabla_{fX} Y = f \nabla_X Y$$

$$\nabla_X (fY) = X[f]Y + f \nabla_X Y,$$

for all $X, Y \in \mathfrak{X}(M)$ and all $f \in C^\infty(M)$. The vector field $\nabla_X Y$ is called the covariant derivative of $Y$ with respect to $X$.

A connection on $M$ is also known as an affine connection on $M$. A basic property of $\nabla$ is that it is a local operator.

**Proposition 11.2.** Let $M$ be a smooth manifold and let $\nabla$ be a connection on $M$. For every open subset $U \subseteq M$, for every vector field $Y \in \mathfrak{X}(M)$, if $Y \equiv 0$ on $U$, then $\nabla_X Y \equiv 0$ on $U$ for all $X \in \mathfrak{X}(M)$; that is, $\nabla$ is a local operator.
11.1. CONNECTIONS ON MANIFOLDS

Proposition 11.2 is a special case of Proposition 29.1 which is proved in Chapter 29. Proposition 11.2 implies that a connection $\nabla$ on $M$ restricts to a connection $\nabla | U$ on every open subset $U \subseteq M$.

It can also be shown that $(\nabla_X Y)(p)$ only depends on $X(p)$; that is, for any two vector fields $X, Y \in \mathfrak{X}(M)$, if $X(p) = Y(p)$ for some $p \in M$, then

$$(\nabla_X Z)(p) = (\nabla_Y Z)(p) \quad \text{for every } Z \in \mathfrak{X}(M).$$

A proof of the above fact is given in Chapter 29 (see Proposition 29.2).

Consequently, for any $p \in M$, the covariant derivative $(\nabla_u Y)(p)$ is well defined for any tangent vector $u \in T_p M$ and any vector field $Y$ defined on some open subset $U \subseteq M$, with $p \in U$.

Observe that on $U$, the $n$-tuple of vector fields $\left(\frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n}\right)$ is a local frame. We can write

$$\nabla \frac{\partial}{\partial x_i} \left(\frac{\partial}{\partial x_j}\right) = \sum_{k=1}^{n} \Gamma^k_{ij} \frac{\partial}{\partial x_k},$$

for some unique smooth functions $\Gamma^k_{ij}$ defined on $U$, called the Christoffel symbols.

We say that a connection $\nabla$ is flat on $U$ iff

$$\nabla_X \left(\frac{\partial}{\partial x_i}\right) = 0, \quad \text{for all } X \in \mathfrak{X}(U), \ 1 \leq i \leq n.$$ 

**Proposition 11.3.** Every smooth manifold $M$ possesses a connection.

**Proof.** We can find a family of charts $(U_\alpha, \varphi_\alpha)$ such that $\{U_\alpha\}_\alpha$ is a locally finite open cover of $M$. If $(f_\alpha)$ is a partition of unity subordinate to the cover $\{U_\alpha\}_\alpha$ and if $\nabla^\alpha$ is the flat connection on $U_\alpha$, then it is immediately verified that

$$\nabla = \sum_\alpha f_\alpha \nabla^\alpha$$

is a connection on $M$. \qed

**Remark:** A connection on $TM$ can be viewed as a linear map

$$\nabla : \mathfrak{X}(M) \longrightarrow \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathfrak{X}(M)),$$

such that, for any fixed $Y \in \mathfrak{X}(M)$, the map $\nabla Y : X \mapsto \nabla X Y$ is $C^\infty(M)$-linear, which implies that $\nabla Y$ is a $(1,1)$ tensor.
11.2 Parallel Transport

The notion of connection yields the notion of parallel transport. First, we need to define the covariant derivative of a vector field along a curve.

**Definition 11.2.** Let $M$ be a smooth manifold and let $\gamma: [a, b] \to M$ be a smooth curve in $M$. A smooth vector field along the curve $\gamma$ is a smooth map $X: [a, b] \to TM$, such that $\pi(X(t)) = \gamma(t)$, for all $t \in [a, b]$ $(X(t) \in T_{\gamma(t)}M)$.

Recall that the curve $\gamma: [a, b] \to M$ is smooth iff $\gamma$ is the restriction to $[a, b]$ of a smooth curve on some open interval containing $[a, b]$.

Since a vector $X$ field along a curve $\gamma$ does not necessarily extend to an open subset of $M$ (for example, if the image of $\gamma$ is dense in $M$), the covariant derivative $(\nabla_{\gamma'} X)_{\gamma(t_0)}$ may not be defined, so we need a proposition showing that the covariant derivative of a vector field along a curve makes sense.

**Proposition 11.4.** Let $M$ be a smooth manifold, let $\nabla$ be a connection on $M$ and $\gamma: [a, b] \to M$ be a smooth curve in $M$. There is a $\mathbb{R}$-linear map $D/dt$, defined on the vector space of smooth vector fields $X$ along $\gamma$, which satisfies the following conditions:

1. For any smooth function $f: [a, b] \to \mathbb{R}$,
   $$\frac{D(fX)}{dt} = \frac{df}{dt} X + f \frac{DX}{dt}.$$

2. If $X$ is induced by a vector field $Z \in \mathfrak{X}(M)$, that is $X(t_0) = Z(\gamma(t_0))$ for all $t_0 \in [a, b]$, then $\frac{DX}{dt}(t_0) = (\nabla_{\gamma'} Z)_{\gamma(t_0)}$.

**Proof.** Since $\gamma([a, b])$ is compact, it can be covered by a finite number of open subsets $U_\alpha$, such that $(U_\alpha, \varphi_\alpha)$ is a chart. Thus, we may assume that $\gamma: [a, b] \to U$ for some chart $(U, \varphi)$. As $\varphi \circ \gamma: [a, b] \to \mathbb{R}^n$, we can write

$$\varphi \circ \gamma(t) = (u_1(t), \ldots, u_n(t)),$$

where each $u_i = pr_i \circ \varphi \circ \gamma$ is smooth. Now, by applying the chain rule it is easy to see that

$$\gamma'(t_0) = \sum_{i=1}^n \frac{du_i}{dt} \left( \frac{\partial}{\partial x_i} \right)_{\gamma(t_0)}.$$

If $(s_1, \ldots, s_n)$ is a frame over $U$, we can write

$$X(t) = \sum_{i=1}^n X_i(t)s_i(\gamma(t)),$$
for some smooth functions $X_i$. For every $t \in [a, b]$, each vector fields $s_j$ over $U$ can be extended to a vector field on $M$ whose restriction to some open subset containing $\gamma(t)$ agrees with $s_j$, so conditions (1) and (2) imply that
\[
\frac{DX}{dt} = \sum_{j=1}^{n} \left( \frac{dX_j}{dt} s_j(\gamma(t)) + X_j(t)\nabla_{\gamma'(t)}(s_j(\gamma(t))) \right).
\]
Since
\[
\gamma'(t) = \sum_{i=1}^{n} \frac{du_i}{dt} \left( \frac{\partial}{\partial x_i} \right)_{\gamma(t)},
\]
there exist some smooth functions $\Gamma^k_{ij}$ (generally different from the Christoffel symbols) so that
\[
\nabla_{\gamma'(t)}(s_j(\gamma(t))) = \nabla \sum_{i=1}^{n} \frac{du_i}{dt} \left( \frac{\partial}{\partial x_i} \right)_{\gamma(t)}(s_j(\gamma(t))) + \frac{du_k}{dt} \nabla \left( \frac{\partial}{\partial x_k} \right)_{\gamma(t)}(s_j(\gamma(t)))
\]
\[
= \sum_{i=1}^{n} \frac{du_i}{dt} \nabla_{\gamma(t)} \left( \frac{\partial}{\partial x_i} \right)(s_j(\gamma(t)))
\]
\[
= \sum_{i,k} \frac{du_i}{dt} \Gamma^k_{ij} s_k(\gamma(t)).
\]
It follows that
\[
\frac{DX}{dt} = \sum_{k=1}^{n} \left( \frac{dX_k}{dt} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} X_j \right) s_k(\gamma(t)).
\]

Conversely, the above expression defines a linear operator $D/dt$, and it is easy to check that it satisfies (1) and (2).

**Definition 11.3.** The operator $D/dt$ is often called *covariant derivative along* $\gamma$ and it is also denoted by $\nabla_{\gamma'(t)}$ or simply $\nabla_{\gamma'}$.

**Definition 11.4.** Let $M$ be a smooth manifold and let $\nabla$ be a connection on $M$. For every curve $\gamma: [a, b] \to M$ in $M$, a vector field $X$ along $\gamma$ is *parallel (along $\gamma$)* iff
\[
\frac{DX}{dt}(s) = 0 \quad \text{for all} \quad s \in [a, b].
\]

If $M$ was embedded in $\mathbb{R}^d$ for some $d$, then to say that $X$ is parallel along $\gamma$ would mean that the directional derivative $(D_{\gamma'}X)(\gamma(t))$ is normal to $T_{\gamma(t)}M$.

The following proposition can be shown using the existence and uniqueness of solutions of ODE’s (in our case, linear ODE’s) and its proof is omitted:
**Proposition 11.5.** Let $M$ be a smooth manifold and let $\nabla$ be a connection on $M$. For every $C^1$ curve $\gamma: [a, b] \to M$ in $M$, for every $t \in [a, b]$ and every $v \in T_{\gamma(t)}M$, there is a unique parallel vector field $X$ along $\gamma$ such that $X(t) = v$.

**Proof.** For the proof of Proposition 11.5 it is sufficient to consider the portions of the curve $\gamma$ contained in some chart. In such a chart $(U, \varphi)$, as in the proof of Proposition 11.4, using a local frame $(s_1, \ldots, s_n)$ over $U$, we have

$$\frac{DX}{dt} = \sum_{k=1}^{n} \left( \frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{du_i}{dt} X_j \right) s_k(\gamma(t)),$$

with $u_i = pr_i \circ \varphi \circ \gamma$. Consequently, $X$ is parallel along our portion of $\gamma$ iff the system of linear ODE’s in the unknowns $X_k$,

$$\frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^k \frac{du_i}{dt} X_j = 0, \quad k = 1, \ldots, n,$$

is satisfied. \hfill $\square$

**Remark:** Proposition 11.5 can be extended to piecewise $C^1$ curves.

**Definition 11.5.** Let $M$ be a smooth manifold and let $\nabla$ be a connection on $M$. For every curve $\gamma: [a, b] \to M$ in $M$, for every $t \in [a, b]$, the **parallel transport from $\gamma(a)$ to $\gamma(t)$ along $\gamma$** is the linear map from $T_{\gamma(a)}M$ to $T_{\gamma(t)}M$ which associates to any $v \in T_{\gamma(a)}M$ the vector $X_v(t) \in T_{\gamma(t)}M$, where $X_v$ is the unique parallel vector field along $\gamma$ with $X_v(a) = v$. See Figure 11.2.

![Figure 11.2: The parallel transport of the red vector field around the spherical triangle ABC.](image-url)
11.3. CONNECTIONS COMPATIBLE WITH A METRIC

The following proposition is an immediate consequence of properties of linear ODE’s:

**Proposition 11.6.** Let $M$ be a smooth manifold and let $\nabla$ be a connection on $M$. For every $C^1$ curve $\gamma: [a,b] \rightarrow M$ in $M$, the parallel transport along $\gamma$ defines for every $t \in [a,b]$ a linear isomorphism $P_\gamma: T_{\gamma(a)}M \rightarrow T_{\gamma(t)}M$, between the tangent spaces $T_{\gamma(a)}M$ and $T_{\gamma(t)}M$.

In particular, if $\gamma$ is a closed curve, that is if $\gamma(a) = \gamma(b) = p$, we obtain a linear isomorphism $P_\gamma$ of the tangent space $T_pM$, called the holonomy of $\gamma$. The holonomy group of $\nabla$ based at $p$, denoted $\text{Hol}_p(\nabla)$, is the subgroup of $\text{GL}(n, \mathbb{R})$ (where $n$ is the dimension of the manifold $M$) given by

$$\text{Hol}_p(\nabla) = \{ P_\gamma \in \text{GL}(n, \mathbb{R}) \mid \gamma \text{ is a closed curve based at } p \}.$$

If $M$ is connected, then $\text{Hol}_p(\nabla)$ depends on the basepoint $p \in M$ up to conjugation, and so $\text{Hol}_p(\nabla)$ and $\text{Hol}_q(\nabla)$ are isomorphic for all $p, q \in M$. In this case, it makes sense to talk about the holonomy group of $\nabla$. By abuse of language, we call $\text{Hol}_p(\nabla)$ the holonomy group of $M$.

11.3 Connections Compatible with a Metric; Levi-Civita Connections

If a Riemannian manifold $M$ has a metric, then it is natural to define when a connection $\nabla$ on $M$ is compatible with the metric.

Given any two vector fields $Y, Z \in \mathfrak{X}(M)$, the smooth function $\langle Y, Z \rangle$ is defined by

$$\langle Y, Z \rangle(p) = \langle Y_p, Z_p \rangle_p,$$

for all $p \in M$.

**Definition 11.6.** Given any metric $\langle -, - \rangle$ on a smooth manifold $M$, a connection $\nabla$ on $M$ is compatible with the metric, for short, a metric connection, iff

$$X(\langle Y, Z \rangle) = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle,$$

for all vector fields, $X, Y, Z \in \mathfrak{X}(M)$.

**Proposition 11.7.** Let $M$ be a Riemannian manifold with a metric $\langle -, - \rangle$. Then $M$ possesses metric connections.

**Proof.** For every chart $(U_\alpha, \varphi_\alpha)$, we use the Gram-Schmidt procedure to obtain an orthonormal frame over $U_\alpha$ and we let $\nabla^\alpha$ be the flat connection over $U_\alpha$. By construction, $\nabla^\alpha$ is compatible with the metric. We finish the argument by using a partition of unity, leaving the details to the reader. \qed
We know from Proposition 11.7 that metric connections on $TM$ exist. However, there are many metric connections on $TM$ and none of them seems more relevant than the others.

It is remarkable that if we require a certain kind of symmetry on a metric connection, then it is uniquely determined. Such a connection is known as the Levi-Civita connection. The Levi-Civita connection can be characterized in several equivalent ways, a rather simple way involving the notion of torsion of a connection.

There are two error terms associated with a connection. The first one is the curvature $R(X, Y) = \nabla_{[X,Y]} + \nabla_Y \nabla_X - \nabla_X \nabla_Y$.

The second natural error term is the torsion $T(X, Y)$ of the connection $\nabla$, given by $T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]$, which measures the failure of the connection to behave like the Lie bracket.

**Proposition 11.8. (Levi-Civita, Version 1)** Let $M$ be any Riemannian manifold. There is a unique, metric, torsion-free connection $\nabla$ on $M$; that is, a connection satisfying the conditions:

$$X(\langle Y, Z \rangle) = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle,$$

$$\nabla_X Y - \nabla_Y X = [X, Y],$$

for all vector fields, $X, Y, Z \in \mathfrak{X}(M)$. This connection is called the Levi-Civita connection (or canonical connection) on $M$. Furthermore, this connection is determined by the Koszul formula

$$2\langle \nabla_X Y, Z \rangle = X(\langle Y, Z \rangle) + Y(\langle X, Z \rangle) - Z(\langle X, Y \rangle)$$

$$- \langle Y, [X, Z] \rangle - \langle X, [Y, Z] \rangle - \langle Z, [Y, X] \rangle.$$  

**Proof.** First we prove uniqueness. Since our metric is a non-degenerate bilinear form, it suffices to prove the Koszul formula. As our connection is compatible with the metric, we have

$$X(\langle Y, Z \rangle) = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle$$

$$Y(\langle X, Z \rangle) = \langle \nabla_Y X, Z \rangle + \langle X, \nabla_Y Z \rangle$$

$$-Z(\langle X, Y \rangle) = -\langle \nabla_Z X, Y \rangle - \langle X, \nabla_Z Y \rangle.$$ 

Adding up the above equation and using the fact that the torsion is zero gives us

$$X(\langle Y, Z \rangle) + Y(\langle X, Z \rangle) - Z(\langle X, Y \rangle)$$

$$= \langle Y, \nabla_X Z - \nabla_Z X \rangle + \langle X, \nabla_Y Z - \nabla_Z Y \rangle + \langle Z, \nabla_X Y + \nabla_Y X \rangle$$

$$= \langle Y, \nabla_X Z - \nabla_Z X \rangle + \langle X, \nabla_Y Z - \nabla_Z Y \rangle$$

$$+ \langle Z, \nabla_Y X - \nabla_X Y \rangle + \langle Z, \nabla_X Y + \nabla_Z Y \rangle$$

$$= \langle Y, [X, Z] \rangle + \langle X, [Y, Z] \rangle + \langle Z, [Y, X] \rangle + 2\langle Z, \nabla_X Y \rangle,$$
which yields the Koszul formula.

Next, we prove existence. We begin by checking that the right-hand side of the Koszul formula is \( C^\infty(M) \)-linear in \( Z \), for \( X \) and \( Y \) fixed. But then, the linear map \( Z \mapsto \langle \nabla_X Y, Z \rangle \) induces a one-form and \( \nabla_X Y \) is the vector field corresponding to it \textit{via} the non-degenerate pairing. It remains to check that \( \nabla \) satisfies the properties of a connection, which it a bit tedious (for example, see Kuhnel [109], Chapter 5, Section D). \( \square \)

In the simple case where \( M = \mathbb{R}^n \) and the metric is the Euclidean inner product on \( \mathbb{R}^n \), any two smooth vector fields \( X, Y \) can be written as

\[
X = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^{n} g_i \frac{\partial}{\partial x_i},
\]

for some smooth functions \( f_i, g_i \), and they can be viewed as smooth functions \( X, Y : \mathbb{R}^n \to \mathbb{R}^n \). Then, it is easy to verify that the Levi-Civita connection is given by

\[
(\nabla_X Y)(p) = dY_p(X(p)), \quad p \in \mathbb{R}^n,
\]

because the right-hand side satisfies all the conditions of Proposition 11.8, and there is a unique such connection. Thus, the Levi-Civita connection induced by the Euclidean inner product on \( \mathbb{R}^n \) is the flat connection.

Remark: In a chart \((U, \varphi)\), if we set

\[
\partial_k g_{ij} = \frac{\partial}{\partial x_k}(g_{ij}),
\]

then it can be shown that the Christoffel symbols of the Levi-Civita connection are given by

\[
\Gamma^k_{ij} = \frac{1}{2} \sum_{l=1}^{n} g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}),
\]

where \((g^{kl})\) is the inverse of the matrix \((g_{kl})\). For example suppose we take the polar coordinate parameterization of the plane given by

\[
x = r \cos \theta \quad y = r \sin \theta,
\]

with \( 0 < \theta < 2\pi, r > 0 \). For any \( p = (r \cos \theta, r \sin \theta) \), a basis for the tangent plane \( T_p\mathbb{R}^2 \) is

\[
\frac{\partial p}{\partial r} = (\cos \theta, \sin \theta) \quad \frac{\partial p}{\partial \theta} = (-r \sin \theta, r \cos \theta).
\]
Since
\[
\langle \frac{\partial p}{\partial r}, \frac{\partial p}{\partial r} \rangle = 1
\]
\[
\langle \frac{\partial p}{\partial r}, \frac{\partial p}{\partial \theta} \rangle = 0
\]
\[
\langle \frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \theta} \rangle = r^2,
\]
we discover that
\[
g = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}, \quad g^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{r^2} \end{pmatrix}.
\]
By associating \( r \) with 1 and \( \theta \) with 2, we discover that
\[
\Gamma^r_{\theta\theta} = \Gamma^1_{22} = -r,
\]
since
\[
\Gamma^i_{jk} = \frac{1}{2} \sum_{l=1}^{2} g^{il} \left( \partial_2 g_{jl} + \partial_2 g_{jl} - \partial_1 g_{jl} \right) \\
= \frac{1}{2} \left[ g^{11}(2 \partial_2 g_{21} - \partial_1 g_{22}) + g^{12}(2 \partial_2 g_{22} - \partial_2 g_{22}) \right] \\
= -\frac{1}{2} g^{11} \partial_1 g_{22} = -\frac{1}{2} \frac{\partial}{\partial r} g_{22} \\
= -\frac{1}{2} \frac{\partial}{\partial r} r^2 = -r.
\]
Similar calculations show that
\[
\Gamma^r_{\theta\theta} = \Gamma^1_{12} = \Gamma^1_{21} = 0
\]
\[
\Gamma^\theta_{r\theta} = \Gamma^2_{12} = \Gamma^2_{21} = \frac{1}{r}
\]
\[
\Gamma^r_{rr} = \Gamma^1_{11} = 0
\]
\[
\Gamma^\theta_{rr} = \Gamma^1_{22} = 0
\]
\[
\Gamma^\theta_{\theta\theta} = \Gamma^2_{22} = 0.
\]
Since
\[
\nabla_{\frac{\partial}{\partial x_i}} \left( \frac{\partial}{\partial x_j} \right) = \sum_{k=1}^{n} \Gamma_{ij}^{k} \frac{\partial}{\partial x_k},
\]
we explicitly calculate the Levi-Civita connection as

\[ \nabla_{\partial_r} \left( \frac{\partial}{\partial r} \right) = \sum_{k=1}^{2} \Gamma_{11}^{k} \frac{\partial}{\partial x_k} = 0 \]

\[ \nabla_{\partial \theta} \left( \frac{\partial}{\partial r} \right) = \sum_{k=1}^{2} \Gamma_{112}^{k} \frac{\partial}{\partial x_k} = \frac{1}{r} \frac{\partial}{\partial \theta} \]

\[ \nabla_{\partial r} \left( \frac{\partial}{\partial \theta} \right) = \sum_{k=1}^{2} \Gamma_{21}^{k} \frac{\partial}{\partial x_k} = \frac{1}{r} \frac{\partial}{\partial \theta} \]

\[ \nabla_{\partial \theta} \left( \frac{\partial}{\partial \theta} \right) = \sum_{k=1}^{2} \Gamma_{22}^{k} \frac{\partial}{\partial x_k} = -r \frac{\partial}{\partial r} \]

It can be shown that a connection is torsion-free iff

\[ \Gamma_{ij}^{k} = \Gamma_{ji}^{k}, \quad \text{for all } i,j,k. \]

We conclude this section with various useful facts about torsion-free or metric connections. First, there is a nice characterization for the Levi-Civita connection induced by a Riemannian manifold over a submanifold.

**Proposition 11.9.** Let \( M \) be any Riemannian manifold and let \( N \) be any submanifold of \( M \) equipped with the induced metric. If \( \nabla^{M} \) and \( \nabla^{N} \) are the Levi-Civita connections on \( M \) and \( N \), respectively, induced by the metric on \( M \), then for any two vector field \( X \) and \( Y \) in \( \mathfrak{X}(M) \) with \( X(p), Y(p) \in T_{p}N \), for all \( p \in N \), we have

\[ \nabla^{N}_{X}Y = (\nabla^{M}_{X}Y)\parallel, \]

where \( (\nabla^{M}_{X}Y)\parallel(p) \) is the orthogonal projection of \( \nabla^{M}_{X}Y(p) \) onto \( T_{p}N \), for every \( p \in N \).

In particular, if \( \gamma \) is a curve on a surface \( M \subseteq \mathbb{R}^{3} \), then a vector field \( X(t) \) along \( \gamma \) is parallel iff \( X'(t) \) is normal to the tangent plane \( T_{\gamma(t)}M \).

If \( \nabla \) is a metric connection, then we can say more about the parallel transport along a curve. Recall from Section 11.2, Definition 11.4, that a vector field \( X \) along a curve \( \gamma \) is parallel iff

\[ \frac{dX}{dt} = 0. \]

The following proposition will be needed:

**Proposition 11.10.** Given any Riemannian manifold \( M \) and any metric connection \( \nabla \) on \( M \), for every curve \( \gamma : [a, b] \to M \) on \( M \), if \( X \) and \( Y \) are two vector fields along \( \gamma \), then

\[ \frac{d}{dt} \langle X(\gamma(t)), Y(\gamma(t)) \rangle = \left\langle \frac{DX}{dt}, Y(\gamma(t)) \right\rangle + \left\langle X(\gamma(t)), \frac{DY}{dt} \right\rangle. \]
Proof. Since
\[ \frac{d}{dt} \langle X(\gamma(t)), Y(\gamma(t)) \rangle = d \langle X, Y \rangle_{\gamma(t)}(\gamma'(t)) = \gamma'(t) \langle X, Y \rangle_{\gamma(t)}, \]
it would be tempting to apply directly the equation
\[ Z(\langle X, Y \rangle) = \langle \nabla_Z X, Y \rangle + \langle X, \nabla_Z Y \rangle \]
asserting the compatibility of the connection with the metric, but this is wrong because
the above equation applies to vector fields \( X, Y \) defined on the whole of \( M \) (or at least
one some open subset of \( M \)), and yet in our situation \( X \) and \( Y \) are only defined along the
curve \( \gamma \), and in general, such vector fields cannot be extended to an open subset of \( M \). This
subtle point seems to have been overlooked in several of the classical texts. Note that Milnor
[125] circumvents this difficulty by defining compatibility in a different way (which turns
out to be equivalent to the notion used here). Our way out is to use charts, as in the proof
of Proposition 11.4; this is the proof method used by O’Neill [138] and Gallot, Hulin and
Lafontaine [73] (Chapter 2), although they leave computations to the reader.

We may assume that \( \gamma: [a, b] \to U \) for some chart \((U, \varphi)\). Then, if \((s_1, \ldots, s_n)\) is a frame
above \( U \), we can write
\[
X(\gamma(t)) = \sum_{i=1}^{n} X_i(t) s_i(\gamma(t)) \\
Y(\gamma(t)) = \sum_{k=1}^{n} Y_k(t) s_k(\gamma(t)),
\]
and as in the proof of Proposition 11.4, we have
\[
\frac{DX}{dt} = \sum_{j=1}^{n} \left( \frac{dX_j}{dt} s_j(\gamma(t)) + X_j(t) \nabla_{\gamma'(t)}(s_j(\gamma(t))) \right) \\
\frac{DY}{dt} = \sum_{l=1}^{n} \left( \frac{dY_l}{dt} s_l(\gamma(t)) + Y_l(t) \nabla_{\gamma'(t)}(s_l(\gamma(t))) \right).
\]
It follows that
\[
\left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle = \sum_{j,k=1}^{n} \frac{dX_j}{dt} Y_k(t) \langle s_j(\gamma(t)), s_k(\gamma(t)) \rangle \\
+ \sum_{j,k=1}^{n} X_j(t) Y_k(t) \langle \nabla_{\gamma'(t)} s_j(\gamma(t)), s_k(\gamma(t)) \rangle \\
+ \sum_{i,l=1}^{n} X_i(t) \frac{dY_l}{dt} \langle s_i(\gamma(t)), s_l(\gamma(t)) \rangle \\
+ \sum_{i,l=1}^{n} X_i(t) Y_l(t) \langle s_i(\gamma(t)), \nabla_{\gamma'(t)} s_l(\gamma(t)) \rangle,
\]

so

\[ \left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle = \sum_{i,k=1}^{n} \left( \frac{dX_i}{dt} Y_k(t) + X_i(t) \frac{dY_k}{dt} \right) \langle s_i(\gamma(t)), s_k(\gamma(t)) \rangle \\
+ \sum_{i,k=1}^{n} X_i(t) Y_k(t) \left( \langle \nabla_{\gamma'(t)} s_i(\gamma(t)), s_k(\gamma(t)) \rangle + \langle s_i(\gamma(t)), \nabla_{\gamma'(t)} s_k(\gamma(t)) \rangle \right). \]

On the other hand, the compatibility of the connection with the metric implies that

\[ \langle \nabla_{\gamma'(t)} s_i(\gamma(t)), s_k(\gamma(t)) \rangle + \langle s_i(\gamma(t)), \nabla_{\gamma'(t)} s_k(\gamma(t)) \rangle = \gamma'(t) \langle s_i, s_k \rangle_{\gamma(t)} = \frac{d}{dt} \langle s_i(\gamma(t)), s_k(\gamma(t)) \rangle, \]

and thus we have

\[ \left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle = \sum_{i,k=1}^{n} \left( \frac{dX_i}{dt} Y_k(t) + X_i(t) \frac{dY_k}{dt} \right) \langle s_i(\gamma(t)), s_k(\gamma(t)) \rangle \\
+ \sum_{i,k=1}^{n} X_i(t) Y_k(t) \frac{d}{dt} \langle s_i(\gamma(t)), s_k(\gamma(t)) \rangle \\
= \frac{d}{dt} \left( \sum_{i,k=1}^{n} X_i(t) Y_k(t) \langle s_i(\gamma(t)), s_k(\gamma(t)) \rangle \right) \\
= \frac{d}{dt} \langle X(\gamma(t)), Y(\gamma(t)) \rangle, \]

as claimed. \(\Box\)

Using Proposition 11.10 we get

**Proposition 11.11.** Given any Riemannian manifold \(M\) and any metric connection \(\nabla\) on \(M\), for every curve \(\gamma: [a,b] \to M\) on \(M\), if \(X\) and \(Y\) are two vector fields along \(\gamma\) that are parallel, then

\[ \langle X, Y \rangle = C, \]

for some constant \(C\). In particular, \(\|X(t)\|\) is constant. Furthermore, the linear isomorphism \(P_\gamma: T_{\gamma(a)} \to T_{\gamma(b)}\) is an isometry.

**Proof.** From Proposition 11.10, we have

\[ \frac{d}{dt} \langle X(\gamma(t)), Y(\gamma(t)) \rangle = \left\langle \frac{DX}{dt}, Y(\gamma(t)) \right\rangle + \left\langle X(\gamma(t)), \frac{DY}{dt} \right\rangle. \]

As \(X\) and \(Y\) are parallel along \(\gamma\), we have \(DX/dt = 0\) and \(DY/dt = 0\), so

\[ \frac{d}{dt} \langle X(\gamma(t)), Y(\gamma(t)) \rangle = 0, \]
which shows that \( \langle X(\gamma(t)), Y(\gamma(t)) \rangle \) is constant. Therefore, for all \( v, w \in T_{\gamma(a)} \), if \( X \) and \( Y \) are the unique vector fields parallel along \( \gamma \) such that \( X(\gamma(a)) = v \) and \( Y(\gamma(a)) = w \) given by Proposition 29.13, we have

\[
\langle P_\gamma(v), P_\gamma(w) \rangle = \langle X(\gamma(b)), Y(\gamma(b)) \rangle = \langle X(\gamma(a)), Y(\gamma(a)) \rangle = \langle v, w \rangle,
\]

which proves that \( P_\gamma \) is an isometry.

In particular, Proposition 11.11 shows that the holonomy group \( \text{Hol}_p(\nabla) \) based at \( p \) is a subgroup of \( O(n) \).
Chapter 12

Geodesics on Riemannian Manifolds

If \((M, g)\) is a Riemannian manifold, then the concept of length makes sense for any piecewise smooth (in fact, \(C^1\)) curve on \(M\). Then, it is possible to define the structure of a metric space on \(M\), where \(d(p, q)\) is the greatest lower bound of the length of all curves joining \(p\) and \(q\). Curves on \(M\) which locally yield the shortest distance between two points are of great interest. These curves called geodesics play an important role and the goal of this chapter is to study some of their properties.

In Section 12.1, we define geodesics and prove some of their basic properties, in particular the fact that they always exist locally. Note that the notion of geodesic only requires a connection on a manifold, since by definition, a geodesic is a curve \(\gamma\) such that \(\gamma'\) is parallel along \(\gamma\), that is

\[
\frac{D\gamma'}{dt} = \nabla_{\gamma'}\gamma' = 0,
\]

where \(\frac{D}{dt}\) be the covariant derivative along \(\gamma\), also denoted \(\nabla_{\gamma'}\) (see Proposition 11.4 and Definition 11.3). Thus, geodesics can be defined in manifolds that are not endowed with a Riemannian metric. However, most useful properties of geodesics involve metric notions, and their proofs use the fact that the connection on the manifold is compatible with the metric and torsion-free. For this reason, we usually assume that we are dealing with Riemannian manifolds equipped with the Levi-Civita connection. We conclude Section 12.1 with the definition of the Hessian of a function defined on a Riemannian manifold, and show how the Hessian can be computed using geodesics.

For every point \(p \in M\) on a manifold \(M\), using geodesics through \(p\) we can define the exponential map \(\exp_p\), which maps a neighborhood of 0 in the tangent space \(T_pM\) back into \(M\); see Section 12.2. The exponential map is a very useful technical tool because it establishes a precise link between the linearization of a manifold by its tangent spaces and the manifold itself. In particular, manifolds for which the exponential map is defined for all \(p \in M\) and all \(v \in T_pM\) can be studied in more depth; see Section 12.3. Such manifolds are called complete. A fundamental theorem about complete manifolds is the theorem of Hopf and Rinow, which we prove in full.
Geodesics are locally distance minimizing, but in general they fail to be distance minimizing if they extend too far. This phenomenon is captured by the subtle notion of cut locus, which we define and study briefly. In Section 12.4, we also discuss briefly various notions of convexity induced by geodesics.

Geodesics between two points \( p \) and \( q \) turn out to be critical points of the energy functional on the path space \( \Omega(p, q) \), the space of all piecewise smooth curves from \( p \) to \( q \). This is an infinite dimensional manifold consisting of functions (curves), so in order to define what it means for a curve \( \omega \) in \( \Omega(p, q) \) to be a critical point of a function \( F \) defined on \( \Omega(p, q) \), we introduce the notion of variation (of a curve). Then, it is possible to obtain a formula giving the derivative \( dE(\tilde{\alpha}(u))/du \mid_{u=0} \) of the energy function \( E \) (with \( E(\omega) = \int_0^1 \|\omega'(t)\|^2 dt \)) applied to a variation \( \tilde{\alpha} \) of a curve \( \omega \) (the first variation formula); see Section 12.5. It turns out that a curve \( \omega \) is a geodesic iff it is a critical point of the energy function (that is, \( dE(\tilde{\alpha}(u))/du \mid_{u=0} = 0 \) for all variations of \( \omega \)). This result provides a fruitful link with the calculus of variations.

Since geodesics are a standard chapter of every differential geometry text, we will omit many proofs and instead give precise pointers to the literature.

Among the many presentations of this subject, in our opinion, Milnor’s account [125] (Part II, Section 11) is still one of the best, certainly by its clarity and elegance. We acknowledge that our presentation was heavily inspired by this beautiful work. We also relied heavily on Gallot, Hulin and Lafontaine [73] (Chapter 2), Do Carmo [60], O’Neill [138], Kuhnel [109], and class notes by Pierre Pansu (see http://www.math.u-psud.fr/%7Epansu/web_dea/resume_dea_04.html in http://www.math.u-psud.fr/~pansu/). Another reference that is remarkable by its clarity and the completeness of its coverage is Postnikov [144].

# 12.1 Geodesics, Local Existence and Uniqueness

Recall the following definitions regarding curves:

**Definition 12.1.** Given any smooth manifold \( M \), a smooth parametric curve (for short, curve) on \( M \) is a smooth map \( \gamma: I \to M \), where \( I \) is some open interval of \( \mathbb{R} \). For a closed interval \([a, b] \subseteq \mathbb{R}\), a map \( \gamma: [a, b] \to M \) is a smooth curve from \( p = \gamma(a) \) to \( q = \gamma(b) \) iff \( \gamma \) can be extended to a smooth curve \( \tilde{\gamma}: (a - \epsilon, b + \epsilon) \to M \), for some \( \epsilon > 0 \). Given any two points \( p, q \in M \), a continuous map \( \gamma: [a, b] \to M \) is a piecewise smooth curve from \( p \) to \( q \) iff

1. There is a sequence \( a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b \) of numbers \( t_i \in \mathbb{R} \), so that each map \( \gamma_i = \gamma \mid [t_i, t_{i+1}] \), called a curve segment, is a smooth curve for \( i = 0, \ldots, k - 1 \).
2. \( \gamma(a) = p \) and \( \gamma(b) = q \).

The set of all piecewise smooth curves from \( p \) to \( q \) is denoted by \( \Omega(M; p, q) \), or briefly by \( \Omega(p, q) \) (or even by \( \Omega \), when \( p \) and \( q \) are understood).
The set $\Omega(M; p, q)$ is an important object sometimes called the *path space* of $M$ (from $p$ to $q$). Unfortunately it is an infinite-dimensional manifold, which makes it hard to investigate its properties.

Observe that at any junction point $\gamma_{i-1}(t_i) = \gamma_i(t_i)$, there may be a jump in the velocity vector of $\gamma$. We let $\gamma'(t_i) = \gamma_i'(t_i)$ and $\gamma''(t_i) = \gamma_{i-1}'(t_i)$.

Let $(M, g)$ be a Riemannian manifold. Given any $p \in M$, for every $v \in T_p M$, the *(Riemannian) norm* of $v$, denoted $\|v\|$, is defined by

$$\|v\| = \sqrt{g_p(v, v)}.$$

The Riemannian inner product $g_p(u, v)$ of two tangent vectors $u, v \in T_p M$ will also be denoted by $\langle u, v \rangle_p$, or simply $\langle u, v \rangle$. Given any curve $\gamma \in \Omega(M; p, q)$, the *length* $L(\gamma)$ of $\gamma$ is defined by

$$L(\gamma) = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \|\gamma'(t)\| \, dt = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \sqrt{g(\gamma'(t), \gamma'(t))} \, dt.$$

It is easy to see that $L(\gamma)$ is unchanged by a monotone reparametrization (that is, a map $h: [a, b] \to [c, d]$ whose derivative $h'$ has a constant sign).

Now, let $M$ be any smooth manifold equipped with an arbitrary connection $\nabla$. For every curve $\gamma$ on $M$, recall that $\frac{D\gamma'}{dt}$ is the associated covariant derivative along $\gamma$, also denoted $\nabla_{\gamma'}$ (see Proposition 11.11 and Definition 11.3).

**Definition 12.2.** Let $M$ be any smooth manifold equipped with a connection $\nabla$. A curve $\gamma: I \to M$ (where $I \subseteq \mathbb{R}$ is any interval) is a *geodesic* iff $\gamma'(t)$ is parallel along $\gamma$; that is, iff

$$\frac{D\gamma'}{dt} = \nabla_{\gamma'} \gamma' = 0.$$

Observe that the notion of geodesic only requires a connection on a manifold, and that geodesics can be defined in manifolds that are not endowed with a Riemannian metric. However, most useful properties of geodesics involve metric notions, and their proofs use the fact that the connection on the manifold is compatible with the metric and torsion-free. Therefore, from on on, we assume that our Riemannian manifold $(M, g)$ is equipped with the Levi-Civita connection.

If $M$ was embedded in $\mathbb{R}^d$, a geodesic would be a curve $\gamma$ such that the acceleration vector $\gamma'' = \frac{D\gamma'}{dt}$ is normal to $T_{\gamma(t)} M$.

Since our connection is compatible with the metric, by Proposition 11.11, $\|\gamma'(t)\| = \sqrt{g(\gamma'(t), \gamma'(t))}$ is constant, say $\|\gamma'(t)\| = c$. If we define the *arc-length* function $s(t)$ relative to $a$, where $a$ is any chosen point in $I$, by

$$s(t) = \int_a^t \sqrt{g(\gamma'(t), \gamma'(t))} \, dt = c(t - a), \quad t \in I,$$
we conclude that for a geodesic $\gamma(t)$, the parameter $t$ is an affine function of the arc-length. When $\epsilon = 1$, which can be achieved by an affine reparametrization, we say that the geodesic is normalized.

The geodesics in $\mathbb{R}^n$ are the straight lines parametrized by constant velocity. The geodesics of the 2-sphere are the great circles, parametrized by arc-length. The geodesics of the Poincaré half-plane are the lines $x = a$ and the half-circles centered on the $x$-axis. The geodesics of an ellipsoid are quite fascinating. They can be completely characterized, and they are parametrized by elliptic functions (see Hilbert and Cohn-Vossen [90], Chapter 4, Section and Berger and Gostiaux [20], Section 10.4.9.5). If $M$ is a submanifold of $\mathbb{R}^n$, geodesics are curves whose acceleration vector $\gamma'' = D\gamma'/dt$ is normal to $M$ (that is, for every $p \in M$, $\gamma''$ is normal to $T_p M$).

In a local chart $(U, \varphi)$, since a geodesic is characterized by the fact that its velocity vector field $\gamma'(t)$ along $\gamma$ is parallel, by Proposition 11.5, it is the solution of the following system of second-order ODE’s in the unknowns $u_k$:

$$\frac{d^2 u_k}{dt^2} + \sum_{ij} \Gamma_{ij}^k \frac{du_i}{dt} \frac{du_j}{dt} = 0, \quad k = 1, \ldots, n,$$

with $u_i = pr_i \circ \varphi \circ \gamma$ ($n = \dim(M)$).

The standard existence and uniqueness results for ODE’s can be used to prove the following proposition (see O’Neill [138], Chapter 3):

**Proposition 12.1.** Let $(M, g)$ be a Riemannian manifold. For every point $p \in M$ and every tangent vector $v \in T_p M$, there is some interval $(-\eta, \eta)$ and a unique geodesic

$$\gamma_v : (-\eta, \eta) \to M,$$

satisfying the conditions

$$\gamma_v(0) = p, \quad \gamma'_v(0) = v.$$

The following proposition is used to prove that every geodesic is contained in a unique maximal geodesic (i.e., with largest possible domain). For a proof, see O’Neill [138] (Chapter 3) or Petersen [140] (Chapter 5, Section 2, Lemma 7).

**Proposition 12.2.** For any two geodesics $\gamma_1 : I_1 \to M$ and $\gamma_2 : I_2 \to M$, if $\gamma_1(a) = \gamma_2(a)$ and $\gamma'_1(a) = \gamma'_2(a)$ for some $a \in I_1 \cap I_2$, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$.

**Remark:** It is easy to check that Propositions 12.1 and 12.2 hold for any smooth manifold equipped with a connection.

Propositions 12.1 and 12.2 imply that for every $p \in M$ and every $v \in T_p M$, there is a unique geodesic, denoted $\gamma_v$, such that $\gamma(0) = p, \gamma'(0) = v$, and the domain of $\gamma$ is the largest
possible, that is, cannot be extended. We call $\gamma_v$ a \textit{maximal geodesic} (with initial conditions $\gamma_v(0) = p$ and $\gamma_v'(0) = v$).

Observe that the system of differential equations satisfied by geodesics has the following homogeneity property: If $t \mapsto \gamma(t)$ is a solution of the above system, then for every constant $c$, the curve $t \mapsto \gamma(ct)$ is also a solution of the system. We can use this fact together with standard existence and uniqueness results for ODE’s to prove the proposition below.

**Proposition 12.3.** Let $(M, g)$ be a Riemannian manifold. For every point $p_0 \in M$, there is an open subset $U \subseteq M$, with $p_0 \in U$, and some $\epsilon > 0$, so that: For every $p \in U$ and every tangent vector $v \in T_p M$, with $\|v\| < \epsilon$, there is a unique geodesic $\gamma_v: (-2, 2) \to M$ satisfying the conditions

$$\gamma_v(0) = p, \quad \gamma_v'(0) = v.$$ 

**Proof.** We follow Milnor [125] (Part II, Section 10, Proposition 10.2). By a standard theorem about the existence and uniqueness of solutions of ODE’s, for every $p_0 \in M$, there is some open subset $U$ of $M$ containing $p_0$ and some numbers $\epsilon_1 > 0$ and $\epsilon_2 > 0$, such that for every $p \in M$ and every $v \in T_p M$ with $\|v\| < \epsilon_1$, there is a unique geodesic $\tilde{\gamma}_v: (-2\epsilon_2, 2\epsilon_2) \to M$ such that $\tilde{\gamma}_v(0) = p$ and $\tilde{\gamma}_v'(0) = v$. For any constant $c \neq 0$, the curve $t \mapsto \tilde{\gamma}_v(ct)$ is a geodesic defined on $(-\eta/c, \eta/c)$ (or $(\eta/c, -\eta/c)$ if $c < 0$) such that $\tilde{\gamma}_v'(0) = cv$. Thus,

$$\tilde{\gamma}_v(ct) = \tilde{\gamma}_{cv}(t), \quad ct \in (-\eta, \eta).$$

Pick $\epsilon > 0$ so that $\epsilon < \epsilon_1\epsilon_2$. Then, if $\|v\| < \epsilon$ and $|t| < 2$, note that

$$\|v/\epsilon_2\| < \epsilon_1 \quad \text{and} \quad |\epsilon_2 t| < 2\epsilon_2.$$ 

Hence, we can define the geodesic $\gamma_v$ by

$$\gamma_v(t) = \tilde{\gamma}_{v/\epsilon_2}(\epsilon_2 t), \quad \|v\| < \epsilon, \quad |t| < 2,$$

and we have $\gamma_v(0) = p$ and $\gamma_v'(0) = v$, which concludes the proof. \qed

**Remark:** Proposition 12.3 holds for a Riemannian manifold equipped with an arbitrary connection.

Besides the notion of the gradient of a function, there is also the notion of Hessian. Now that we have geodesics at our disposal, we also have a method to compute the Hessian, a task which is generally quite complex.

Given a smooth function $f: M \to \mathbb{R}$ on a Riemannian manifold $M$ equipped with the Levi-Civita connection, recall that the gradient $\text{grad} \ f$ of $f$ is the vector field uniquely defined by the condition

$$\langle (\text{grad} \ f)_p, u \rangle_p = df_p(u) = u(f), \quad \text{for all} \ u \in T_p M \ \text{and all} \ p \in M.$$
Definition 12.3. The \textit{Hessian} \( \text{Hess}(f) \) (or \( \nabla^2(f) \)) of a function \( f \in C^\infty(M) \) is defined by

\[
\text{Hess}(f)(X,Y) = X(Y(f)) - (\nabla_X Y)(f) = X(df(Y)) - df(\nabla_X Y),
\]
for all vector fields \( X,Y \in \mathfrak{X}(M) \).

Since \( \nabla \) is torsion-free, we get

\[
\nabla_X Y(f) - \nabla_Y X(f) = [X,Y](f) = X(Y(f)) - Y(X(f)),
\]
which in turn implies

\[
\text{Hess}(f)(X,Y) = X(Y(f)) - (\nabla_X Y)(f) = Y(X(f)) - (\nabla_Y X)(f) = \text{Hess}(f)(Y,X),
\]
which means that the Hessian is \textit{symmetric}.

Proposition 12.4. The Hessian is given by the equation

\[
\text{Hess}(f)(X,Y) = \langle \nabla_X (\text{grad} f), Y \rangle, \quad X,Y \in \mathfrak{X}(M).
\]

\textbf{Proof.} We have

\[
\begin{align*}
X(Y(f)) & = X(df(Y)) \\
& = X(\langle \text{grad} f, Y \rangle) \\
& = \langle \nabla_X (\text{grad} f), Y \rangle + \langle \text{grad} f, \nabla_X Y \rangle \\
& = \langle \nabla_X (\text{grad} f), Y \rangle + (\nabla_X Y)(f)
\end{align*}
\]

which yields

\[
\langle \nabla_X (\text{grad} f), Y \rangle = X(Y(f)) - (\nabla_X Y)(f) = \text{Hess}(f)(X,Y),
\]
as claimed. \qed

In the simple case where \( M = \mathbb{R}^n \) and the metric is the usual Euclidean inner product on \( \mathbb{R}^n \), we can easily compute the Hessian of a function \( f : \mathbb{R}^n \to \mathbb{R} \). For any two vector fields

\[
X = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^{n} y_i \frac{\partial}{\partial x_i},
\]

with \( x_i, y_i \in \mathbb{R} \), we have \( \nabla_X Y = dY(X) = 0 \) (\( x_i, y_i \) are constants and the Levi-Civita connection induced by the Euclidean inner product is the flat connection), so \( \text{Hess}(f)(X,Y) = X(Y(f)) \) and if we write \( x^\top = (x_1, \ldots, x_n)^\top \) and \( y^\top = (y_1, \ldots, y_n)^\top \), it is easy to see that

\[
\text{Hess}(f)_p(X,Y) = x^\top H_p y,
\]

where \( H_p \) is the matrix

\[
H_p = \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(p) \right),
\]
the usual Hessian matrix of the function \( f \) at \( p \).

In the general case of a Riemannian manifold \((M, \langle -, - \rangle)\), given any function \( f \in C^\infty(M) \), for any \( p \in M \) and for any \( u \in T_pM \), the value of the Hessian \( \text{Hess}(f)_p(u, u) \) can be computed using geodesics. Indeed, for any geodesic \( \gamma : [0, \epsilon] \to M \) such that \( \gamma(0) = p \) and \( \gamma'(0) = u \), we have

\[
\text{Hess}(f)_p(u, u) = \gamma'(\gamma'(f)) - (\nabla_{\gamma'} \gamma')(f) = \gamma'(\gamma'(f)),
\]

since \( \nabla_{\gamma'} \gamma' = 0 \) because \( \gamma \) is a geodesic, and

\[
\gamma'(\gamma'(f)) = \gamma'(df(\gamma')) = \gamma' \left( \frac{d}{dt} f(\gamma(t)) \right)_{|t=0} = \frac{d^2}{dt^2} f(\gamma(t))_{|t=0}.
\]

Therefore, we have

\[
\text{Hess}(f)_p(u, u) = \left. \frac{d^2}{dt^2} f(\gamma(t)) \right|_{t=0}.
\]

Since the Hessian is a symmetric bilinear form, we obtain \( \text{Hess}(f)_p(u, v) \) by polarization; that is,

\[
\text{Hess}(f)_p(u, v) = \frac{1}{2} (\text{Hess}(f)_p(u + v, u + v) - \text{Hess}(f)_p(u, u) - \text{Hess}(f)_p(v, v)).
\]

Let us find the Hessian of the function \( f : \text{SO}(3) \to \mathbb{R} \) defined in the second example of Section 7.5, with

\[
f(R) = (u^\top R v)^2.
\]

We found that

\[
df_R(X) = 2u^\top X vu^\top R v, \quad X \in R\text{so}(3)
\]

and that the gradient is given by

\[
(\text{grad}(f))_R = u^\top R v R(R^\top vu^\top - vu^\top R).
\]

To compute the Hessian, we use the curve \( \gamma(t) = Re^{tB} \), where \( B \in \text{so}(3) \). Indeed, it can be shown (see Section 17.3, Proposition 17.19) that the metric induced by the inner product

\[
\langle B_1, B_2 \rangle = \text{tr}(B_1^\top B_2) = -\text{tr}(B_1 B_2)
\]
on \( \text{so}(n) \) is bi-invariant, and so the curve \( \gamma \) is a geodesic.

First we compute

\[
(f(\gamma(t)))'(t) = ((u^\top Re^{tB} v)^2)'(t) = 2u^\top Re^{tB} vu^\top R Be^{tB} v,
\]
and then
\[
\text{Hess}(f)_R(RB, RB) = (f(\gamma(t)))''(0) = (2u^\top Re^{\top B} vu^\top RBe^{\top B} v)'(0) = 2u^\top RBvu^\top RBv + 2u^\top Rvu^\top RBBv = 2u^\top RBvu^\top RBv + 2u^\top Rvu^\top RBR^\top RBv.
\]

By polarization, we obtain
\[
\text{Hess}(f)_R(X, Y) = 2u^\top Xvu^\top Yv + u^\top Rvu^\top X R^\top Yv + u^\top Rvu^\top Y R^\top Xv,
\]
with \(X, Y \in Rso(3)\).

### 12.2 The Exponential Map

The idea behind the exponential map is to parametrize a smooth manifold \(M\) locally near any \(p \in M\) in terms of a map from the tangent space \(T_p M\) to the manifold, this map being defined in terms of geodesics.

**Definition 12.4.** Let \(M\) be a smooth manifold equipped with some arbitrary connection. For every \(p \in M\), let \(D(p)\) (or simply, \(D\)) be the open subset of \(T_p M\) given by
\[
D(p) = \{ v \in T_p M \mid \gamma_v(1) \text{ is defined} \},
\]
where \(\gamma_v\) is the unique maximal geodesic with initial conditions \(\gamma_v(0) = p\) and \(\gamma'_v(0) = v\).

The **exponential map** is the map \(\exp_p : D(p) \to M\) given by
\[
\exp_p(v) = \gamma_v(1).
\]

It is easy to see that \(D(p)\) is *star-shaped* (with respect to \(p\)), which means that if \(w \in D(p)\), then the line segment \(\{tw \mid 0 \leq t \leq 1\}\) is contained in \(D(p)\). In view of the fact that if \(\gamma_v: (-\eta, \eta) \to M\) is a geodesic through \(p\) with initial velocity \(v\), then for any \(c \neq 0\),
\[
\gamma_v(ct) = \gamma_{cv}(t), \quad ct \in (-\eta, \eta),
\]
we have
\[
\exp_p(tv) = \gamma_{tv}(1) = \gamma_v(t), \quad t \in D(p),
\]
so the curve
\[
t \mapsto \exp_p(tv), \quad tv \in D(p),
\]
is the geodesic \(\gamma_v\) through \(p\) such that \(\gamma'_v(0) = v\). Such geodesics are called **radial geodesics**.

In a Riemannian manifold with the Levi-Civita connection, the point \(\exp_p(tv)\) is obtained by running along the geodesic \(\gamma_v\) an arc length equal to \(t \|v\|\), starting from \(p\). If the tangent
12.2. THE EXPONENTIAL MAP

Figure 12.1: The image of $v$ under $\exp_p$

vector $tv$ at $p$ is a flexible wire, the exponential map wraps the wire along the geodesic curve without stretching its length. See Figure 12.1.

In general, $\mathcal{D}(p)$ is a proper subset of $T_pM$. For example, if $U$ is a bounded open subset of $\mathbb{R}^n$, since we can identify $T_pU$ with $\mathbb{R}^n$ for all $p \in U$, then $\mathcal{D}(p) \subseteq U$, for all $p \in U$.

**Definition 12.5.** A smooth manifold $M$ equipped with an arbitrary connection is **geodesically complete** iff $\mathcal{D}(p) = T_pM$ for all $p \in M$; that is, the exponential $\exp_p(v)$ is defined for all $p \in M$ and for all $v \in T_pM$.

Equivalently, $(M, g)$ is geodesically complete iff every geodesic can be extended indefinitely.

Geodesically complete Riemannian manifolds (with the Levi-Civita connection) have nice properties, some of which will be investigated later.

Observe that $d(\exp_p)_0 = \text{id}_{T_pM}$. This is because, for every $v \in \mathcal{D}(p)$, the map $t \mapsto \exp_p(tv)$ is the geodesic $\gamma_v$, and

$$\frac{d}{dt}(\gamma_v(t))|_{t=0} = v = \frac{d}{dt}(\exp_p(tv))|_{t=0} = d(\exp_p)_0(v).$$
It follows from the inverse function theorem that \( \exp_p \) is a diffeomorphism from some open ball in \( T_p M \) centered at 0 to \( M \). By using the curve \( t \mapsto (t + 1)v \) passing through \( v \) in \( T_p M \) and with initial velocity \( v \in T_v(T_p M) \approx T_p M \), we get
\[
\left. \frac{d}{dt} (\gamma_v(t + 1)) \right|_{t=0} = \gamma'_v(1).
\]

The following stronger proposition plays a crucial role in the proof of the Hopf-Rinow Theorem; see Theorem 12.14.

**Proposition 12.5.** Let \((M, g)\) be a Riemannian manifold. For every point \( p \in M \), there is an open subset \( W \subseteq M \), with \( p \in W \), and a number \( \epsilon > 0 \), so that:

1. Any two points \( q_1, q_2 \) of \( W \) are joined by a unique geodesic of length \( < \epsilon \).
2. This geodesic depends smoothly upon \( q_1 \) and \( q_2 \); that is, if \( t \mapsto \exp_{q_1}(tv) \) is the geodesic joining \( q_1 \) and \( q_2 \) \((0 \leq t \leq 1)\), then \( v \in T_{q_1} M \) depends smoothly on \((q_1, q_2)\).
3. For every \( q \in W \), the map \( \exp_q \) is a diffeomorphism from the open ball \( B(0, \epsilon) \subseteq T_q M \) to its image \( U_q = \exp_q(B(0, \epsilon)) \subseteq M \), with \( W \subseteq U_q \) and \( U_q \) open.

**Proof.** We follow Milnor [125] (Chapter II, Section 10, Lemma 10.3). Let
\[
U = \{(q, v) \in TM \mid q \in U, v \in T_q M, \|v\| < \epsilon_1 \},
\]
where the open subset \( U \) of \( M \) and \( \epsilon_1 \) are given by Proposition 12.3, for the point \( p_0 = p \in M \). Then, we can define the map \( \Phi : U \to M \times M \) by
\[
\Phi(q, v) = (q, \exp_q(v)).
\]
I claim that \( d_{(p, 0)} \Phi \) is invertible, which implies that \( \Phi \) is a local diffeomorphism near \((p, 0)\). If we pick a chart \((V, \varphi)\) at \( p \), then we have the chart \((V \times V, \varphi \times \varphi)\) at \((p, p) = \Phi(p, 0)\) in \( M \times M \), and since
\[
\left. \frac{d}{dt} (\gamma_v(t + 1)) \right|_{t=0} = \gamma'_v(1),
\]
it is easy to check that in the basis of \( T_p M \times T_p M \) consisting of the pairs \((\frac{\partial}{\partial x_i})_p, (\frac{\partial}{\partial x_j})_p\), the Jacobian matrix of \( d_{(p, 0)} \Phi \) is equal to
\[
\begin{pmatrix}
  I & I \\
  0 & I
\end{pmatrix}.
\]
By the inverse function theorem, there is an open subset \( U' \) contained in \( U \) with \((p, 0) \in U' \) and an open subset \( W' \) of \( M \times M \) containing \((p, p)\) such that \( \Phi \) is a diffeomorphism between \( U' \) and \( W' \). We may assume that there is some open subset \( U' \) of \( U \) containing \( p \) and some \( \epsilon > 0 \) such that \( \epsilon < \epsilon_1 \) and
\[
U' = \{(q, v) \mid q \in U', v \in T_q M, \|v\| < \epsilon \} = \bigcup_{q \in U'} \{q\} \times B(0, \epsilon).
\]
Now, if we choose a smaller open subset $W$ containing $p$ such that $W \times W \subseteq \mathcal{W}$, because $\Phi$ is a diffeomorphism on $\mathcal{U}'$, we have

$$\{q\} \times W \subseteq \Phi(\{q\} \times B(0, \epsilon)),$$

for all $q \in W$. From the definition of $\Phi$, we have $W \subseteq \exp_q(B(0, \epsilon))$, and $\exp_p$ is a diffeomorphism on $B(0, \epsilon) \subseteq T_qM$, which proves part (3).

Given any two points $q_1, q_2 \in W$, since $\Phi$ is a diffeomorphism between $\mathcal{U}'$ and $\mathcal{W}'$ with $W \times W \subseteq \mathcal{W}'$, there is a unique $v \in T_{q_1}M$ such that $\|v\| < \epsilon$ and $\Phi(q_1, v) = (q_1, q_2)$; that is, $\exp_{q_1}(v) = q_2$, which means that $t \mapsto \exp_{q_1}(tv)$ is the unique geodesic from $q_1$ to $q_2$, which proves (1).

Finally, since $(q_1, v) = \Phi^{-1}(q_1, q_2)$ and $\Phi$ is a diffeomorphism, (2) holds.

**Remark:** Except for the part of statement (1) about the length of geodesics having length $< \epsilon$, Proposition 12.5 holds for a Riemannian manifold equipped with an arbitrary connection.

For any $q \in M$, an open neighborhood of $q$ of the form $U_q = \exp_p(B(0, \epsilon))$ where $\exp_p$ is a diffeomorphism from the open ball $B(0, \epsilon)$ onto $U_q$, is called a normal neighborhood.

For the rest of this chapter, we assume that we are dealing with Riemannian manifolds equipped with the Levi-Civita connection.

**Remark:** The proof of the previous proposition can be sharpened to prove that for any $p \in M$, there is some $\beta > 0$ such that any two points $q_1, q_2 \in \exp(B(0, \beta))$, there is a unique geodesic from $q_1$ to $q_2$ that stays within $\exp(B(0, \beta))$; see Do Carmo [60] (Chapter 3, Proposition 4.2). We say that $\exp(B(0, \beta))$ is strongly convex. The least upper bound of these $\beta$ is called the convexity radius at $p$.

**Definition 12.6.** Let $(M, g)$ be a Riemannian manifold. For every point $p \in M$, the injectivity radius of $M$ at $p$, denoted $i(p)$, is the least upper bound of the numbers $r > 0$ such that $\exp_p$ is a diffeomorphism on the open ball $B(0, r) \subseteq T_pM$. The injectivity radius $i(M)$ of $M$ is the greatest lower bound of the numbers $i(p)$, where $p \in M$.

For every $p \in M$, we get a chart $(U_p, \varphi)$, where $U_p = \exp_p(B(0, i(p)))$ and $\varphi = \exp^{-1}$, called a normal chart. If we pick any orthonormal basis $(e_1, \ldots, e_n)$ of $T_pM$, then the $x_i$’s, with $x_i = pr_i \circ \exp^{-1}$ and $pr_i$ the projection onto $\mathbb{R}e_i$, are called normal coordinates at $p$ (here, $n = \dim(M)$). These are defined up to an isometry of $T_pM$. The following proposition shows that Riemannian metrics do not admit any local invariants of order one. The proof is left as an exercise.

**Proposition 12.6.** Let $(M, g)$ be a Riemannian manifold. For every point $p \in M$, in normal coordinates at $p$,

$$g \left( \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right)_p = \delta_{ij} \quad \text{and} \quad \Gamma^k_{ij}(p) = 0.$$
For the next proposition known as Gauss Lemma, we need to define polar coordinates on $T_pM$. If $n = \dim(M)$, observe that the map $(0, \infty) \times S^{n-1} \to T_pM - \{0\}$ given by
\[(r,v) \mapsto rv, \quad r > 0, \quad v \in S^{n-1}\]
is a diffeomorphism, where $S^{n-1}$ is the sphere of radius $r = 1$ in $T_pM$. Then, the map $(0, i(p)) \times S^{n-1} \to U_p - \{p\}$ given by
\[(r,v) \mapsto \exp_p(rv), \quad 0 < r < i(p), \quad v \in S^{n-1}\]
is also a diffeomorphism.

**Proposition 12.7.** (Gauss Lemma) Let $(M,g)$ be a Riemannian manifold. For every point $p \in M$, the images $\exp_p(S(0,r))$ of the spheres $S(0,r) \subseteq T_pM$ centered at 0 by the exponential map $\exp_p$ are orthogonal to the radial geodesics $r \mapsto \exp_p(rv)$ through $p$ for all $r < i(p)$, with $v \in S^{n-1}$. This means that for any differentiable curve $t \mapsto v(t)$ on the unit sphere $S^{n-1}$, the corresponding curve on $M$
\[t \mapsto \exp_p(rv(t)) \quad \text{with } r \text{ fixed},\]
is orthogonal to the radial geodesic
\[r \mapsto \exp_p(rv(t)) \quad \text{with } t \text{ fixed } (0 < r < i(p)).\]

*See Figure 12.2. Furthermore, in polar coordinates, the pull-back metric $\exp^*g$ induced on $T_pM$ is of the form
\[\exp^*g = dr^2 + g_r,\]
where $g_r$ is a metric on the unit sphere $S^{n-1}$, with the property that $g_r/r^2$ converges to the standard metric on $S^{n-1}$ (induced by $\mathbb{R}^n$) when $r$ goes to zero (here, $n = \dim(M)$).

*Proof sketch.* We follow Milnor; see [125], Chapter II, Section 10. Pick any curve $t \mapsto v(t)$ on the unit sphere $S^{n-1}$. The first statement can be restated in terms of the parametrized surface
\[f(r,t) = \exp_p(rv(t));\]
we must prove that
\[\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0,\]
for all $(r,t)$. However, as we are using the Levi-Civita connection, which is compatible with the metric, we have
\[\frac{\partial}{\partial r} \left( \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right) = \left( \frac{D}{\partial r}, \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right) + \left( \frac{\partial f}{\partial r}, \frac{D}{\partial r}, \frac{\partial f}{\partial t} \right).\]
The first expression on the right-hand side of (†) is zero since the curves
\[r \mapsto f(r,t)\]
are geodesics. For the second expression, first observe that
\[
\left\langle \frac{\partial f}{\partial r}, D_t \frac{\partial f}{\partial r} \right\rangle = \frac{1}{2} \frac{\partial}{\partial t} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial r} \right\rangle = 0,
\]
since \(1 = \|v(t)\| = \|\partial f/\partial r\|\), since the velocity vector of a geodesic has constant norm. Next, note that if we can prove that
\[
D_t \frac{\partial f}{\partial r} = D_r \frac{\partial f}{\partial t},
\]
then
\[
0 = \left\langle \frac{\partial f}{\partial r}, D_t \frac{\partial f}{\partial r} \right\rangle = \left\langle \frac{\partial f}{\partial r}, D_r \frac{\partial f}{\partial t} \right\rangle,
\]
so the second expression on the right-hand side of (†) is also zero. The equation
\[
D_t \frac{\partial f}{\partial r} = D_r \frac{\partial f}{\partial t}
\]
follows from the fact that the Levi-Civita connection is torsion-free. The details of the computation are given in Do Carmo [60] (Chapter 3, Lemma 3.4).

Since the right-hand side of (†) is zero,
\[
\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle
\]
is independent of \( r \). But, for \( r = 0 \), we have
\[ f(0,t) = \exp_p(0) = p, \]
hence
\[ \frac{\partial f}{\partial t}(0,t) = 0 \]
and thus,
\[ \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0 \]
for all \( r, t \), which concludes the proof of the first statement.

The orthogonality of \( \frac{\partial f}{\partial r} \) and \( \frac{\partial f}{\partial t} \) implies that the pullback metric \( \exp^* g \) induced on \( T_pM \) is of the form \( \exp^* g = dr^2 + g_r \), where \( g_r \) is a metric on the unit sphere \( S^{n-1} \). For the proof that \( g_r/r^2 \) converges to the standard metric on \( S^{n-1} \), see Pansu’s class notes, Chapter 3, Section 3.5.

Observe that the proof of Gauss Lemma (Proposition 12.7) uses the fact that the connection is compatible with the metric and torsion-free.

**Remark:** if \( v(t) \) is a curve on \( S^{n-1} \) such that \( v(0) = v \) and \( v'(0) = w_N \) (with \( \|v\| < i(p) \)), then
\[ \frac{\partial f}{\partial r}(1,0) = (d\exp_p)_v(v), \quad \frac{\partial f}{\partial t}(1,0) = (d\exp_p)_v(w_N), \]
and Gauss lemma can be stated as
\[ \langle (d\exp_p)_v(v), (d\exp_p)_v(w_N) \rangle = \langle v, w_N \rangle = 0. \]
This is how Gauss lemma is stated in Do Carmo [60] (Chapter 3, Lemma 3.5).

The next three results use the fact that the connection is compatible with the metric and torsion-free. Consider any piecewise smooth curve
\[ \omega: [a,b] \to U_p - \{p\}. \]
We can write each point \( \omega(t) \) uniquely as
\[ \omega(t) = \exp_p(r(t)v(t)), \]
with \( 0 < r(t) < i(p), v(t) \in T_pM \) and \( \|v(t)\| = 1 \).
Proposition 12.8. Let \((M, g)\) be a Riemannian manifold. We have
\[
\int_a^b \|\omega'(t)\| \, dt \geq |r(b) - r(a)|,
\]
where equality holds only if the function \(r\) is monotone and the function \(v\) is constant. Thus, the shortest path joining two concentric spherical shells \(\exp_p(S(0, r_1))\) and \(\exp_p(S(0, r_2))\) is a radial geodesic.

Proof. (After Milnor, see [125], Chapter II, Section 10.) Again, let \(f(r, t) = \exp_p(rv(t))\) so that \(\omega(t) = f(r(t), t)\). Then,
\[
\frac{d\omega}{dt} = \frac{\partial f}{\partial r} r'(t) + \frac{\partial f}{\partial t}.
\]
The proof of the previous proposition showed that the two vectors on the right-hand side are orthogonal and since \(\|\partial f/\partial r\| = 1\), this gives
\[
\left\| \frac{d\omega}{dt} \right\|^2 = |r'(t)|^2 + \left\| \frac{\partial f}{\partial t} \right\|^2 \geq |r'(t)|^2
\]
where equality holds only if \(\partial f/\partial t = 0\); hence only if \(v'(t) = 0\). Thus,
\[
\int_a^b \left\| \frac{d\omega}{dt} \right\| \, dt \geq \int_a^b |r'(t)| \, dt \geq |r(b) - r(a)|
\]
where equality holds only if \(r(t)\) is monotone and \(v(t)\) is constant.

We now get the following important result from Proposition 12.7 and Proposition 12.8, namely that geodesics are locally lengthwise minimizing curves.

Theorem 12.9. Let \((M, g)\) be a Riemannian manifold. Let \(W\) and \(\epsilon\) be as in Proposition 12.5 and let \(\gamma: [0, 1] \rightarrow M\) be the geodesic of length \(< \epsilon\) joining two points \(q_1, q_2\) of \(W\). For any other piecewise smooth path \(\omega\) joining \(q_1\) and \(q_2\), we have
\[
\int_0^1 \|\gamma'(t)\| \, dt \leq \int_0^1 \|\omega'(t)\| \, dt,
\]
where equality holds only if the images \(\omega([0, 1])\) and \(\gamma([0, 1])\) coincide. Thus, \(\gamma\) is the shortest path from \(q_1\) to \(q_2\).

Proof. (After Milnor, see [125], Chapter II, Section 10.) Consider any piecewise smooth path \(\omega\) from \(q_1 = \gamma(0)\) to some point
\[
q_2 = \exp_{q_1}(rv) \in U_{q_1},
\]
where \(0 < r < \epsilon\) and \(\|v\| = 1\). Then, for any \(\delta\) with \(0 < \delta < r\), the path \(\omega\) must contain a segment joining the spherical shell of radius \(\delta\) to the spherical shell of radius \(r\), and lying between these two shells. The length of this segment will be at least \(r - \delta\); hence if we let \(\delta\) go to zero, the length of \(\omega\) will be at least \(r\). If \(\omega([0, 1]) \neq \gamma([0, 1])\), we easily obtain a strict inequality. \(\square\)
Here is an important consequence of Theorem 12.9.

**Corollary 12.10.** Let $(M, g)$ be a Riemannian manifold. If $\omega: [0, b] \to M$ is any curve parametrized by arc-length and $\omega$ has length less than or equal to the length of any other curve from $\omega(0)$ to $\omega(b)$, then $\omega$ is a geodesic.

Proof. Consider any segment of $\omega$ lying within an open set $W$ as above, and having length $< \epsilon$. By Theorem 12.9, this segment must be a geodesic. Hence, the entire curve is a geodesic. □

Corollary 12.10 together with the fact that isometries preserve geodesics can be used to determine the geodesics in various spaces, for example in the Poincaré half-plane.

**Definition 12.7.** Let $(M, g)$ be a Riemannian manifold. A geodesic $\gamma: [a, b] \to M$ is minimal iff its length is less than or equal to the length of any other piecewise smooth curve joining its endpoints.

Theorem 12.9 asserts that any sufficiently small segment of a geodesic is minimal. On the other hand, a long geodesic may not be minimal. For example, a great circle arc on the unit sphere is a geodesic. If such an arc has length greater than $\pi$, then it is not minimal. Minimal geodesics are generally not unique. For example, any two antipodal points on a sphere are joined by an infinite number of minimal geodesics.

A **broken geodesic** is a piecewise smooth curve as in Definition 12.1, where each curve segment is a geodesic.

**Proposition 12.11.** A Riemannian manifold $(M, g)$ is connected iff any two points of $M$ can be joined by a broken geodesic.

Proof. Assume $M$ is connected, pick any $p \in M$, and let $S_p \subseteq M$ be the set of all points that can be connected to $p$ by a broken geodesic. For any $q \in M$, choose a normal neighborhood $U$ of $q$. If $q \in S_p$, then it is clear that $U \subseteq S_p$. On the other hand, if $q \notin S_p$, then $U \subseteq M - S_p$. Therefore, $S_p \neq \emptyset$ is open and closed, so $S_p = M$. The converse is obvious. □

**Remark:** Proposition 12.11 holds for a smooth manifold equipped with any connection.

In general, if $M$ is connected, then it is not true that any two points are joined by a geodesic. However, this will be the case if $M$ is a geodesically complete Riemannian manifold equipped with the Levi-Civita connection, as we will see in the next section.

Next we will see that a Riemannian metric induces a distance on the manifold whose induced topology agrees with the original metric.
12.3 Complete Riemannian Manifolds, the Hopf-Rinow Theorem and the Cut Locus

Every connected Riemannian manifold \((M, g)\) is a metric space in a natural way. Furthermore, \(M\) is a complete metric space iff \(M\) is geodesically complete. In this section, we explore briefly some properties of complete Riemannian manifolds equipped with the Levi-Civita connection.

Proposition 12.12. Let \((M, g)\) be a connected Riemannian manifold. For any two points \(p, q \in M\), let \(d(p, q)\) be the greatest lower bound of the lengths of all piecewise smooth curves joining \(p\) to \(q\). Then, \(d\) is a metric on \(M\) and the topology of the metric space \((M, d)\) coincides with the original topology of \(M\).

A proof of the above proposition can be found in Gallot, Hulin and Lafontaine [73] (Chapter 2, Proposition 2.91) or O'Neill [138] (Chapter 5, Proposition 18).

The distance \(d\) is often called the Riemannian distance on \(M\). For any \(p \in M\) and any \(\epsilon > 0\), the metric ball of center \(p\) and radius \(\epsilon\) is the subset \(B_\epsilon(p) \subseteq M\) given by

\[
B_\epsilon(p) = \{ q \in M \mid d(p, q) < \epsilon \}.
\]

The next proposition follows easily from Proposition 12.5 (Milnor [125], Section 10, Corollary 10.8).

Proposition 12.13. Let \((M, g)\) be a connected Riemannian manifold. For any compact subset \(K \subseteq M\), there is a number \(\delta > 0\) so that any two points \(p, q \in K\) with distance \(d(p, q) < \delta\) are joined by a unique geodesic of length less than \(\delta\). Furthermore, this geodesic is minimal and depends smoothly on its endpoints.

Recall from Definition 12.5 that \((M, g)\) is geodesically complete iff the exponential map \(v \mapsto \exp_p(v)\) is defined for all \(p \in M\) and for all \(v \in T_pM\). We now prove the following important theorem due to Hopf and Rinow (1931):

Theorem 12.14. (Hopf-Rinow) Let \((M, g)\) be a connected Riemannian manifold. If there is a point \(p \in M\) such that \(\exp_p\) is defined on the entire tangent space \(T_pM\), then any point \(q \in M\) can be joined to \(p\) by a minimal geodesic. As a consequence, if \(M\) is geodesically complete, then any two points of \(M\) can be joined by a minimal geodesic.

Proof. We follow Milnor’s proof in [125], Chapter 10, Theorem 10.9. Pick any two points \(p, q \in M\) and let \(r = d(p, q)\). By Proposition 12.5, there is some \(\epsilon > 0\), such that the exponential map is a diffeomorphism between the open ball \(B(0, \epsilon)\) and its image \(U_p = \exp_p(B(0, \epsilon))\). For \(\delta < \epsilon\), let \(S = \exp_p(S(0, \delta))\), where \(S(0, \delta)\) is the sphere of radius \(\delta\). Since \(S \subseteq U_p\) is compact, there is some point

\[
p_0 = \exp_p(\delta v), \quad \text{with} \quad \|v\| = 1,
\]
on \( S \) for which the distance to \( q \) is minimized. We will prove that
\[
\exp_p(rv) = q,
\]
which will imply that the geodesic \( \gamma \) given by \( \gamma(t) = \exp_p(tv) \) is actually a minimal geodesic from \( p \) to \( q \) (with \( t \in [0, r] \)). Here we use the fact that the exponential \( \exp_p \) is defined everywhere on \( T_p M \). See Figure 12.3.

\[ \begin{align*}
\text{Figure 12.3: An illustration of the preceding paragraph} \\
\end{align*} \]

The proof amounts to showing that a point which moves along the geodesic \( \gamma \) must get closer and closer to \( q \). In fact, for each \( t \in [\delta, r] \), we prove
\[
d(\gamma(t), q) = r - t. \tag{\ast_t}
\]
We get the proof by setting \( t = r \).

First, we prove (\ast_\delta). Since every path from \( p \) to \( q \) must pass through \( S \), by the choice of \( p_0 \), we have
\[
r = d(p, q) = \min_{s \in S} \{ d(p, s) + d(s, q) \} = \delta + d(p_0, q).\]
Therefore, \( d(p_0, q) = r - \delta \), and since \( p_0 = \gamma(\delta) \), this proves \((\ast_\delta)\).

Define \( t_0 \in [\delta, r] \) by

\[
t_0 = \sup \{ t \in [\delta, r] \mid d(\gamma(t), q) = r - t \}.
\]

As the set \( \{ t \in [\delta, r] \mid d(\gamma(t), q) = r - t \} \) is closed, it contains its upper bound \( t_0 \), so the equation \((\ast_{t_0})\) also holds. We claim that if \( t_0 < r \), then we obtain a contradiction.

As we did with \( p \), there is some small \( \delta' > 0 \) so that if \( S' = \text{exp}_{\gamma(t_0)}(B(0, \delta')) \), then there is some point \( p'_0 \) on \( S' \) with minimum distance from \( q \) and \( p'_0 \) is joined to \( \gamma(t_0) \) by a minimal geodesic. See Figure 12.4.

![Figure 12.4: An illustration of the preceding paragraph](image)

We have

\[
r - t_0 = d(\gamma(t_0), q) = \min_{s \in S'} \{ d(\gamma(t_0), s) + d(s, q) \} = \delta' + d(p'_0, q),
\]

hence

\[
\begin{align*}
    d(p'_0, q) &= r - t_0 - \delta'.
\end{align*}
\]

We claim that \( p'_0 = \gamma(t_0 + \delta') \).

By the triangle inequality and using \((\dagger)\) (recall that \( d(p, q) = r \)), we have

\[
d(p, p'_0) \geq d(p, q) - d(p'_0, q) = t_0 + \delta'.
\]

But, a path of length precisely \( t_0 + \delta' \) from \( p \) to \( p'_0 \) is obtained by following \( \gamma \) from \( p \) to \( \gamma(t_0) \), and then following a minimal geodesic from \( \gamma(t_0) \) to \( p'_0 \). Since this broken geodesic has minimal length, by Corollary 12.10, it is a genuine (unbroken) geodesic, and so it coincides with \( \gamma \). But then, as \( p'_0 = \gamma(t_0 + \delta') \), equality \((\dagger)\) becomes \((\ast_{t_0+\delta'})\), namely

\[
    d(\gamma(t_0 + \delta'), q) = r - (t_0 + \delta'),
\]

contradicting the maximality of \( t_0 \). Therefore, we must have \( t_0 = r \), and \( q = \exp_p(rv) \), as desired. \( \square \)
Remark: Theorem 12.14 is proved in nearly every book on Riemannian geometry. Among those, we mention Gallot, Hulin and Lafontaine [73] (Chapter 2, Theorem 2.103) and O’Neill [138] (Chapter 5, Lemma 24). Since the proof of Theorem 12.14 makes crucial use of Corollary 12.10, which itself relies on the fact that the connection is symmetric and torsion-free, Theorem 12.14 only holds for the Levi-Civita connection.

Theorem 12.14 implies the following result (often known as the Hopf-Rinow Theorem):

**Theorem 12.15.** Let \((M,g)\) be a connected, Riemannian manifold. The following statements are equivalent:

1. The manifold \((M,g)\) is geodesically complete; that is, for every \(p \in M\), every geodesic through \(p\) can be extended to a geodesic defined on all of \(\mathbb{R}\).

2. For every point \(p \in M\), the map \(\exp_p\) is defined on the entire tangent space \(T_pM\).

3. There is a point \(p \in M\), such that \(\exp_p\) is defined on the entire tangent space \(T_pM\).

4. Any closed and bounded subset of the metric space \((M,d)\) is compact.

5. The metric space \((M,d)\) is complete (that is, every Cauchy sequence converges).

Proofs of Theorem 12.15 can be found in Gallot, Hulin and Lafontaine [73] (Chapter 2, Corollary 2.105) and O’Neill [138] (Chapter 5, Theorem 21).

In view of Theorem 12.15, a connected Riemannian manifold \((M,g)\) is geodesically complete iff the metric space \((M,d)\) is complete. We will refer simply to \(M\) as a *complete Riemannian manifold* (it is understood that \(M\) is connected). Also, by (4), every compact, Riemannian manifold is complete. If we remove any point \(p\) from a Riemannian manifold \(M\), then \(M - \{p\}\) is not complete, since every geodesic that formerly went through \(p\) yields a geodesic that can’t be extended.

Assume \((M,g)\) is a complete Riemannian manifold. Given any point \(p \in M\), it is interesting to consider the subset \(U_p \subseteq T_pM\) consisting of all \(v \in T_pM\) such that the geodesic

\[
t \mapsto \exp_p(tv)
\]

is a minimal geodesic up to \(t = 1 + \epsilon\), for some \(\epsilon > 0\). The subset \(U_p\) is open and star-shaped, and it turns out that \(\exp_p\) is a diffeomorphism from \(U_p\) onto its image \(\exp_p(U_p)\) in \(M\). The left-over part \(M - \exp_p(U_p)\) (if nonempty) is actually equal to \(\exp_p(\partial U_p)\), and it is an important subset of \(M\) called the *cut locus of \(p\)*. The following proposition is needed to establish properties of the cut locus:

**Proposition 12.16.** Let \((M,g)\) be a complete Riemannian manifold. For any geodesic \(\gamma: [0,a] \to M\) from \(p = \gamma(0)\) to \(q = \gamma(a)\), the following properties hold:

1. If there is no geodesic shorter than \(\gamma\) between \(p\) and \(q\), then \(\gamma\) is minimal on \([0,a]\).
(ii) If there is another geodesic of the same length as $\gamma$ between $p$ and $q$, then $\gamma$ is no longer minimal on any larger interval, $[0, a + \varepsilon]$.

(iii) If $\gamma$ is minimal on any interval $I$, then $\gamma$ is also minimal on any subinterval of $I$.

Proof. Part (iii) is an immediate consequence of the triangle inequality. As $M$ is complete, by the Hopf-Rinow Theorem, there is a minimal geodesic from $p$ to $q$, so $\gamma$ must be minimal too. This proves part (i). For part (ii), assume that $\omega$ is another geodesic from $p$ to $q$ of the same length as $\gamma$ and that $\gamma$ is defined in $[0, a + \varepsilon]$ some some $\varepsilon > 0$. Since $\gamma$ and $\omega$ are assumed to be distinct curves, the curve $\varphi: [0, a + \varepsilon] \to M$ given by

$$
\varphi(t) = \begin{cases} 
\omega(t) & 0 \leq t \leq a \\
\gamma(t) & a \leq t \leq a + \varepsilon
\end{cases}
$$

is not smooth at $t = a$, since otherwise $\gamma$ and $\omega$ would be equal on their common domain, since Proposition 12.1 implies there is a unique geodesic through $q$ with initial condition $v = \gamma'(a) = \omega'(a)$. Pick $\varepsilon'$ so that $0 < \varepsilon' < \min \{\varepsilon, a\}$, and consider the points $q_1 = \varphi(a - \varepsilon')$ and $q_2 = \varphi(a + \varepsilon')$. By Hopf-Rinow’s theorem, there is a minimal geodesic $\psi$ from $q_1$ to $q_2$, and since the portion of $\varphi$ from $q_1$ to $q_2$ is not smooth, the length of $\psi$ is strictly smaller than the length of the segment of $\varphi$ from $q_1$ to $q_2$. But then, the curve $\tilde{\varphi}$ obtained by concatenating the segment of $\omega$ from $p$ to $q_1$ and $\psi$ from $q_1$ to $q_2$ is strictly shorter that the curve obtained by concatenating the curve segment $\omega$ from $p$ to $q$ with the curve segment $\gamma$ from $q$ to $q_2$. See Figure 12.5.

Figure 12.5: The geodesics $\omega$, $\gamma$, $\psi$, and the path $\tilde{\varphi}$

However, the length of the curve segment $\omega$ from $p$ to $q$ is equal to length of the curve segment $\gamma$ from $p$ to $q$. This proves that $\tilde{\varphi}$ from $p$ to $q_2$ is strictly shorter than $\gamma$ from $p$ to $q_2$, so $\gamma$ is no longer minimal beyond $q$. \qed
Again, assume \((M, g)\) is a complete Riemannian manifold and let \(p \in M\) be any point. For every \(v \in T_pM\), let
\[
I_v = \{ s \in \mathbb{R} \cup \{\infty\} \mid \text{the geodesic } t \mapsto \exp_p(tv) \text{ is minimal on } [0, s] \}.
\]
It is easy to see that \(I_v\) is a closed interval, so \(I_v = [0, \rho(v)]\) (with \(\rho(v)\) possibly infinite). It can be shown that if \(w = \lambda v\), then \(\rho(v) = \lambda \rho(w)\), so we can restrict our attention to unit vectors \(v\). It can also be shown that the map \(\rho: S^{n-1} \to \mathbb{R}\) is continuous, where \(S^{n-1}\) is the unit sphere of center 0 in \(T_pM\), and that \(\rho(v)\) is bounded below by a strictly positive number.

**Definition 12.8.** Let \((M, g)\) be a complete Riemannian manifold and let \(p \in M\) be any point. Define \(U_p\) by
\[
U_p = \{ v \in T_pM \mid \rho \left(\frac{v}{\|v\|}\right) > \|v\| \} = \{ v \in T_pM \mid \rho(v) > 1 \},
\]
and the cut locus of \(p\) by
\[
\text{Cut}(p) = \exp_p(\partial U_p) = \{ \exp_p(\rho(v)v) \mid v \in S^{n-1} \}.
\]

The set \(U_p\) is open and star-shaped. The boundary \(\partial U_p\) of \(U_p\) in \(T_pM\) is sometimes called the tangential cut locus of \(p\) and is denoted \(\tilde{\text{Cut}}(p)\).

**Remark:** The cut locus was first introduced for convex surfaces by Poincaré (1905) under the name *ligne de partage*. According to Do Carmo [60] (Chapter 13, Section 2), for Riemannian manifolds, the cut locus was introduced by J.H.C. Whitehead (1935). But it was Klingenberg (1959) who revived the interest in the cut locus and showed its usefulness.

**Proposition 12.17.** Let \((M, g)\) be a complete Riemannian manifold. For any point \(p \in M\), the sets \(\exp_p(U_p)\) and \(\text{Cut}(p)\) are disjoint and
\[
M = \exp_p(U_p) \cup \text{Cut}(p).
\]

**Proof.** From the Hopf-Rinow Theorem, for every \(q \in M\), there is a minimal geodesic \(t \mapsto \exp_p(vt)\) such that \(\exp_p(v) = q\). This shows that \(\rho(v) \geq 1\), so \(v \in U_p\) and
\[
M = \exp_p(U_p) \cup \text{Cut}(p).
\]
It remains to show that this is a disjoint union. Assume \(q \in \exp_p(U_p) \cap \text{Cut}(p)\). Since \(q \in \exp_p(U_p)\), there is a geodesic \(\gamma\) such that \(\gamma(0) = p, \gamma(a) = q, \) and \(\gamma\) is minimal on \([0, a + \epsilon]\), for some \(\epsilon > 0\). On the other hand, as \(q \in \text{Cut}(p)\), there is some geodesic \(\tilde{\gamma}\) with \(\tilde{\gamma}(0) = p, \tilde{\gamma}(b) = q, \) \(\tilde{\gamma}\) minimal on \([0, b]\), but \(\tilde{\gamma}\) not minimal after \(b\). As \(\gamma\) and \(\tilde{\gamma}\) are both minimal from \(p\) to \(q\), they have the same length from \(p\) to \(q\). But then, as \(\gamma\) and \(\tilde{\gamma}\) are distinct, by Proposition 12.16 (ii), the geodesic \(\gamma\) can’t be minimal after \(q\), a contradiction. \(\square\)
Observe that the injectivity radius \( i(p) \) of \( M \) at \( p \) is equal to the distance from \( p \) to the cut locus of \( p \):
\[
i(p) = d(p, \text{Cut}(p)) = \inf_{q \in \text{Cut}(p)} d(p, q).
\]
Consequently, the injectivity radius \( i(M) \) of \( M \) is given by
\[
i(M) = \inf_{p \in M} d(p, \text{Cut}(p)).
\]
If \( M \) is compact, it can be shown that \( i(M) > 0 \). It can also be shown using Jacobi fields that \( \exp_p \) is a diffeomorphism from \( \mathcal{U}_p \) onto its image \( \exp_p(\mathcal{U}_p) \). Thus, \( \exp_p(\mathcal{U}_p) \) is diffeomorphic to an open ball in \( \mathbb{R}^n \) (where \( n = \dim(M) \)) and the cut locus is closed. Hence, the manifold \( M \) is obtained by gluing together an open \( n \)-ball onto the cut locus of a point. In some sense the topology of \( M \) is “contained” in its cut locus.

Given any sphere \( S^{n-1} \), the cut locus of any point \( p \) is its antipodal point \( \{-p\} \). For more examples, consult Gallot, Hulin and Lafontaine [73] (Chapter 2, Section 2C7), Do Carmo [60] (Chapter 13, Section 2) or Berger [19] (Chapter 6). In general, the cut locus is very hard to compute. In fact, even for an ellipsoid, the determination of the cut locus of an arbitrary point was a matter of conjecture for a long time. This conjecture was finally settled around 2011.

### 12.4 Convexity, Convexity Radius

Proposition 12.5 shows that if \( (M, g) \) is a Riemannian manifold, then for every point \( p \in M \), there is an open subset \( W \subseteq M \) with \( p \in W \) and a number \( \epsilon > 0 \), so that any two points \( q_1, q_2 \) of \( W \) are joined by a unique geodesic of length < \( \epsilon \). However, there is no guarantee that this unique geodesic between \( q_1 \) and \( q_2 \) stays inside \( W \). Intuitively this says that \( W \) may not be convex.

The notion of convexity can be generalized to Riemannian manifolds, but there are some subtleties. In this short section we review various definition of convexity found in the literature and state one basic result. Following Sakai [150] (Chapter IV, Section 5), we make the following definition:

**Definition 12.9.** Let \( C \subseteq M \) be a nonempty subset of some Riemannian manifold \( M \).

1. The set \( C \) is called **strongly convex** iff for any two points \( p, q \in C \), there exists a unique minimal geodesic \( \gamma \) from \( p \) to \( q \) in \( M \) and \( \gamma \) is contained in \( C \).

2. If for every point \( p \in \mathcal{C} \), there is some \( \epsilon(p) > 0 \) so that \( C \cap B_{\epsilon(p)}(p) \) is strongly convex, then we say that \( C \) is **locally convex** (where \( B_{\epsilon(p)}(p) \) is the metric ball of center 0 and radius \( \epsilon(p) \)).

3. The set \( C \) is called **totally convex** iff for any two points \( p, q \in C \), all geodesics from \( p \) to \( q \) in \( M \) are contained in \( C \).
It is clear that if $C$ is strongly convex or totally convex, then $C$ is locally convex. If $M$ is complete and any two points are joined by a unique geodesic, then the three conditions of Definition 12.9 are equivalent. The next Proposition will show that a metric ball with sufficiently small radius is strongly convex.

**Definition 12.10.** For any $p \in M$, the **convexity radius at** $p$, denoted $r(p)$, is the least upper bound of the numbers $r > 0$ such that for any metric ball $B_r(q)$, if $B_r(q) \subseteq B_r(p)$, then $B_r(q)$ is strongly convex and every geodesic contained in $B_r(p)$ is a minimal geodesic joining its endpoints. The **convexity radius of** $M$ $r(M)$ is the greatest lower bound of the set $\{r(p) \mid p \in M\}$.

Note that it is possible that $r(M) = 0$ if $M$ is not compact.

The following proposition is proved in Sakai [150] (Chapter IV, Section 5, Theorem 5.3).

**Proposition 12.18.** If $M$ is a Riemannian manifold, then $r(p) > 0$ for every $p \in M$, and the map $p \mapsto r(p) \in \mathbb{R}_+ \cup \{\infty\}$ is continuous. Furthermore, if $r(p) = \infty$ for some $p \in M$, then $r(q) = \infty$ for all $q \in M$.

That $r(p) > 0$ is also proved in Do Carmo [60] (Chapter 3, Section 4, Proposition 4.2). More can be said about the structure of connected locally convex subsets of $M$; see Sakai [150] (Chapter IV, Section 5).

**Remark:** The following facts are stated in Berger [19] (Chapter 6):

(1) If $M$ is compact, then the convexity radius $r(M)$ is strictly positive.

(2) $r(M) \leq \frac{1}{2} i(M)$, where $i(M)$ is the injectivity radius of $M$.

Berger also points out that if $M$ is compact, then the existence of a finite cover by convex balls can used to triangulate $M$. This method was proposed by Hermann Karcher (see Berger [19], Chapter 3, Note 3.4.5.3).

### 12.5 The Calculus of Variations Applied to Geodesics; The First Variation Formula

In this section, we consider a Riemannian manifold $(M, g)$ equipped with the Levi-Civita connection. The path space $\Omega(p, q)$ was introduced in Definition 12.1. It is an “infinite dimensional” manifold. By analogy with finite dimensional manifolds, we define a kind of tangent space to $\Omega(p, q)$ at a “point” $\omega$. In this section, it is convenient to assume that paths in $\Omega(p, q)$ are parametrized over the interval $[0, 1]$. 
12.5. THE CALCULUS OF VARIATIONS APPLIED TO GEODESICS

Figure 12.6: The point $\omega$ in $\Omega(p, q)$ and its associated tangent vector, the blue vector field. Each blue vector is contained in a tangent space for $\omega(t)$.

**Definition 12.11.** For every “point” $\omega \in \Omega(p, q)$, we define the “tangent space” $T_\omega \Omega(p, q)$ to $\Omega(p, q)$ at $\omega$, as the space of all piecewise smooth vector fields $W$ along $\omega$, for which $W(0) = W(1) = 0$. See Figure 12.6.

Now, if $F: \Omega(p, q) \to \mathbb{R}$ is a real-valued function on $\Omega(p, q)$, it is natural to ask what the induced “tangent map”

$$dF_\omega: T_\omega \Omega(p, q) \to \mathbb{R},$$

should mean (here, we are identifying $T_F(\omega)\mathbb{R}$ with $\mathbb{R}$). Observe that $\Omega(p, q)$ is not even a topological space so the answer is far from obvious!

In the case where $f: M \to \mathbb{R}$ is a function on a manifold, there are various equivalent ways to define $df$, one of which involves curves. For every $v \in T_p M$, if $\alpha: (-\epsilon, \epsilon) \to M$ is a curve such that $\alpha(0) = p$ and $\alpha'(0) = v$, then we know that

$$df_p(v) = \frac{d(f(\alpha(t))))}{dt} \bigg|_{t=0}.$$  

We may think of $\alpha$ as a small variation of $p$. Recall that $p$ is a critical point of $f$ iff $df_p(v) = 0$, for all $v \in T_p M$.

Rather than attempting to define $dF_\omega$ (which requires some conditions on $F$), we will mimic what we did with functions on manifolds and define what is a critical path of a function $F: \Omega(p, q) \to \mathbb{R}$, using the notion of variation. Now, geodesics from $p$ to $q$ are special paths in $\Omega(p, q)$, and they turn out to be the critical paths of the energy function

$$E^b_a(\omega) = \int_a^b \|\omega'(t)\|^2 dt,$$

where $\omega \in \Omega(p, q)$, and $0 \leq a < b \leq 1$. 
Definition 12.12. Given any path \( \omega \in \Omega(p,q) \), a variation of \( \omega \) (keeping endpoints fixed) is a function \( \tilde{\alpha}: (-\epsilon, \epsilon) \to \Omega(p,q) \), for some \( \epsilon > 0 \), such that:

1. \( \tilde{\alpha}(0) = \omega \)

2. There is a subdivision \( 0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1 \) of \([0, 1]\) so that the map

\[
\alpha: (-\epsilon, \epsilon) \times [0, 1] \to M
\]

defined by \( \alpha(u, t) = \tilde{\alpha}(u)(t) \) is smooth on each strip \((-\epsilon, \epsilon) \times [t_i, t_{i+1}]\), for \( i = 0, \ldots, k-1 \).

If \( U \) is an open subset of \( \mathbb{R}^n \) containing the origin and if we replace \((-\epsilon, \epsilon)\) by \( U \) in the above, then \( \tilde{\alpha}: U \to \Omega(p,q) \) is called an \( n \)-parameter variation of \( \omega \).

The function \( \alpha \) is also called a variation of \( \omega \). Since each \( \tilde{\alpha}(u) \) belongs to \( \Omega(p,q) \), note that

\[
\alpha(u, 0) = p, \quad \alpha(u, 1) = q, \quad \text{for all } u \in (-\epsilon, \epsilon).
\]

The function \( \tilde{\alpha} \) may be considered as a “smooth path” in \( \Omega(p,q) \), since for every \( u \in (-\epsilon, \epsilon) \), the map \( \tilde{\alpha}(u) \) is a curve in \( \Omega(p,q) \) called a curve in the variation (or longitudinal curve of the variation).

The “tangent vector” \( \frac{d\tilde{\alpha}}{du}(0) \in T_\omega \Omega(p,q) \) is defined to be the vector field \( W \) along \( \omega \) given by

\[
W_t = \left. \frac{\partial \alpha}{\partial u}(u, t) \right|_{u=0}.
\]

By definition,

\[
\frac{d\tilde{\alpha}}{du}(0)_t = W_t, \quad t \in [0, 1].
\]

Clearly, \( W \in T_\omega \Omega(p,q) \). In particular, \( W(0) = W(1) = 0 \). The vector field \( W \) is also called the variation vector field associated with the variation \( \alpha \). See Figure 12.7.

Besides the curves in the variation \( \tilde{\alpha}(u) \) (with \( u \in (-\epsilon, \epsilon) \)), for every \( t \in [0, 1] \), we have a curve \( \alpha_t: (-\epsilon, \epsilon) \to M \), called a transversal curve of the variation, defined by

\[
\alpha_t(u) = \tilde{\alpha}(u)(t),
\]

and \( W_t \) is equal to the velocity vector \( \alpha'_t(0) \) at the point \( \omega(t) = \alpha_t(0) \). For \( \epsilon \) sufficiently small, the vector field \( W_t \) is an infinitesimal model of the variation \( \tilde{\alpha} \).

We can show that for any \( W \in T_\omega \Omega(p,q) \), there is a variation \( \tilde{\alpha}: (-\epsilon, \epsilon) \to \Omega(p,q) \) which satisfies the conditions

\[
\tilde{\alpha}(0) = \omega, \quad \frac{d\tilde{\alpha}}{du}(0) = W.
\]
Figure 12.7: A variation of $\omega$ in $\mathbb{R}^2$ with transversal curve $\alpha_t(u)$. The blue vector field is the variational vector field $W_t$.

**Sketch of the proof.** By the compactness of $\omega([0,1])$, it is possible to find a $\delta > 0$ so that $\exp_{\omega(t)}$ is defined for all $t \in [0,1]$ and all $v \in T_{\omega(t)}M$, with $\|v\| < \delta$. Then, if

$$N = \max_{t \in [0,1]} \|W_t\|,$$

for any $\epsilon$ such that $0 < \epsilon < \frac{\delta}{N}$, it can be shown that

$$\tilde{\alpha}(u)(t) = \exp_{\omega(t)}(uW_t)$$

works (for details, see Do Carmo [60], Chapter 9, Proposition 2.2).

As we said earlier, given a function $F: \Omega(p,q) \to \mathbb{R}$, we do not attempt to define the differential $dF_\omega$, but instead the notion of critical path.

**Definition 12.13.** Given a function $F: \Omega(p,q) \to \mathbb{R}$, we say that a path $\omega \in \Omega(p,q)$ is a critical path for $F$ iff

$$\left.\frac{dF(\tilde{\alpha}(u))}{du}\right|_{u=0} = 0,$$

for every variation $\tilde{\alpha}$ of $\omega$ (which implies that the derivative $\left.\frac{dF(\tilde{\alpha}(u))}{du}\right|_{u=0}$ is defined for every variation $\tilde{\alpha}$ of $\omega$).

For example, if $F$ takes on its minimum on a path $\omega_0$ and if the derivatives $\frac{dF(\tilde{\alpha}(u))}{du}$ are all defined, then $\omega_0$ is a critical path of $F$.

We will apply the above to two functions defined on $\Omega(p,q)$:
(1) The energy function (also called action integral)

\[ E^b_a(\omega) = \int_a^b \|\omega'(t)\|^2 \, dt. \]

(We write \( E = E^1_0 \).)

(2) The arc-length function

\[ L^b_a(\omega) = \int_a^b \|\omega'(t)\| \, dt. \]

The quantities \( E^b_a(\omega) \) and \( L^b_a(\omega) \) can be compared as follows: if we apply the Cauchy-Schwarz's inequality,

\[ \left( \int_a^b f(t)g(t)dt \right)^2 \leq \left( \int_a^b f^2(t)dt \right) \left( \int_a^b g^2(t)dt \right) \]

with \( f(t) \equiv 1 \) and \( g(t) = \|\omega'(t)\| \), we get

\[ (L^b_a(\omega))^2 \leq (b - a)E^b_a, \]

where equality holds iff \( g \) is constant; that is, iff the parameter \( t \) is proportional to arc-length.

Now suppose that there exists a minimal geodesic \( \gamma \) from \( p \) to \( q \). Then,

\[ E(\gamma) = L(\gamma)^2 \leq L(\omega)^2 \leq E(\omega), \]

where the equality \( L(\gamma)^2 = L(\omega)^2 \) holds only if \( \omega \) is also a minimal geodesic, possibly reparametrized. On the other hand, the equality \( L(\omega) = E(\omega)^2 \) can hold only if the parameter is proportional to arc-length along \( \omega \). This proves that \( E(\gamma) < E(\omega) \) unless \( \omega \) is also a minimal geodesic. We just proved:

**Proposition 12.19.** Let \((M, g)\) be a complete Riemannian manifold. For any two points \( p, q \in M \), if \( d(p, q) = \delta \), then the energy function \( E: \Omega(p, q) \to \mathbb{R} \) takes on its minimum \( \delta^2 \) precisely on the set of minimal geodesics from \( p \) to \( q \).

Next, we are going to show that the critical paths of the energy function are exactly the geodesics. For this we need the first variation formula.

Let \( \tilde{\alpha}: (-\epsilon, \epsilon) \to \Omega(p, q) \) be a variation of \( \omega \), and let

\[ W_t = \left. \frac{\partial \alpha}{\partial u} (u, t) \right|_{u=0} \]

be its associated variation vector field. Furthermore, let

\[ V_t = \frac{d\omega}{dt} = \omega'(t), \]
12.5. THE CALCULUS OF VARIATIONS APPLIED TO GEODESICS

The velocity vector field of \( \omega \), and

\[
\Delta_t V = V_{t_+} - V_{t_-},
\]

the discontinuity in the velocity vector at \( t \), which is nonzero only for \( t = t_i \), with \( 0 < t_i < 1 \)
(see the definition of \( \gamma'(t_i)_+ \) and \( \gamma'(t_i)_- \) just after Definition 12.1). See Figure 12.8.

**Theorem 12.20. (First Variation Formula)** For any path \( \omega \in \Omega(p,q) \), we have

\[
\frac{1}{2} \left. \frac{dE(\bar{\omega}(u))}{du} \right|_{u=0} = -\sum_i \langle W_t, \Delta_t V \rangle - \int_0^1 \left\langle W_t, \frac{D}{dt} V_t \right\rangle dt,
\]

where \( \bar{\omega} : (-\epsilon, \epsilon) \to \Omega(p,q) \) is any variation of \( \omega \).

**Proof.** (After Milnor, see [125], Chapter II, Section 12, Theorem 12.2.) By Proposition 11.10, we have

\[
\frac{\partial}{\partial u} \left( \frac{\partial \alpha}{\partial t} \cdot \frac{\partial \alpha}{\partial t} \right) = 2 \left( \frac{D}{dt} \frac{\partial \alpha}{\partial u} \cdot \frac{\partial \alpha}{\partial t} \right).
\]

Therefore,

\[
\frac{dE(\bar{\omega}(u))}{du} = \frac{d}{du} \int_0^1 \left( \frac{\partial \alpha}{\partial t} \cdot \frac{\partial \alpha}{\partial t} \right) dt = 2 \int_0^1 \left( \frac{D}{dt} \frac{\partial \alpha}{\partial u} \cdot \frac{\partial \alpha}{\partial t} \right) dt.
\]

Now, because we are using the Levi-Civita connection, which is torsion-free, it is not hard to prove that

\[
\frac{D}{dt} \frac{\partial \alpha}{\partial u} = \frac{D}{\partial u} \frac{\partial \alpha}{\partial t},
\]

so

\[
\frac{dE(\bar{\omega}(u))}{du} = 2 \int_0^1 \left( \frac{D}{\partial u} \frac{\partial \alpha}{\partial u} \cdot \frac{\partial \alpha}{\partial t} \right) dt.
\]

We can choose \( 0 = t_0 < t_1 < \cdots < t_k = 1 \) so that \( \alpha \) is smooth on each strip \( (\epsilon, \epsilon) \times [t_{i-1}, t_i] \).

Then, we can “integrate by parts” on \( [t_{i-1}, t_i] \) as follows: The equation

\[
\frac{\partial}{\partial t} \left( \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right) = \left( \frac{D}{\partial t} \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right) + \left( \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right)
\]
implies that
\[
\int_{t_{i-1}}^{t_i} \left\langle \frac{D}{\partial t} \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle dt = \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle \bigg|_{t=(t_i)^-} - \int_{t_{i-1}}^{t_i} \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle dt.
\]

Adding up these formulae for \(i = 1, \ldots, k - 1\) and using the fact that \(\frac{\partial \alpha}{\partial u} = 0\) for \(t = 0\) and \(t = 1\), we get
\[
\frac{1}{2} \frac{dE(\tilde{\alpha}(u))}{du} = -\sum_{i=1}^{k-1} \left\langle \frac{\partial \alpha}{\partial u}, \Delta t_i \frac{\partial \alpha}{\partial t} \right\rangle - \int_0^1 \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt.
\]

Setting \(u = 0\), we obtain the formula
\[
\frac{1}{2} \frac{dE(\tilde{\alpha}(u))}{du} \bigg|_{u=0} = -\sum_i \langle W_t, \Delta t_i V \rangle - \int_0^1 \left\langle W_t, \frac{D}{dt} V_t \right\rangle dt,
\]
as claimed. \(\square\)

**Remark:** The reader will observe that the proof used the fact that the connection is compatible with the metric and torsion-free.

Intuitively, the first term on the right-hand side shows that varying the path \(\omega\) in the direction of decreasing “kink” tends to decrease \(E\). The second term shows that varying the curve in the direction of its acceleration vector \(\frac{D}{dt} \omega'(t)\) also tends to reduce \(E\).

A geodesic \(\gamma\) (parametrized over \([0, 1]\)) is smooth on the entire interval \([0, 1]\) and its acceleration vector \(\frac{D}{dt} \gamma'(t)\) is identically zero along \(\gamma\). This gives us half of

**Theorem 12.21.** Let \((M, g)\) be a Riemannian manifold. For any two points \(p, q \in M\), a path \(\omega \in \Omega(p, q)\) (parametrized over \([0, 1]\)) is critical for the energy function \(E\) iff \(\omega\) is a geodesic.

**Proof.** From the first variation formula, it is clear that a geodesic is a critical path of \(E\).

Conversely, assume \(\omega\) is a critical path of \(E\). There is a variation \(\tilde{\alpha}\) of \(\omega\) such that its associated variation vector field is of the form
\[
W(t) = f(t) \frac{D}{dt} \omega'(t),
\]
with \(f(t)\) smooth and positive except that it vanishes at the \(t_i\)’s. For this variation we get
\[
\frac{1}{2} \frac{dE(\tilde{\alpha}(u))}{du} \bigg|_{u=0} = -\int_0^1 f(t) \left\langle \frac{D}{dt} \omega'(t), \frac{D}{dt} \omega'(t) \right\rangle dt.
\]
This expression is zero iff
\[
\frac{D}{dt} \omega'(t) = 0 \quad \text{on } [0, 1].
\]
Hence, the restriction of \( \omega \) to each \([t_i, t_{i+1}]\) is a geodesic.

It remains to prove that \( \omega \) is smooth on the entire interval \([0, 1]\). For this, pick a variation \( \tilde{\alpha} \) such that

\[
W(t_i) = \Delta t_i V.
\]

Then, we have

\[
\frac{1}{2} \frac{dE(\tilde{\alpha}(u))}{du} \bigg|_{u=0} = -\sum_{i=1}^{k} \langle \Delta t_i V, \Delta t_i V \rangle.
\]

If the above expression is zero, then \( \Delta t_i V = 0 \) for \( i = 1, \ldots, k - 1 \), which means that \( \omega \) is \( C^1 \) everywhere on \([0, 1]\). By the uniqueness theorem for ODE’s, \( \omega \) must be smooth everywhere on \([0, 1]\), and thus, it is an unbroken geodesic.

**Remark:** If \( \omega \in \Omega(p, q) \) is parametrized by arc-length, then it is easy to prove that

\[
\frac{dL(\tilde{\alpha}(u))}{du} \bigg|_{u=0} = \frac{1}{2} \frac{dE(\tilde{\alpha}(u))}{du} \bigg|_{u=0}.
\]

As a consequence, a path \( \omega \in \Omega(p, q) \) is critical for the arc-length function \( L \) iff it can be reparametrized so that it is a geodesic (see Gallot, Hulin and Lafontaine [73], Chapter 3, Theorem 3.31).

In order to go deeper into the study of geodesics, we need Jacobi fields and the “second variation formula,” both involving a curvature term. Therefore, we now proceed with a more thorough study of curvature on Riemannian manifolds.
Chapter 13

Curvature in Riemannian Manifolds

Since the notion of curvature can be defined for curves and surfaces, it is natural to wonder whether it can be generalized to manifolds of dimension \( n \geq 3 \). Such a generalization does exist and was first proposed by Riemann. However, Riemann’s seminal paper published in 1868 two years after his death only introduced the sectional curvature, and did not contain any proofs or any general methods for computing the sectional curvature. Fifty years or so later, the idea emerged that the curvature of a Riemannian manifold \( M \) should be viewed as a measure \( R(X,Y)Z \) of the extent to which the operator \( (X,Y) \mapsto \nabla_X \nabla_Y Z \) is symmetric, where \( \nabla \) is a connection on \( M \) (where \( X,Y,Z \) are vector fields, with \( Z \) fixed). It turns out that the operator \( R(X,Y)Z \) is \( C^\infty(M) \)-linear in all of its three arguments, so for all \( p \in M \), it defines a trilinear map

\[
R_p: T_pM \times T_pM \times T_pM \longrightarrow T_pM.
\]

The curvature operator \( R \) is a rather complicated object, so it is natural to seek a simpler object. Fortunately, there is a simpler object, namely the sectional curvature \( K(u,v) \), which arises from \( R \) through the formula

\[
K(u,v) = \langle R(u,v)u, v \rangle,
\]

for linearly independent unit vectors \( u, v \). When \( \nabla \) is the Levi-Civita connection induced by a Riemannian metric on \( M \), it turns out that the curvature operator \( R \) can be recovered from the sectional curvature. Another important notion of curvature is the \textit{Ricci curvature} \( \text{Ric}(x,y) \), which arises as the trace of the linear map \( v \mapsto R(x,v)y \). The curvature operator \( R \), sectional curvature, and Ricci curvature, are introduced in the first three sections of this chapter.

In Section 12.5, we discovered that the geodesics are exactly the critical paths of the energy functional (Theorem 12.21). A deeper understanding is achieved by investigating the second derivative of the energy functional at a critical path (a geodesic). By analogy with the Hessian of a real-valued function on \( \mathbb{R}^n \), it is possible to define a bilinear functional

\[
I_\gamma : T_\gamma \Omega(p,q) \times T_\gamma \Omega(p,q) \rightarrow \mathbb{R}
\]
when $\gamma$ is a critical point of the energy function $E$ (that is, $\gamma$ is a geodesic). This bilinear form is usually called the index form. In order to define the functional $I_{\gamma}$ (where $\gamma$ is a geodesic), we introduce 2-parameter variations, which generalize the variations given by Definition 12.12. Then, we derive the second variation formula, which gives an expression for the second derivative $\partial^2((E \circ \tilde{\alpha})/\partial u_1 \partial u_2)(u_1, u_2) \mid_{(0,0)}$, where $\tilde{\alpha}$ is a 2-variation of a geodesic $\gamma$. Remarkably, this expression contains a curvature term $R(V,W_1)V$, where $W_1(t) = (\partial \alpha/\partial u_1)(0,0,t)$ and $V(t) = \gamma'(t)$. The second variation formula allows us to show that the index form $I(W_1, W_2)$ is well-defined, and symmetric bilinear. When $\gamma$ is a minimal geodesic, $I$ is positive semi-definite. For any geodesic $\gamma$, we define the index of

$$I: T_{\gamma}\Omega(p,q) \times T_{\gamma}\Omega(p,q) \to \mathbb{R}$$

as the maximum dimension of a subspace of $T_{\gamma}\Omega(p,q)$ on which $I$ is negative definite. Section 13.4 is devoted to the second variation formula and the definition of the index form.

In Section 13.5, we define Jacobi fields and study some of their properties. Given a geodesic $\gamma \in \Omega(p,q)$, a vector field $J$ along $\gamma$ is a Jacobi field iff it satisfies the Jacobi differential equation

$$\frac{D^2 J}{dt^2} + R(\gamma', J)\gamma' = 0.$$  

We prove that Jacobi fields are exactly the vector fields that belong to the nullspace of the index form $I$. Jacobi fields also turn out to arise from special variations consisting of geodesics (geodesic variations). We define the notion of conjugate points along a geodesic. We show that the derivative of the exponential map is expressible in terms of a Jacobi field and characterize the critical points of the exponential in terms of conjugate points.

Section 13.6 presents some applications of Jacobi fields and the second variation formula to topology. We prove

(1) Hadamard and Cartan’s Theorem about complete manifolds of non-positive sectional curvature.

(2) Myers’ Theorem about complete manifolds of Ricci curvature bounded from below by a positive number.

We also state the famous Morse Index Theorem.

In Section 13.7, we revisit the cut locus and prove more properties about it using Jacobi fields.

### 13.1 The Curvature Tensor

As we said above, if $M$ is a Riemannian manifold and if $\nabla$ is a connection on $M$, the Riemannian curvature $R(X,Y)Z$ measures the extent to which the operator $(X,Y) \mapsto \nabla_X \nabla_Y Z$
13.1. THE CURVATURE TENSOR

is symmetric (for any fixed $Z$). The Riemannian curvature also measures the defect of symmetry of the operator $\nabla^2_{X,Y}Z$ given by

$$\nabla^2_{X,Y}Z = \nabla_X (\nabla_Y Z) - \nabla_{\nabla_X Y} Z,$$

and called the second covariant derivative of $Z$ with respect to $X$ and $Y$. In fact, we will show that if $\nabla$ is the Levi-Civita connection,

$$R(X,Y)Z = \nabla^2_{Y,X}Z - \nabla^2_{X,Y}Z.$$

The Riemannian curvature is a special instance of the notion of curvature of a connection on a vector bundle. This approach is discussed in Chapter 29, but the present chapter can be read and understood independently of Chapter 29.

If $(M, \langle -, - \rangle)$ is a Riemannian manifold of dimension $n$, and if the connection $\nabla$ on $M$ is the flat connection, which means that

$$\nabla_X \left( \frac{\partial}{\partial x_i} \right) = 0, \quad i = 1, \ldots, n,$$

for every chart $(U, \varphi)$ and all $X \in \mathfrak{X}(U)$, since every vector field $Y$ on $U$ can be written uniquely as

$$Y = \sum_{i=1}^{n} Y_i \frac{\partial}{\partial x_i}$$

for some smooth functions $Y_i$ on $U$, for every other vector field $X$ on $U$, because the connection is flat and by the Leibniz property of connections, we have

$$\nabla_X \left( Y_i \frac{\partial}{\partial x_i} \right) = X(Y_i) \frac{\partial}{\partial x_i} + Y_i \nabla_X \left( \frac{\partial}{\partial x_i} \right) = X(Y_i) \frac{\partial}{\partial x_i}.$$

Then, it is easy to check that the above implies that

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z,$$

for all $X,Y,Z \in \mathfrak{X}(M)$. Consequently, it is natural to define the deviation of a connection from the flat connection by the quantity

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

for all $X,Y,Z \in \mathfrak{X}(M)$. The above defines a function

$$R: \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \rightarrow \mathfrak{X}(M),$$

which is clearly skew-symmetric in $X$ and $Y$. This function turns out to be $C^\infty(M)$-linear in $X,Y,Z$. 
Proposition 13.1. Let $M$ be a manifold with any connection $\nabla$. The function

$$ R: \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \longrightarrow \mathfrak{X}(M) $$

given by

$$ R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z $$

is $C^\infty(M)$-linear in $X,Y,Z$, and skew-symmetric in $X$ and $Y$. As a consequence, for any $p \in M$, $(R(X,Y)Z)_p$ depends only on $X(p), Y(p), Z(p)$.

Proof. Let us check $C^\infty(M)$-linearity in $Z$. Additivity is clear. For any function $f \in C^\infty(M)$, we have

$$ \nabla_Y \nabla_X (fZ) = \nabla_Y (X(f)Z + f \nabla_X Z) $$

$$ = Y(X(f))Z + X(f) \nabla_Y Z + Y(f) \nabla_X Z + f \nabla_Y \nabla_X Z. $$

It follows that

$$ \nabla_X \nabla_Y (fZ) - \nabla_Y \nabla_X (fZ) = X(Y(f))Z + Y(f) \nabla_X Z + X(f) \nabla_Y Z + f \nabla_X \nabla_Y Z $$

$$ - Y(X(f))Z - X(f) \nabla_Y Z - Y(f) \nabla_X Z - f \nabla_Y \nabla_X Z $$

$$ = (XY - YX)(f)Z + f(\nabla_X \nabla_Y - \nabla_Y \nabla_X)Z, $$

hence

$$ R(X,Y)(fZ) = \nabla_X \nabla_Y (fZ) - \nabla_Y \nabla_X (fZ) - \nabla_{[X,Y]} (fZ) $$

$$ = (XY - YX)(f)Z + f(\nabla_X \nabla_Y - \nabla_Y \nabla_X)Z - [X,Y](f)Z - f \nabla_{[X,Y]} Z $$

$$ = (XY - YX - [X,Y])(f)Z + f(\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]})Z $$

$$ = fR(X,Y)Z. $$

Let us now check $C^\infty(M)$-linearity in $Y$. Additivity is clear. For any function $f \in C^\infty(M)$, recall that

$$ [X,fY] = X(f)Y + f[X,Y]. $$

Then,

$$ R(X,fY)Z = \nabla_X \nabla_{fY} Z - \nabla_{fY} \nabla_X Z - \nabla_{[X,fY]} Z $$

$$ = \nabla_X (f\nabla_Y Z) - f \nabla_Y \nabla_X Z - X(f) \nabla_Y Z - f \nabla_{[X,Y]} Z $$

$$ = X(f) \nabla_Y Z + f \nabla_X \nabla_Y Z - f \nabla_Y \nabla_X Z - X(f) \nabla_Y Z - f \nabla_{[X,Y]} Z $$

$$ = f(\nabla_X \nabla_Y Z - f \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z) $$

$$ = fR(X,Y)Z. $$

Since $R$ is skew-symmetric in $X$ and $Y$, $R$ is also $C^\infty(M)$-linear in $X$. For any chart $(U, \varphi)$, we can express the vector fields $X, Y, Z$ uniquely as

$$ X = \sum_{i=1}^n X_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{j=1}^n Y_j \frac{\partial}{\partial x_j}, \quad Z = \sum_{k=1}^n Z_k \frac{\partial}{\partial x_k}. $$
13.1. THE CURVATURE TENSOR

for some smooth functions \(X_i, Y_j, Z_k \in C^\infty(U)\), and by \(C^\infty(U)\)-linearity, we have

\[
R(X, Y)Z = \sum_{i,j,k} R\left(X_i \frac{\partial}{\partial x_i}, Y_j \frac{\partial}{\partial x_j}\right) \left(Z_k \frac{\partial}{\partial x_k}\right).
\]

Evaluated at \(p\), we get

\[
(R(X, Y)Z)_p = \sum_{i,j,k} X_i(p)Y_j(p)Z_k(q) \left(R\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) \left(\frac{\partial}{\partial x_k}\right)\right)_p,
\]

an expression that depends only on the values of the functions \(X_i, Y_j, Z_k\) at \(p\).

It follows that \(R\) defines for every \(p \in M\) a trilinear map

\[
R_p: T_pM \times T_pM \times T_pM \rightarrow T_pM.
\]

(In fact, \(R\) defines a \((1,3)\)-tensor.)

If our manifold is a Riemannian manifold \((M, \langle -, - \rangle)\) equipped with a connection, experience shows that it is useful to consider the family of quadrilinear forms (unfortunately!) also denoted \(R\), given by

\[
R_p(x, y, z, w) = \langle R_p(x, y)z, w \rangle_p,
\]

as well as the expression \(R_p(x, y, y, x)\), which, for an orthonormal pair of vectors \((x, y)\), is known as the sectional curvature \(K_p(x, y)\).

This last expression brings up a dilemma regarding the choice for the sign of \(R\). With our present choice, the sectional curvature \(K_p(x, y)\) is given by \(K_p(x, y) = R_p(x, y, y, x)\), but many authors define \(K\) as \(K_p(x, y) = R_p(x, y, x, y)\). Since \(R(X, Y)\) is skew-symmetric in \(X, Y\), the latter choice corresponds to using \(-R(X, Y)\) instead of \(R(X, Y)\), that is, to define \(R(X, Y)Z\) by

\[
R(X, Y)Z = \nabla_{[X,Y]}Z + \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z.
\]

As pointed out by Milnor [125] (Chapter II, Section 9), the latter choice for the sign of \(R\) has the advantage that, in coordinates, the quantity \(\langle R(\partial/\partial x_h, \partial/\partial x_i)\partial/\partial x_j, \partial/\partial x_k\rangle\) coincides with the classical Ricci notation, \(R_{hijk}\). Gallot, Hulin and Lafontaine [73] (Chapter 3, Section A.1) give other reasons supporting this choice of sign. Clearly, the choice for the sign of \(R\) is mostly a matter of taste and we apologize to those readers who prefer the first choice but we will adopt the second choice advocated by Milnor and others (including O’Neill [138] and Do Carmo [60]), we make the following formal definition:
Definition 13.1. Let \((M, \langle -,- \rangle)\) be a Riemannian manifold equipped with the Levi-Civita connection. The curvature tensor is the family of trilinear functions \(R_p : T_p M \times T_p M \times T_p M \to T_p M\) defined by
\[
R_p(x, y)z = \nabla_{[X,Y]}Z + \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z,
\]
for every \(p \in M\) and for any vector fields \(X, Y, Z \in \mathfrak{X}(M)\) such that \(x = X(p), y = Y(p), \) and \(z = Z(p)\). The family of quadrilinear forms associated with \(R\), also denoted \(R\), is given by
\[
R_p(x, y, z, w) = \langle R_p(x, y)z, w \rangle_p,
\]
for all \(p \in M\) and all \(x, y, z, w \in T_p M\).

Following common practice in mathematics, in the interest of keeping notation to a minimum, we often write \(R(x, y, z, w)\) instead of \(R_p(x, y, z, w)\). Since \(x, y, z, w \in T_p M\), this abuse of notation rarely causes confusion.

**Remark:** The curvature tensor \(R\) is indeed a \((1, 3)\)-tensor, and the associated family of quadrilinear forms is a \((0, 4)\)-tensor.

Locally in a chart, we write
\[
R \left( \frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i} \right) \frac{\partial}{\partial x_j} = \sum_l R^l_{jhi} \frac{\partial}{\partial x_l}
\]
and
\[
R_{hijk} = \left\langle R \left( \frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i} \right) \frac{\partial}{\partial x_j}, \frac{\partial}{\partial x_k} \right\rangle = \sum_l g_{lk} R^l_{jhi}.
\]
The coefficients \(R^l_{jhi}\) can be expressed in terms of the Christoffel symbols \(\Gamma^k_{ij}\), by a rather unfriendly formula; see Gallot, Hulin and Lafontaine [73] (Chapter 3, Section 3.A.3) or O’Neill [138] (Chapter III, Lemma 38). Since we have adopted O’Neill’s conventions for the order of the subscripts in \(R^l_{jhi}\), here is the formula from O’Neill:
\[
R^l_{jhi} = \partial_i \Gamma^l_{hj} - \partial_h \Gamma^l_{ij} + \sum_m \Gamma^l_{im} \Gamma^m_{hj} - \sum_m \Gamma^l_{hm} \Gamma^m_{ij}.
\]
For example, in the case of the sphere \(S^2\), we parametrize as
\[
\begin{align*}
x &= \sin \theta \cos \varphi \\
y &= \sin \theta \sin \varphi \\
z &= \cos \theta,
\end{align*}
\]
over the domain to \(\{(\theta, \varphi) \mid 0 < \theta < \pi, 0 < \varphi < 2\pi\}\). For the basis \((u(\theta, \varphi), v(\theta, \varphi))\) of the the tangent plane \(T_p S^2\) at \(p = (\sin \theta \cos \varphi, \sin \theta \sin \varphi, \cos \theta)\), where
\[
\begin{align*}
u(\theta, \varphi) &= \frac{\partial}{\partial \varphi} = (\cos \theta \cos \varphi, \cos \theta \sin \varphi, -\sin \theta) \\
u(\theta, \varphi) &= \frac{\partial}{\partial \varphi} = (\sin \theta \sin \varphi, \sin \theta \cos \varphi, 0),
\end{align*}
\]
we found that the metric on $T_pS^2$ is given by the matrix

$$g_p = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2 \theta \end{pmatrix};$$

see Section 10.2. Note that

$$g_p^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2 \theta} \end{pmatrix}.$$ 

Since the Christoffel symbols of the Levi-Civita connection are given by

$$\Gamma^k_{ij} = \frac{1}{2} \sum_{l=1}^n g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}),$$

(see Section 11.3), we discover that the only nonzero Christoffel symbols are

$$\Gamma^2_{12} = \Gamma^2_{\theta \varphi} = \Gamma^2_{\varphi \theta} = \frac{1}{2} \sum_{l=1}^2 g^{2l}(\partial_1 g_{2l} + \partial_2 g_{1l} - \partial_l g_{12}) = \frac{1}{2} g^{22} \partial_1 g_{22} = \frac{1}{2} \left( \frac{1}{\sin^2 \theta} \cdot \frac{\partial}{\partial \theta} \sin^2 \theta \right) = \frac{\cos \theta}{\sin \theta},$$

$$\Gamma^1_{22} = \Gamma^\theta_{\varphi \varphi} = \frac{1}{2} \sum_{l=1}^2 g^{1l}(\partial_2 g_{2l} + \partial_2 g_{1l} - \partial_l g_{22}) = -\frac{1}{2} \partial_1 g_{22} = -\frac{\partial}{\partial \theta} \sin^2 \theta = -\sin \theta \cos \theta,$$

where we have set $\theta \to 1$ and $\varphi \to 2$. The only nonzero Riemann curvature tensor components are

$$R^1_{212} = R^\theta_{\varphi \theta \varphi} = \partial_2 \Gamma^1_{12} - \partial_1 \Gamma^1_{22} + \sum_{m=1}^2 \Gamma^1_{1m} \Gamma^m_{12} - \sum_{m=1}^2 \Gamma^1_{2m} \Gamma^m_{22} = \partial_2 \Gamma^1_{12} - \partial_1 \Gamma^1_{22} + \sum_{m=1}^2 \Gamma^1_{1m} \Gamma^m_{12} - \sum_{m=1}^2 \Gamma^1_{2m} \Gamma^m_{22} = -\sin^2 \theta$$

$$R^1_{221} = R^\varphi_{\theta \varphi \theta} = \partial_1 \Gamma^1_{22} - \partial_2 \Gamma^1_{12} + \sum_{m=1}^2 \Gamma^1_{1m} \Gamma^m_{22} - \sum_{m=1}^2 \Gamma^1_{2m} \Gamma^m_{12} = -\sin^2 \theta$$

$$R^2_{112} = R^\varphi_{\theta \varphi \theta} = \partial_2 \Gamma^2_{11} - \partial_1 \Gamma^2_{12} + \sum_{m=1}^2 \Gamma^2_{1m} \Gamma^m_{11} - \sum_{m=1}^2 \Gamma^2_{2m} \Gamma^m_{21} = -\frac{\partial}{\partial \theta} \left( \frac{\cos \theta}{\sin \theta} \right) - \Gamma^2_{12} \Gamma^2_{12} = -\frac{\partial}{\partial \theta} \cot \theta - \frac{\cos^2 \theta}{\sin^2 \theta} = \frac{1 - \cos^2 \theta}{\sin^2 \theta} = 1$$

$$R^2_{121} = R^\varphi_{\theta \varphi \theta} = \partial_1 \Gamma^2_{12} - \partial_2 \Gamma^2_{11} + \sum_{m=1}^2 \Gamma^2_{1m} \Gamma^m_{21} - \sum_{m=1}^2 \Gamma^2_{2m} \Gamma^m_{11} = -R^2_{112} = -1,$$
CHAPTER 13. CURVATURE IN RIEMANNIAN MANIFOLDS

while the only nonzero components of the associated quadrilinear form are

\[ R_{1221} = \sum_{l=1}^{2} g_{l1} R_{211}^l = g_{11} R_{211}^1 = -\sin^2 \theta \]

\[ R_{2121} = \sum_{l=1}^{2} g_{l1} R_{221}^l = g_{11} R_{221}^1 = \sin^2 \theta \]

\[ R_{1212} = \sum_{l=1}^{2} g_{l2} R_{121}^l = g_{22} R_{121}^2 = \sin^2 \theta \]

\[ R_{2112} = \sum_{l=1}^{2} g_{l2} R_{122}^l = g_{22} R_{122}^2 = -\sin^2 \theta. \]

There is another way of defining the curvature tensor which is useful for comparing second covariant derivatives of one-forms. For any fixed vector field \( Z \), the map \( Y \mapsto \nabla Y Z \) from \( \mathfrak{X}(M) \) to \( \mathfrak{X}(M) \) is a \( C^\infty(M) \)-linear map that we will denote \( \nabla Z \) (this is a \( (1,1) \) tensor). The covariant derivative \( \nabla_X \nabla_Y Z \) of \( \nabla Z \) is defined by

\[ (\nabla_X (\nabla_Z))(Y) = \nabla_X (\nabla_Y Z) - (\nabla_{\nabla_X Y})Z. \]

Usually, \((\nabla_X (\nabla_Z))(Y)\) is denoted by \( \nabla^2_{X,Y} Z \), and

\[ \nabla^2_{X,Y} Z = \nabla_X (\nabla_Y Z) - \nabla_{\nabla_X Y} Z \]

is called the second covariant derivative of \( Z \) with respect to \( X \) and \( Y \). Then, we have

\[ \nabla^2_{Y,X} Z - \nabla^2_{X,Y} Z = \nabla_Y (\nabla_X Z) - \nabla_{\nabla_Y X} Z - \nabla_X (\nabla_Y Z) + \nabla_{\nabla_X Y} Z \]

\[ = \nabla_Y (\nabla_X Z) - \nabla_X (\nabla_Y Z) + \nabla_{\nabla_Y X - \nabla_{[X,Y]} Z} \]

\[ = \nabla_Y (\nabla_X Z) - \nabla_X (\nabla_Y Z) + \nabla_{[X,Y]} Z \]

\[ = R(X,Y)Z, \]

since \( \nabla_X Y - \nabla_Y X = [X,Y] \), as the Levi-Civita connection is torsion-free. Therefore, the curvature tensor can also be defined by

\[ R(X,Y)Z = \nabla^2_{Y,X} Z - \nabla^2_{X,Y} Z. \]

We already know that the curvature tensor has some symmetry properties, for example \( R(y,x)z = -R(x,y)z \), but when it is induced by the Levi-Civita connection, it has more remarkable properties stated in the next proposition.

**Proposition 13.2.** For a Riemannian manifold \( (M, \langle -, - \rangle) \) equipped with the Levi-Civita connection, the curvature tensor satisfies the following properties for every \( p \in M \) and for all \( x, y, z, w \in T_p M \):
13.1. THE CURVATURE TENSOR

(1) \( R(x, y)z = -R(y, x)z \)

(2) (First Bianchi Identity) \( R(x, y)z + R(y, z)x + R(z, x)y = 0 \)

(3) \( R(x, y, z, w) = -R(x, y, w, z) \)

(4) \( R(x, y, z, w) = R(z, w, x, y) \).

Proof. The proof of Proposition 13.2 uses the fact that \( R_p(x, y)z = R(X, Y)Z \), for any vector fields \( X, Y, Z \) such that \( x = X(p) \), \( y = Y(p) \) and \( Z = Z(p) \). In particular, \( X, Y, Z \) can be chosen so that their pairwise Lie brackets are zero (choose a coordinate system and give \( X, Y, Z \) constant components). Part (1) is already known. Part (2) follows from the fact that the Levi-Civita connection is torsion-free and is equivalent to the Jacobi identity for Lie brackets. In particular

\[
R(x, y)z + R(y, z)x + R(z, x)y = \nabla_{[x, y]}Z + \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[y, z]}X + \nabla_Z \nabla_Y X - \nabla_Y \nabla_Z X + \nabla_{[z, x]}Y \\
+ \nabla_X \nabla_Z Y - \nabla_Z \nabla_X Y \\
= \nabla_{[x, y]}Z + \nabla_Y (\nabla_X Z - \nabla_Z X) + \nabla_X (\nabla_Z Y - \nabla_Y Z) + \nabla_Z (\nabla_Y X - \nabla_X Y) \\
+ \nabla_{[y, z]}X + \nabla_{[z, x]}Y \\
= \nabla_{[x, y]}Z + \nabla_Y [X, Z] + \nabla_X [Z, Y] + \nabla_Z [Y, X] + \nabla_{[y, z]}X + \nabla_{[z, x]}Y \\
= \nabla_{[x, y]}Z - \nabla_Z [X, Y] + \nabla_{[y, z]}X - \nabla_X [Y, Z] + \nabla_{[z, x]}Y - \nabla_Y [Z, X] \\
= [[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0, \quad \text{by Proposition 8.4.}
\]

Parts (3) and (4) are a little more tricky. Complete proofs can be found in Milnor [125] (Chapter II, Section 9), O’Neill [138] (Chapter III) and Kuhnel [109] (Chapter 6, Lemma 6.3).

Part (3) of Proposition 13.2 can be interpreted as the fact that for every \( p \in M \) and all \( x, y \in T_pM \), the linear map \( z \mapsto R(x, y)z \) (from \( T_pM \) to itself) is skew-symmetric. Indeed, for all \( z, w \in T_pM \), we have

\[
\langle R(x, y)z, w \rangle = R(x, y, z, w) = -R(x, y, w, z) = -\langle R(x, y)w, z \rangle = -\langle z, R(x, y)w \rangle.
\]

The next proposition will be needed in the proof of the second variation formula. If \( \alpha : U \to M \) is a parametrized surface, where \( U \) is some open subset of \( \mathbb{R}^2 \), we say that a vector field \( V \in \mathcal{X}(M) \) is a vector field along \( \alpha \) iff \( V(x, y) \in T_{\alpha(x, y)}M \), for all \( (x, y) \in U \). For any smooth vector field \( V \) along \( \alpha \), we also define the covariant derivatives \( DV/\partial x \) and \( DV/\partial y \) as follows: For each fixed \( y_0 \), if we restrict \( V \) to the curve

\[
x \mapsto \alpha(x, y_0)
\]

we obtain a vector field \( V_{y_0} \) along this curve, and we set

\[
\frac{DV}{\partial x} (x, y_0) = \frac{DV_{y_0}}{dx}.
\]
Then, we let \( y_0 \) vary so that \( (x, y_0) \in U \), and this yields \( DV/\partial x \). We define \( DV/\partial y \) in a similar manner, using a fixed \( x_0 \).

**Proposition 13.3.** For a Riemannian manifold \((M, \langle -,-\rangle)\) equipped with the Levi-Civita connection, for every parametrized surface \( \alpha : \mathbb{R}^2 \rightarrow M \), for every vector field \( V \in \mathfrak{X}(M) \) along \( \alpha \), we have

\[
\frac{D}{\partial y} \frac{D}{\partial x} V - \frac{D}{\partial x} \frac{D}{\partial y} V = R\left( \frac{\partial \alpha}{\partial x}, \frac{\partial \alpha}{\partial y} \right) V.
\]

**Proof.** Express both sides in local coordinates in a chart and make use of the identity

\[
\nabla \frac{\partial}{\partial x_j} \nabla \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_k} - \nabla \frac{\partial}{\partial x_i} \nabla \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_k} = R\left( \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right) \frac{\partial}{\partial x_k}.
\]

**Remark:** Since the Levi-Civita connection is torsion-free, it is easy to check that

\[
\frac{D}{\partial x} \frac{\partial \alpha}{\partial y} = \frac{D}{\partial y} \frac{\partial \alpha}{\partial x}.
\]

We used this identity in the proof of Theorem 12.20.

The curvature tensor is a rather complicated object. Thus, it is quite natural to seek simpler notions of curvature. The sectional curvature is indeed a simpler object, and it turns out that the curvature tensor can be recovered from it.

### 13.2 Sectional Curvature

Basically, the sectional curvature is the curvature of two-dimensional sections of our manifold. Given any two vectors \( u, v \in T_p M \), recall by Cauchy-Schwarz that

\[
\langle u, v \rangle_p^2 \leq \langle u, u \rangle_p \langle v, v \rangle_p,
\]

with equality iff \( u \) and \( v \) are linearly dependent. Consequently, if \( u \) and \( v \) are linearly independent, we have

\[
\langle u, u \rangle_p \langle v, v \rangle_p - \langle u, v \rangle_p^2 \neq 0.
\]

In this case, we claim that the ratio

\[
K_p(u, v) = \frac{R_p(u, v, u, v)}{\langle u, u \rangle_p \langle v, v \rangle_p - \langle u, v \rangle_p^2} = \frac{\langle R_p(u, v)u, v \rangle}{\langle u, u \rangle_p \langle v, v \rangle_p - \langle u, v \rangle_p^2}
\]

is independent of the plane \( \Pi \) spanned by \( u \) and \( v \).

If \( (x, y) \) is another basis of \( \Pi \), then

\[
x = au + bv \\
y = cu + dv.
\]
We get
\[ \langle x, x \rangle_p \langle y, y \rangle_p - \langle x, y \rangle_p^2 = (ad - bc)^2 (\langle u, u \rangle_p \langle v, v \rangle_p - \langle u, v \rangle_p^2), \]
and similarly,
\[ R_p(x, y, x, y) = \langle R_p(x, y) x, y \rangle_p = (ad - bc)^2 R_p(u, v, u, v), \]
which proves our assertion.

**Definition 13.2.** Let \( (M, \langle - , - \rangle) \) be any Riemannian manifold equipped with the Levi-Civita connection. For every \( p \in T_p M \), for every 2-plane \( \Pi \subseteq T_p M \), the sectional curvature \( K_p(\Pi) \) of \( \Pi \) is given by
\[ K_p(\Pi) = K_p(x, y) = \frac{R_p(x, y, x, y)}{\langle x, x \rangle_p \langle y, y \rangle_p - \langle x, y \rangle_p^2}, \]
for any basis \((x, y)\) of \( \Pi \).

As in the case of the curvature tensor, in order to keep notation to a minimum we often write \( K(\Pi) \) instead of \( K_p(\Pi) \) (or \( K(x, y) \)) instead of \( K_p(x, y) \)). Since \( \Pi \subseteq T_p M \) \((x, y \in T_p M)\) for some \( p \in M \), this rarely causes confusion.

Let us take a moment to compute the sectional curvature of \( S^2 \). By using the notation from Section 13.1 we find that
\[ K \left( \frac{\partial}{\partial \theta}, \frac{\partial}{\partial \phi} \right) = \frac{R \left( \frac{\partial}{\partial \theta}, \frac{\partial}{\partial \phi}, \frac{\partial}{\partial \theta}, \frac{\partial}{\partial \phi} \right)}{\sin^2 \theta} = \frac{R_{1212}}{\sin^2 \theta} = 1. \]

Observe that if \((x, y)\) is an orthonormal basis, then the denominator is equal to 1. The expression \( R_p(x, y, x, y) \) is often denoted \( \kappa_p(x, y) \). Remarkably, \( \kappa_p \) determines \( R_p \). We denote the function \( p \mapsto \kappa_p \) by \( \kappa \). We state the following proposition without proof:

**Proposition 13.4.** Let \((M, \langle - , - \rangle)\) be any Riemannian manifold equipped with the Levi-Civita connection. The function \( \kappa \) determines the curvature tensor \( R \). Thus, the knowledge of all the sectional curvatures determines the curvature tensor. Moreover, for all \( p \in M \), for all \( x, y, w, z \in T_p M \), we have
\[ 6 \langle R(x, y) z, w \rangle = \kappa(x + w, y + z) - \kappa(x, y + z) - \kappa(w, y + z) \]
\[ - \kappa(y + w, x + z) + \kappa(y, x + z) + \kappa(w, x + z) \]
\[ - \kappa(x + w, y) + \kappa(x, y) + \kappa(w, y) \]
\[ - \kappa(x + w, z) + \kappa(x, z) + \kappa(w, z) \]
\[ + \kappa(y + w, x) - \kappa(y, x) - \kappa(w, x) \]
\[ + \kappa(y + w, z) - \kappa(y, z) - \kappa(w, z). \]
For a proof of this formidable equation, see Kuhnel [109] (Chapter 6, Theorem 6.5). A different proof of the above proposition (without an explicit formula) is also given in O’Neill [138] (Chapter III, Corollary 42).

Let
\[ R_1(x, y)z = \langle x, z \rangle y - \langle y, z \rangle x. \]
Observe that
\[ \langle R_1(x, y)x, y \rangle = \langle x, x \rangle \langle y, y \rangle - \langle x, y \rangle^2. \]

As a corollary of Proposition 13.4, we get:

**Proposition 13.5.** Let \((M, \langle -, - \rangle)\) be any Riemannian manifold equipped with the Levi-Civita connection. If the sectional curvature \(K(\Pi)\) does not depend on the plane \(\Pi\) but only on \(p \in M\), in the sense that \(K\) is a scalar function \(K : M \to \mathbb{R}\), then
\[ R = K(p)R_1. \]

**Proof.** By hypothesis,
\[ \kappa_p(x, y) = K(p)(\langle x, x \rangle_p \langle y, y \rangle_p - \langle x, y \rangle^2_p), \]
for all \(x, y\). As the right-hand side of the formula in Proposition 13.4 consists of a sum of terms, we see that the right-hand side is equal to \(K\) times a similar sum with \(\kappa\) replaced by
\[ \langle R_1(x, y)x, y \rangle = \langle x, x \rangle \langle y, y \rangle - \langle x, y \rangle^2, \]
so it is clear that \(R = K(p)R_1\).  

In particular, in dimension \(n = 2\), the assumption of Proposition 13.5 holds and \(K\) is the well-known Gaussian curvature for surfaces.

**Definition 13.3.** A Riemannian manifold \((M, \langle -, - \rangle)\) is said to have constant (resp. negative, resp. positive) curvature iff its sectional curvature is constant (resp. negative, resp. positive).

In dimension \(n \geq 3\), we have the following somewhat surprising theorem due to F. Schur:

**Proposition 13.6.** (F. Schur, 1886) Let \((M, \langle -, - \rangle)\) be a connected Riemannian manifold. If \(\dim(M) \geq 3\) and if the sectional curvature \(K(\Pi)\) does not depend on the plane \(\Pi \subseteq T_pM\) but only on the point \(p \in M\), then \(K\) is constant (i.e., does not depend on \(p\)).

The proof, which is quite beautiful, can be found in Kuhnel [109] (Chapter 6, Theorem 6.7).
13.2. SECTIONAL CURVATURE

If we replace the metric \( g = \langle - , - \rangle \) by the metric \( \tilde{g} = \lambda \langle - , - \rangle \) where \( \lambda > 0 \) is a constant, some simple calculations show that the Christoffel symbols and the Levi-Civita connection are unchanged, as well as the curvature tensor, but the sectional curvature is changed, with

\[ \tilde{K} = \lambda^{-1} K. \]

As a consequence, if \( M \) is a Riemannian manifold of constant curvature, by rescaling the metric, we may assume that either \( K = -1 \), or \( K = 0 \), or \( K = +1 \). Here are standard examples of spaces with constant curvature.

1. The sphere \( S^n \subseteq \mathbb{R}^{n+1} \) with the metric induced by \( \mathbb{R}^{n+1} \), where \( S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\} \).

The sphere \( S^n \) has constant sectional curvature \( K = +1 \). This can be shown by using the fact that the stabilizer of the action of \( \text{SO}(n+1) \) on \( S^n \) is isomorphic to \( \text{SO}(n) \). Then, it is easy to see that the action of \( \text{SO}(n) \) on \( T_pS^n \) is transitive on 2-planes and from this, it follows that \( K = 1 \) (for details, see Gallot, Hulin and Lafontaine [73] (Chapter 3, Proposition 3.14).

2. Euclidean space \( \mathbb{R}^{n+1} \) with its natural Euclidean metric. Of course, \( K = 0 \).

3. The hyperbolic space \( \mathcal{H}_n^+(1) \) from Definition 6.1. Recall that this space is defined in terms of the Lorentz inner product \( \langle - , - \rangle_1 \) on \( \mathbb{R}^{n+1} \), given by

\[ \langle (x_1, \ldots, x_{n+1}), (y_1, \ldots, y_{n+1}) \rangle_1 = -x_1y_1 + \sum_{i=2}^{n+1} x_iy_i. \]

By definition, \( \mathcal{H}_n^+(1) \), written simply \( H^n \), is given by

\[ H^n = \{ x = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_1 = -1, x_1 > 0 \}. \]

Given any point \( p = (x_1, \ldots, x_{n+1}) \in H^n \), since a tangent vector at \( p \) is defined as \( x'(0) \) for any curve \( x : (-\epsilon, \epsilon) \to H^n \) with \( x(0) = p \), we note that

\[ \frac{d}{dt} \langle x(t), x(t) \rangle_1 = 2 \langle x'(t), x(t) \rangle_1 = \frac{d}{dt} (-1) = 0, \]

which by setting \( t = 0 \) implies that the set of tangent vectors \( u \in T_pH^n \) are given by the equation

\[ \langle p, u \rangle_1 = 0; \]

that is, \( T_pH^n \) is orthogonal to \( p \) with respect to the Lorentz inner-product. Since \( p \in H^n \), we have \( \langle p, p \rangle_1 = -1 \), that is, \( p \) is timelike, so by Proposition 6.8, all vectors in \( T_pH^n \) are spacelike; that is,

\[ \langle u, u \rangle_1 > 0, \quad \text{for all } u \in T_pH^n, \ u \neq 0. \]
Therefore, the restriction of $\langle -, - \rangle_1$ to $T_p H^n$ is positive, definite, which means that it is a metric on $T_p H^n$. The space $H^n$ equipped with this metric $g_H$ is called hyperbolic space and it has constant curvature $K = -1$. This can be shown by using the fact that the stabilizer of the action of $\SO_0(n, 1)$ on $H^n$ is isomorphic to $\SO(n)$ (see Proposition 6.9). Then, it is easy to see that the action of $\SO(n)$ on $T_p H^n$ is transitive on 2-planes and from this, it follows that $K = -1$ (for details, see Gallot, Hulin and Lafontaine [73] (Chapter 3, Proposition 3.14).

There are other isometric models of $H^n$ that are perhaps intuitively easier to grasp but for which the metric is more complicated. For example, there is a map $\PD: B^n \to H^n$ where $B^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ is the open unit ball in $\mathbb{R}^n$, given by

$$\PD(x) = \left(1 + \|x\|^2, \frac{2x}{1 - \|x\|^2}, \frac{2x}{1 - \|x\|^2}\right).$$

It is easy to check that $\langle \PD(x), \PD(x) \rangle_1 = -1$ and that $\PD$ is bijective and an isometry. One also checks that the pull-back metric $g_{\PD} = \PD^* g_H$ on $B^n$ is given by

$$g_{\PD} = \frac{4}{(1 - \|x\|^2)^2} (dx_1^2 + \cdots + dx_n^2).$$

The metric $g_{\PD}$ is called the conformal disc metric, and the Riemannian manifold $(B^n, g_{\PD})$ is called the Poincaré disc model or conformal disc model. The metric $g_{\PD}$ is proportional to the Euclidean metric, and thus angles are preserved under the map $\PD$. Another model is the Poincaré half-plane model $\{x \in \mathbb{R}^n \mid x_1 > 0\}$, with the metric

$$g_{\PH} = \frac{1}{x_1^2} (dx_1^2 + \cdots + dx_n^2).$$

We already encountered this space for $n = 2$.

The metrics for $S^n$, $\mathbb{R}^{n+1}$, and $H^n$ have a nice expression in polar coordinates, but we prefer to discuss the Ricci curvature next.

### 13.3 Ricci Curvature

The Ricci tensor is another important notion of curvature. It is mathematically simpler than the sectional curvature (since it is symmetric) and it plays an important role in the theory of gravitation as it occurs in the Einstein field equations. The Ricci tensor is an example of contraction, in this case, the trace of a linear map. Recall that if $f: E \to E$ is a linear map from a finite-dimensional Euclidean vector space to itself, given any orthonormal basis $(e_1, \ldots, e_n)$, we have

$$\text{tr}(f) = \sum_{i=1}^n \langle f(e_i), e_i \rangle.$$
Definition 13.4. Let \((M, (-, -))\) be a Riemannian manifold (equipped with the Levi-Civita connection). The Ricci curvature \(\text{Ric} \) of \(M\) is the \((0, 2)\)-tensor defined as follows: For every \(p \in M\), for all \(x, y \in T_pM\), set \(\text{Ric}_p(x, y)\) to be the trace of the endomorphism \(v \mapsto R_p(x, v)y\). With respect to any orthonormal basis \((e_1, \ldots, e_n)\) of \(T_pM\), we have
\[
\text{Ric}_p(x, y) = \sum_{j=1}^n \langle R_p(x, e_j)y, e_j \rangle_p = \sum_{j=1}^n R_p(x, e_j, y, e_j).
\]
The scalar curvature \(S\) of \(M\) is the trace of the Ricci curvature; that is, for every \(p \in M\),
\[
S(p) = \sum_{i \neq j} R_p(e_i, e_j, e_i) = \sum_{i \neq j} K_p(e_i, e_j),
\]
where \(K_p(e_i, e_j)\) denotes the sectional curvature of the plane spanned by \(e_i, e_j\).

In the interest of keeping notation to a minimum, we often write \(\text{Ric}(x, y)\) instead of \(\text{Ric}_p(x, y)\).

In a chart the Ricci curvature is given by
\[
R_{ij} = \text{Ric} \left( \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right) = \sum_m R_{ijm}^m,
\]
and the sectional curvature is given by
\[
S(p) = \sum_{i,j} g^{ij} R_{ij},
\]
where \((g^{ij})\) is the inverse of the Riemann metric matrix \((g_{ij})\). See O’Neill, pp. 87-88 [138].

For \(S^2\), the calculations of Section 13.1 imply that
\[
\text{Ric} \left( \frac{\partial}{\partial \theta}, \frac{\partial}{\partial \phi} \right) = R_{12} = \sum_{m=1}^2 R_{12m} = R_{112}^1 + R_{122}^2 = 0
\]
\[
\text{Ric} \left( \frac{\partial}{\partial \phi}, \frac{\partial}{\partial \theta} \right) = R_{21} = \sum_{m=1}^2 R_{21m} = R_{211}^1 + R_{212}^2 = 0
\]
\[
\text{Ric} \left( \frac{\partial}{\partial \theta}, \frac{\partial}{\partial \phi} \right) = R_{11} = \sum_{m=1}^2 R_{11m} = R_{111}^1 + R_{112}^1 = 1
\]
\[
\text{Ric} \left( \frac{\partial}{\partial \phi}, \frac{\partial}{\partial \phi} \right) = R_{22} = \sum_{m=1}^2 R_{22m} = R_{221}^1 + R_{222}^2 = \sin^2 \theta,
\]
and that
\[
S(p) = \sum_{i=1}^2 \sum_{j=1}^2 g^{ij} R_{ij} = g^{11} R_{11} + g^{12} R_{12} + g^{21} R_{21} + g^{22} R_{22}
= 1 \cdot 1 + \frac{1}{\sin^2 \theta} \cdot \sin^2 \theta = 2.
\]
In view of Proposition 13.2 (4), the Ricci curvature is symmetric. The tensor Ric is a \((0, 2)\)-tensor but it can be interpreted as a \((1, 1)\)-tensor as follows: We let \(\text{Ric}_p^\#\) be the \((1, 1)\)-tensor given by
\[
\langle \text{Ric}_p^\# u, v \rangle_p = \text{Ric}_p(u, v),
\]
for all \(u, v \in T_pM\). Then it is easy to see that
\[
S(p) = \text{tr}(\text{Ric}_p^\#).
\]
This is why we said (by abuse of language) that \(S\) is the trace of \(\text{Ric}\). Observe that if \((e_1, \ldots, e_n)\) is any orthonormal basis of \(T_pM\), as
\[
\text{Ric}_p(u, v) = \sum_{j=1}^n R_p(u, e_j, v, e_j)
\]
\[
= \sum_{j=1}^n R_p(e_j, u, e_j, v)
\]
\[
= \sum_{j=1}^n \langle R_p(e_j, u) e_j, v \rangle_p,
\]
we have
\[
\text{Ric}_p^\#(u) = \sum_{j=1}^n R_p(e_j, u) e_j.
\]

Observe that in dimension \(n = 2\), we get \(S(p) = 2K(p)\). Therefore, in dimension 2, the scalar curvature determines the curvature tensor. In dimension \(n = 3\), it turns out that the Ricci tensor completely determines the curvature tensor, although this is not obvious. We will come back to this point later.

Since \(\text{Ric}(x, y)\) is symmetric, \(\text{Ric}(x, x)\) determines \(\text{Ric}(x, y)\) completely (Use the polarization identity for a symmetric bilinear form, \(\varphi\):
\[
2\varphi(x, y) = \Phi(x + y) - \Phi(x) - \Phi(y),
\]
with \(\Phi(x) = \varphi(x, x)\)). Observe that for any orthonormal frame \((e_1, \ldots, e_n)\) of \(T_pM\), using the definition of the sectional curvature \(K\), we have
\[
\text{Ric}(e_1, e_1) = \sum_{i=1}^n \langle R(e_1, e_i) e_1, e_i \rangle = \sum_{i=2}^n K(e_1, e_i).
\]
Thus, \(\text{Ric}(e_1, e_1)\) is the sum of the sectional curvatures of any \(n - 1\) orthogonal planes orthogonal to \(e_1\) (a unit vector).

For a Riemannian manifold with constant sectional curvature, we see that
\[
\text{Ric}(x, x) = (n - 1)Kg(x, x), \quad S = n(n - 1)K,
\]
where $g = \langle -, - \rangle$ is the metric on $M$. Indeed, if $K$ is constant, then we know that $R = KR_1$, and so

\[
\text{Ric}(x, x) = K \sum_{i=1}^{n} g(R_1(x, e_i)x, e_i)
\]

\[
= K \sum_{i=1}^{n} g((x, x)e_i - (e_i, x)e_i, e_i)
\]

\[
= K \sum_{i=1}^{n} (g(e_i, e_i)g(x, x) - g(e_i, x)^2)
\]

\[
= K(n g(x, x) - \sum_{i=1}^{n} g(e_i, x)^2)
\]

\[
= (n - 1)K g(x, x).
\]

Spaces for which the Ricci tensor is proportional to the metric are called Einstein spaces.

**Definition 13.5.** A Riemannian manifold $(M, g)$ is called an *Einstein space* iff the Ricci curvature is proportional to the metric $g$; that is:

\[
\text{Ric}(x, y) = \lambda g(x, y),
\]

for some function $\lambda: M \to \mathbb{R}$.

If $M$ is an Einstein space, observe that $S = n\lambda$.

**Remark:** For any Riemannian manifold $(M, g)$, the quantity

\[
G = \text{Ric} - \frac{S}{2}g
\]

is called the *Einstein tensor* (or *Einstein gravitation tensor* for space-times spaces). The Einstein tensor plays an important role in the theory of general relativity. For more on this topic, see Kuhnel [109] (Chapters 6 and 8) O’Neill [138] (Chapter 12).

## 13.4 The Second Variation Formula and the Index Form

As in previous sections, we assume that all our manifolds are Riemannian manifolds equipped with the Levi-Civita connection. In Section 12.5, we discovered that the geodesics are exactly the critical paths of the energy functional (Theorem 12.21). For this, we derived the First Variation Formula (Theorem 12.20). It is not too surprising that a deeper understanding
is achieved by investigating the second derivative of the energy functional at a critical path (a geodesic). By analogy with the Hessian of a real-valued function on $\mathbb{R}^n$, it is possible to define a bilinear functional

$$I_\gamma: T_\gamma \Omega(p, q) \times T_\gamma \Omega(p, q) \to \mathbb{R}$$

when $\gamma$ is a critical point of the energy function $E$ (that is, $\gamma$ is a geodesic). This bilinear form is usually called the index form. Note that Milnor denotes $I_\gamma$ by $E^{**}$ and refers to it as the Hessian of $E$, but this is a bit confusing since $I_\gamma$ is only defined for critical points, whereas the Hessian is defined for all points, critical or not.

Now, if $f: M \to \mathbb{R}$ is a real-valued function on a finite-dimensional manifold $M$ and if $p$ is a critical point of $f$, which means that $df_p = 0$, it turns out that there is a symmetric bilinear map $I_f: T_p M \times T_p M \to \mathbb{R}$ such that

$$I_f(X(p), Y(p)) = X_p(Yf) = Y_p(Xf),$$

for all vector fields $X, Y \in \mathfrak{X}(M)$. To show this, observe that for any two vector field $X, Y$,

$$X_p(Yf) - Y_p(Xf) = ([X, Y])_p(f) = df_p([X, Y]_p) = 0,$$

since $p$ is a critical point, namely $df_p = 0$. It follows that the function $I_f: T_p M \times T_p M \to \mathbb{R}$ defined by

$$I_f(X(p), Y(p)) = X_p(Yf)$$

is bilinear and symmetric. Furthermore, $I_f(u, v)$ can be computed as follows: for any $u, v \in T_p M$, for any smooth map $\alpha: \mathbb{R}^2 \to M$ such that

$$\alpha(0, 0) = p, \quad \frac{\partial \alpha}{\partial x}(0, 0) = u, \quad \frac{\partial \alpha}{\partial y}(0, 0) = v,$$

we have

$$I_f(u, v) = \frac{\partial^2 (f \circ \alpha)(x, y)}{\partial x \partial y} \bigg|_{(0,0)} = \frac{\partial \alpha}{\partial x} \left( \frac{\partial \alpha}{\partial y} (f \circ \alpha) \right)_{(0,0)}.$$

The above suggests that in order to define

$$I_\gamma: T_\gamma \Omega(p, q) \times T_\gamma \Omega(p, q) \to \mathbb{R},$$

that is to define $I_\gamma(W_1, W_2)$, where $W_1, W_2 \in T_\gamma \Omega(p, q)$ are vector fields along $\gamma$ (with $W_1(0) = W_2(0) = 0$ and $W_1(1) = W_2(1) = 0$), we consider 2-parameter variations

$$\alpha: U \times [0, 1] \to M,$$

where $U$ is an open subset of $\mathbb{R}^2$ with $(0, 0) \in U$, such that

$$\alpha(0, 0, t) = \gamma(t), \quad \frac{\partial \alpha}{\partial u_1}(0, 0, t) = W_1(t), \quad \frac{\partial \alpha}{\partial u_2}(0, 0, t) = W_2(t).$$
13.4. THE SECOND VARIATION FORMULA AND THE INDEX FORM

Figure 13.1: A 2-parameter variation \( \alpha \). The pink curve with its associated velocity field is \( \alpha(0, 0, t) = \gamma(t) \). The blue vector field is \( W_1(t) \) while the green vector field is \( W_2(t) \).

See Figure 13.1.

Then, we set

\[
I_\gamma(W_1, W_2) = \frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} \bigg|_{(0,0)},
\]

where \( \tilde{\alpha} \in \Omega(p, q) \) is the path given by

\[
\tilde{\alpha}(u_1, u_2)(t) = \alpha(u_1, u_2, t).
\]

For simplicity of notation, the above derivative is often written as \( \frac{\partial^2 E}{\partial u_1 \partial u_2} (0, 0) \).

To prove that \( I_\gamma(W_1, W_2) \) is actually well-defined, we need the following result:

**Theorem 13.7. (Second Variation Formula)** Let \( \alpha : U \times [0, 1] \to M \) be a 2-parameter variation of a geodesic \( \gamma \in \Omega(p, q) \), with variation vector fields \( W_1, W_2 \in T_\gamma \Omega(p, q) \) given by

\[
W_1(t) = \frac{\partial \alpha}{\partial u_1} (0, 0, t), \quad W_2(t) = \frac{\partial \alpha}{\partial u_2} (0, 0, t), \quad \alpha(0, 0, t) = \gamma(t).
\]
Then we have the formula
\[
\frac{1}{2} \left. \frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} \right|_{(0,0)} = - \sum_i \left\langle W_2(t), \Delta \left( \frac{D W_1}{dt} \right) \right\rangle - \int_0^1 \left\langle W_2, \frac{D^2 W_1}{dt^2} + R(V, W_1)V \right\rangle dt,
\]
where \( V(t) = \gamma'(t) \) is the velocity field,
\[
\Delta \left( \frac{D W_1}{dt} \right) = \frac{D W_1}{dt}(t_+) - \frac{D W_1}{dt}(t_-)
\]
is the jump in \( \frac{D W_1}{dt} \) at one of its finitely many points of discontinuity in \((0,1)\), and \( E \) is the energy function on \( \Omega(p,q) \).

Proof. (After Milnor, see [125], Chapter II, Section 13, Theorem 13.1.) By the First Variation Formula (Theorem 12.20), we have
\[
\frac{1}{2} \frac{\partial E(\tilde{\alpha}(u_1, u_2))}{\partial u_2} = - \sum_i \left\langle \frac{\partial \alpha}{\partial u_2}, \Delta \frac{\partial \alpha}{\partial t} \right\rangle - \int_0^1 \left\langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt.
\]
Thus, we get
\[
\frac{1}{2} \frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} = - \sum_i \left\langle \frac{D}{\partial u_1} \frac{\partial \alpha}{\partial u_2}, \Delta \frac{\partial \alpha}{\partial t} \right\rangle - \sum_i \left\langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial u_1} \Delta \frac{\partial \alpha}{\partial t} \right\rangle - \int_0^1 \left\langle \frac{D}{\partial u_1} \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt - \int_0^1 \left\langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial u_1} \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} \right\rangle dt.
\]
Let us evaluate this expression for \((u_1, u_2) = (0,0)\). Since \( \gamma = \tilde{\alpha}(0,0) \) is an unbroken geodesic, we have
\[
\Delta \frac{\partial \alpha}{\partial t} = 0, \quad \frac{D}{\partial t} \frac{\partial \alpha}{\partial t} = 0,
\]
so that the first and third term are zero. As
\[
\frac{D}{\partial u_1} \frac{\partial \alpha}{\partial t} = \frac{D}{\partial t} \frac{\partial \alpha}{\partial u_1},
\]
(see the remark just after Proposition 13.3), we can rewrite the second term and we get
\[
\frac{1}{2} \frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} (0,0) = - \sum_i \left\langle W_2, \Delta \frac{D}{\partial t} W_1 \right\rangle - \int_0^1 \left\langle W_2, \frac{D}{\partial u_1} \frac{D}{\partial t} V \right\rangle dt. \quad (*)
\]
In order to interchange the operators \( \frac{D}{\partial u_1} \) and \( \frac{D}{\partial t} \), we need to bring in the curvature tensor. Indeed, by Proposition 13.3, we have
\[
\frac{D}{\partial u_1} \frac{D}{\partial t} V - \frac{D}{\partial t} \frac{D}{\partial u_1} V = R \left( \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u_1} \right) V = R(V, W_1)V.
\]
Together with the equation
\[
\frac{D}{\partial u_1} V = \frac{D}{\partial u_1} \frac{\partial \alpha}{\partial t} = \frac{D}{\partial t} \frac{\partial \alpha}{\partial u_1} = \frac{D}{\partial t} W_1,
\]
this yields
\[
\frac{D}{\partial u_1} \frac{D}{\partial t} V = \frac{D^2 W_1}{dt^2} + R(V, W_1)V.
\]
Substituting this last expression in (*), we get the Second Variation Formula.

Theorem 13.7 shows that the expression
\[
\frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2}
\]
only depends on the variation fields \(W_1\) and \(W_2\), and thus \(I_\gamma(W_1, W_2)\) is actually well-defined. If no confusion arises, we write \(I(W_1, W_2)\) for \(I_\gamma(W_1, W_2)\).

Proposition 13.8. Given any geodesic \(\gamma \in \Omega(p, q)\), the map \(I: T_\gamma \Omega(p, q) \times T_\gamma \Omega(p, q) \rightarrow \mathbb{R}\) defined so that for all \(W_1, W_2 \in T_\gamma \Omega(p, q)\),
\[
I(W_1, W_2) = \frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2}
\]
only depends on \(W_1\) and \(W_2\) and is bilinear and symmetric, where \(\alpha: U \times [0, 1] \rightarrow M\) is any 2-parameter variation, with
\[
\alpha(0, 0, t) = \gamma(t), \quad \frac{\partial \alpha}{\partial u_1}(0, 0, t) = W_1(t), \quad \frac{\partial \alpha}{\partial u_2}(0, 0, t) = W_2(t).
\]

Proof. We already observed that the Second Variation Formula implies that \(I(W_1, W_2)\) is well defined. This formula also shows that \(I\) is bilinear. As
\[
\frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} = \frac{\partial^2 (E \circ \tilde{\alpha})(u_1, u_2)}{\partial u_2 \partial u_1},
\]
\(I\) is symmetric (but this is not obvious from the right-hand side of the Second Variation Formula).

On the diagonal, \(I(W, W)\) can be described in terms of a 1-parameter variation of \(\gamma\). In fact,
\[
I(W, W) = \left. \frac{d^2 E(\tilde{\alpha})}{du^2} \right|_{(0)},
\]
where \(\tilde{\alpha}: (-\epsilon, \epsilon) \rightarrow \Omega(p, q)\) denotes any variation of \(\gamma\) with variation vector field \(\frac{d\tilde{\alpha}}{du}(0)\) equal to \(W\). To prove this equation it is only necessary to introduce the 2-parameter variation
\[
\tilde{\beta}(u_1, u_2) = \tilde{\alpha}(u_1 + u_2),
\]
and to observe that
\[ \frac{\partial \tilde{\beta}}{\partial u_i} \frac{d\tilde{\alpha}}{du} \quad \frac{\partial^2 (E \circ \tilde{\beta})}{\partial u_1 \partial u_2} = \frac{d^2 (E \circ \tilde{\alpha})}{du^2}. \]

As an application of the above remark we have the following result:

**Proposition 13.9.** If \( \gamma \in \Omega(p,q) \) is a minimal geodesic, then the bilinear index form \( I \) is positive semi-definite, which means that \( I(W,W) \geq 0 \) for all \( W \in T_\gamma \Omega(p,q) \).

**Proof.** The inequality
\[ E(\tilde{\alpha}(u)) \geq E(\gamma) = E(\tilde{\alpha}(0)) \]
implies that
\[ \frac{d^2 E(\tilde{\alpha})}{du^2} (0) \geq 0, \]
which is exactly what needs to be proved. \( \square \)

For any geodesic \( \gamma \), if we define the index of
\[ I : T_\gamma \Omega(p,q) \times T_\gamma \Omega(p,q) \to \mathbb{R} \]
as the maximum dimension of a subspace of \( T_\gamma \Omega(p,q) \) on which \( I \) is negative definite, then Proposition 13.9 says that the index of \( I \) is zero (for the minimal geodesic \( \gamma \)). It turns out that the index of \( I \) is finite for any geodesic, \( \gamma \) (this is a consequence of the Morse Index Theorem).

### 13.5 Jacobi Fields and Conjugate Points

Jacobi fields arise naturally when considering the expression involved under the integral sign in the Second Variation Formula and also when considering the derivative of the exponential. In this section, all manifolds under consideration are Riemannian manifold equipped with the Levi-Civita connection.

If \( B : E \times E \to \mathbb{R} \) is a symmetric bilinear form defined on some vector space \( E \) (possibly infinite dimensional), recall that the nullspace of \( B \) is the subset \( \text{null}(B) \) of \( E \) given by
\[ \text{null}(B) = \{ u \in E \mid B(u,v) = 0, \text{ for all } v \in E \}. \]
The nullity \( \nu \) of \( B \) is the dimension of its nullspace. The bilinear form \( B \) is nondegenerate iff \( \text{null}(B) = (0) \) iff \( \nu = 0 \). If \( U \) is a subset of \( E \), we say that \( B \) is positive definite (resp. negative definite) on \( U \) iff \( B(u,u) > 0 \) (resp. \( B(u,u) < 0 \)) for all \( u \in U \), with \( u \neq 0 \). The index of \( B \) is the maximum dimension of a subspace of \( E \) on which \( B \) is negative definite. We will determine the nullspace of the symmetric bilinear form
\[ I : T_\gamma \Omega(p,q) \times T_\gamma \Omega(p,q) \to \mathbb{R}, \]
13.5. JACOBI FIELDS AND CONJUGATE POINTS

where $\gamma$ is a geodesic from $p$ to $q$ in some Riemannian manifold $M$. Now, if $W$ is a vector field in $T_\gamma\Omega(p,q)$ and $W$ satisfies the equation

$$\frac{D^2W}{dt^2} + R(V, W)V = 0,$$

where $V(t) = \gamma'(t)$ is the velocity field of the geodesic $\gamma$, since $W$ is smooth along $\gamma$, (because $\gamma$ is a geodesic and consists of a single smooth curve), it is obvious from the Second Variation Formula that

$$I(W, W) = 0, \quad \text{for all } W \in T_\gamma\Omega(p,q).$$

Therefore, any vector field in the nullspace of $I$ must satisfy equation $(\ast)$. Such vector fields are called Jacobi fields.

**Definition 13.6.** Given a geodesic $\gamma \in \Omega(p,q)$, a vector field $J$ along $\gamma$ is a Jacobi field iff it satisfies the Jacobi differential equation

$$\frac{D^2J}{dt^2} + R(\gamma', J)\gamma' = 0.$$

The equation of Definition 13.6 is a linear second-order differential equation that can be transformed into a more familiar form by picking some orthonormal parallel vector fields $X_1, \ldots, X_n$ along $\gamma$. To do this, pick any orthonormal basis $(e_1, \ldots, e_n)$ in $T_pM$, with $e_1 = \gamma'(0)/\|\gamma'(0)\|$, and use parallel transport along $\gamma$ to get $X_1, \ldots, X_n$. Then, we can write $J = \sum_{i=1}^n y_i X_i$, for some smooth functions $y_i$, and the Jacobi equation becomes the system of second-order linear ODE’s

$$\frac{d^2y_i}{dt^2} + \sum_{j=1}^n R(\gamma', X_j, \gamma', X_i)y_j = 0, \quad 1 \leq i \leq n.$$  

As an illustration of how to derive the preceding system of equations, suppose $J = y_1 X_1 + y_2 X_2$. Since $\frac{DX_i}{dt} = 0$ for all $i$, we find

$$\frac{D}{dt} J = \frac{dy_1}{dt} X_1 + y_1 \frac{DX_1}{dt} + \frac{dy_2}{dt} X_2 + y_2 \frac{DX_2}{dt} = \frac{dy_1}{dt} X_1 + \frac{dy_2}{dt} X_2,$$

and hence

$$\frac{D^2}{dt^2} J = \frac{d^2y_1}{dt^2} X_1 + \frac{dy_1}{dt} \frac{DX_1}{dt} + \frac{d^2y_2}{dt^2} X_2 + \frac{dy_2}{dt} \frac{DX_2}{dt} = \frac{d^2y_1}{dt^2} X_1 + \frac{d^2y_2}{dt^2} X_2.$$  

We now compute

$$R(\gamma', J)\gamma' = R(\gamma', y_1 X_1 + y_2 X_2)\gamma' = y_1 R(\gamma', X_1)\gamma' + y_2 R(\gamma', X_2)\gamma' = c_1 X_1 + c_2 X_2,$$
where the \( c_i \) are smooth functions determined as follows: since \( \langle X_1, X_1 \rangle = 1 = \langle X_2, X_2 \rangle \) and \( \langle X_1, X_2 \rangle = 0 \), we find that
\[
c_1 = \langle y_1 R(\gamma', X_1)\gamma' + y_2 R(\gamma', X_2)\gamma', X_1 \rangle = y_1 (R(\gamma', X_1)\gamma', X_1) + y_2 (R(\gamma', X_2)\gamma', X_1) \\
= y_1 R(\gamma', X_1, \gamma', X_1) + y_2 R(\gamma', X_2, \gamma', X_1)
\]
and that
\[
c_2 = \langle y_1 R(\gamma', X_1)\gamma' + y_2 R(\gamma', X_2)\gamma', X_2 \rangle = y_1 (R(\gamma', X_1)\gamma', X_2) + y_2 (R(\gamma', X_2)\gamma', X_2) \\
= y_1 R(\gamma', X_1, \gamma', X_2) + y_2 R(\gamma', X_2, \gamma', X_2).
\]

These calculations show that the coefficient of \( X_1 \) is
\[
\frac{d^2 y_1}{dt^2} + c_1 = \frac{d^2 y_1}{dt^2} + y_1 R(\gamma', X_1, \gamma', X_1) + y_2 R(\gamma', X_2, \gamma', X_1) \\
= \frac{d^2 y_1}{dt^2} + \sum_{j=1}^{2} R(\gamma', X_j, \gamma', X_1)y_j,
\]
while the coefficient of \( X_2 \) is
\[
\frac{d^2 y_2}{dt^2} + c_2 = \frac{d^2 y_2}{dt^2} + y_1 R(\gamma', X_1, \gamma', X_2) + y_2 R(\gamma', X_2, \gamma', X_2) \\
= \frac{d^2 y_2}{dt^2} + \sum_{j=1}^{2} R(\gamma', X_j, \gamma', X_2)y_j.
\]

Setting these two coefficients equal to zero gives the systems of equations provided by \((*)\).

By the existence and uniqueness theorem for ODE’s, for every pair of vectors \( u, v \in T_pM \), there is a unique Jacobi fields \( J \) so that \( J(0) = u \) and \( \frac{DJ}{dt}(0) = v \). Since \( T_pM \) has dimension \( n \), it follows that the dimension of the space of Jacobi fields along \( \gamma \) is \( 2n \).

If \( J(0) \) and \( \frac{DJ}{dt}(0) \) are orthogonal to \( \gamma'(0) \), then \( J(t) \) is orthogonal to \( \gamma'(t) \) for all \( t \in [0, 1] \). To show this, recall that by the remark after Proposition 13.2, the linear map \( z \mapsto R(x, y)z \) is skew symmetric. As a consequence, it is a standard fact of linear algebra that \( R(x, y)z \) is orthogonal to \( z \). Since \( X_1 \) is obtained by parallel transport along \( \gamma \) starting with \( X_1(0) \) collinear to \( \gamma'(0) \), the vector \( X_1(t) \) is collinear to \( \gamma'(t) \), and since \( R(\gamma', X_j)\gamma' \) is orthogonal to \( \gamma' \), we have
\[
R(\gamma', X_j, \gamma', X_1) = \langle R(\gamma', X_j)\gamma', X_1 \rangle = 0.
\]

But then, the ODE for \( \frac{d^2 y_2}{dt^2} \) yields
\[
\frac{d^2 y_2}{dt^2} = 0,
\]
Since
\[
J(0) = y_1 e_1 + \sum_{j=2}^{n} y_j(0)e_i = y_1(0)\frac{\gamma'(0)}{||\gamma'(0)||} + \sum_{j=2}^{n} y_j(0)e_i,
\]
we find that
\[ 0 = \langle J(0), \gamma'(0) \rangle = \langle J(0), \|\gamma'(0)\| e_1 \rangle = \|\gamma'(0)\| y_1(0) e_1 = \|\gamma'(0)\| y_1(0), \]
and hence conclude that \( y_1(0) = 0 \). Since
\[ \frac{DJ}{dt}(0) = \frac{dy_1}{dt}(0) e_1 + \sum_{j=2}^{n} \frac{dy_j}{dt}(0) e_j = \frac{dy_1}{dt}(0) \frac{\gamma'(0)}{\|\gamma'(0)\|} + \sum_{j=2}^{n} \frac{dy_j}{dt}(0) e_j, \]
we again discover that
\[ 0 = \langle \frac{DJ}{dt}(0), \gamma'(0) \rangle = \langle \frac{DJ}{dt}(0), \|\gamma'(0)\| e_1 \rangle = \|\gamma'(0)\| \frac{dy_1}{dt}(0), \]
and conclude that \( \frac{dy_1}{dt}(0) = 0 \). Because \( y_1(0) = 0 \) and \( \frac{dy_1}{dt}(0) = 0 \), the ODE \( \frac{d^2 y_1}{dt^2} = 0 \) implies that \( y_1(t) = 0 \) for all \( t \in [0, 1] \). Furthermore, if \( J \) is orthogonal to \( \gamma \), which means that \( J(t) \) is orthogonal to \( \gamma'(t) \) for all \( t \in [0, 1] \), then \( \frac{DJ}{dt} \) is also orthogonal to \( \gamma \). Indeed, as \( \gamma \) is a geodesic, \( \frac{d\gamma'}{dt} = 0 \) and
\[ 0 = \frac{d}{dt} \langle J, \gamma' \rangle = \langle \frac{DJ}{dt}, \gamma' \rangle + \langle J, \frac{D\gamma'}{dt} \rangle = \langle \frac{DJ}{dt}, \gamma' \rangle. \]
Therefore, the dimension of the space of Jacobi fields normal to \( \gamma \) is \( 2n - 2 \) and is of the form \( J = \sum_{i=2}^{n} y_i X_i \). These facts prove part of the following

**Proposition 13.10.** If \( \gamma \in \Omega(p, q) \) is a geodesic in a Riemannian manifold of dimension \( n \), then the following properties hold:

(1) For all \( u, v \in T_p M \), there is a unique Jacobi fields \( J \) so that \( J(0) = u \) and \( \frac{DJ}{dt}(0) = v \). Consequently, the vector space of Jacobi fields has dimension \( 2n \).

(2) The subspace of Jacobi fields orthogonal to \( \gamma \) has dimension \( 2n - 2 \). The vector fields \( \gamma' \) and \( t \mapsto t \gamma'(t) \) are Jacobi fields that form a basis of the subspace of Jacobi fields parallel to \( \gamma \) (that is, such that \( J(t) \) is collinear with \( \gamma'(t) \), for all \( t \in [0, 1] \).) See Figure 13.2.

(3) If \( J \) is a Jacobi field, then \( J \) is orthogonal to \( \gamma \) iff there exist \( a, b \in [0, 1] \), with \( a \neq b \), so that \( J(a) \) and \( J(b) \) are both orthogonal to \( \gamma \) iff there is some \( a \in [0, 1] \) so that \( J(a) \) and \( \frac{DJ}{dt}(a) \) are both orthogonal to \( \gamma \).

(4) For any two Jacobi fields \( X, Y \) along \( \gamma \), the expression \( \langle \nabla_\gamma X, Y \rangle - \langle \nabla_\gamma Y, X \rangle \) is a constant, and if \( X \) and \( Y \) vanish at some point on \( \gamma \), then \( \langle \nabla_\gamma X, Y \rangle - \langle \nabla_\gamma Y, X \rangle = 0 \).
Figure 13.2: An orthogonal Jacobi field $J$ for a three dimensional manifold $M$. Note that $J$ is in the plane spanned by $X_2$ and $X_3$, while $X_1$ is in the direction of the velocity field.

**Proof.** We already proved (1) and part of (2). If $J$ is parallel to $\gamma$, then $J(t) = f(t)\gamma(t)'$ and the Jacobi equation becomes
\[
\frac{d^2 f}{dt^2} = 0.
\]
Therefore,
\[
J(t) = (\alpha + \beta t)\gamma'(t).
\]
It is easily shown that $\gamma'$ and $t \mapsto t\gamma'(t)$ are linearly independent (as vector fields).

To prove (3), using the Jacobi equation, observe that
\[
\frac{d^2}{dt^2} \langle J, \gamma' \rangle = \langle \frac{D^2 J}{dt^2}, \gamma' \rangle = -R(J, \gamma', \gamma', \gamma') = 0.
\]
Therefore,
\[ \langle J, \gamma' \rangle = \alpha + \beta t \]
and the result follows. We leave (4) as an exercise. \( \square \)

Following Milnor, we will show that the Jacobi fields in \( T_\gamma \Omega(p, q) \) are exactly the vector fields in the nullspace of the index form \( I \). First, we define the important notion of conjugate points.

**Definition 13.7.** Let \( \gamma \in \Omega(p, q) \) be a geodesic. Two distinct parameter values \( a, b \in [0, 1] \) with \( a < b \) are **conjugate along \( \gamma \)** iff there is some Jacobi field \( J \), not identically zero, such that \( J(a) = J(b) = 0 \). The dimension \( k \) of the space \( J_{a,b} \) consisting of all such Jacobi fields is called the **multiplicity** (or **order of conjugacy**) of \( a \) and \( b \) as conjugate parameters. We also say that the points \( p_1 = \gamma(a) \) and \( p_2 = \gamma(b) \) are **conjugate**.

**Remark:** As remarked by Milnor and others, as \( \gamma \) may have self-intersections, the above definition is ambiguous if we replace \( a \) and \( b \) by \( p_1 = \gamma(a) \) and \( p_2 = \gamma(b) \), even though many authors make this slight abuse. Although it makes sense to say that the points \( p_1 \) and \( p_2 \) are conjugate, the space of Jacobi fields vanishing at \( p_1 \) and \( p_2 \) is not well defined. Indeed, if \( p_1 = \gamma(a) \) for distinct values of \( a \) (or \( p_2 = \gamma(b) \) for distinct values of \( b \)), then we don’t know which of the spaces, \( J_{a,b} \), to pick. We will say that some points \( p_1 \) and \( p_2 \) on \( \gamma \) are conjugate iff there are parameter values, \( a < b \), such that \( p_1 = \gamma(a) \), \( p_2 = \gamma(b) \), and \( a \) and \( b \) are conjugate along \( \gamma \).

However, for the endpoints \( p \) and \( q \) of the geodesic segment \( \gamma \), we may assume that \( p = \gamma(0) \) and \( q = \gamma(1) \), so that when we say that \( p \) and \( q \) are conjugate we consider the space of Jacobi fields vanishing for \( t = 0 \) and \( t = 1 \). This is the definition adopted Gallot, Hulin and Lafontaine [73] (Chapter 3, Section 3E).

In view of Proposition 13.10 (3), the Jacobi fields involved in the definition of conjugate points are orthogonal to \( \gamma \). The dimension of the space of Jacobi fields such that \( J(a) = 0 \) is obviously \( n \), since the only remaining parameter determining \( J \) is \( \frac{dJ}{dt}(a) \). Furthermore, the Jacobi field \( t \mapsto (t-a)\gamma'(t) \) vanishes at \( a \) but not at \( b \), so the multiplicity of conjugate parameters (points) is at most \( n - 1 \).

For example, if \( M \) is a flat manifold, that is if its curvature tensor is identically zero, then the Jacobi equation becomes
\[ \frac{D^2J}{dt^2} = 0. \]
It follows that \( J \equiv 0 \), and thus, there are no conjugate points. More generally, the Jacobi equation can be solved explicitly for spaces of constant curvature; see Do Carmo [60] (Chapter 5, Example 2.3).

**Theorem 13.11.** Let \( \gamma \in \Omega(p, q) \) be a geodesic. A vector field \( W \in T_\gamma \Omega(p, q) \) belongs to the nullspace of the index form \( I \) iff \( W \) is a Jacobi field. Hence, \( I \) is degenerate if \( p \) and \( q \) are conjugate. The nullity of \( I \) is equal to the multiplicity of \( p \) and \( q \).
Proof. (After Milnor [125], Theorem 14.1). We already observed that a Jacobi field vanishing at 0 and 1 belongs to the nullspace of $I$.

Conversely, assume that $W_1 \in T_\gamma \Omega(p,q)$ belongs to the nullspace of $I$. Pick a subdivision $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0,1]$ so that $W_1 \mid [t_i, t_{i+1}]$ is smooth for all $i = 0, \ldots, k-1$, and let $f : [0,1] \to [0,1]$ be a smooth function which vanishes for the parameter values $t_0, \ldots, t_k$ and is strictly positive otherwise. Then, if we let

$$W_2(t) = f(t) \left( \frac{D^2W_1}{dt^2} + R(\gamma', W_1)\gamma' \right)_t,$$

by the Second Variation Formula, we get

$$0 = -\frac{1}{2} I(W_1, W_2) = \sum 0 + \int_0^1 f(t) \left\| \frac{D^2W_1}{dt^2} + R(\gamma', W_1)\gamma' \right\|^2 dt.$$

Consequently, $W_1 \mid [t_i, t_{i+1}]$ is a Jacobi field for all $i = 0, \ldots, k-1$.

Now, let $W'_2 \in T_\gamma \Omega(p,q)$ be a field such that

$$W'_2(t_i) = \Delta_{t_i} \frac{DW_1}{dt}, \quad i = 1, \ldots, k-1.$$

We get

$$0 = -\frac{1}{2} I(W_1, W'_2) = \sum_{i=1}^{k-1} \left\| \Delta_{t_i} \frac{DW_1}{dt} \right\|^2 + \int_0^1 0 dt.$$

Hence, $\frac{DW_1}{dt}$ has no jumps. Now, a solution $W_1$ of the Jacobi equation is completely determined by the vectors $W_1(t_i)$ and $\frac{DW_1}{dt}(t_i)$, so the $k$ Jacobi fields $W_1 \mid [t_i, t_{i+1}]$ fit together to give a Jacobi field $W_1$ which is smooth throughout $[0,1]$. $\square$

Theorem 13.11 implies that the nullity of $I$ is finite, since the vector space of Jacobi fields vanishing at 0 and 1 has dimension at most $n$. In fact, we observed that the dimension of this space is at most $n-1$.

**Corollary 13.12.** The nullity $\nu$ of $I$ satisfies $0 \leq \nu \leq n-1$, where $n = \text{dim}(M)$.

Jacobi fields turn out to be induced by certain kinds of variations called *geodesic variations*.

**Definition 13.8.** Given a geodesic $\gamma \in \Omega(p,q)$, a *geodesic variation* of $\gamma$ is a smooth map

$$\alpha : (-\epsilon, \epsilon) \times [0,1] \to M,$$

such that

1. $\alpha(0,t) = \gamma(t)$, for all $t \in [0,1]$. 


(2) For every \( u \in (-\epsilon, \epsilon) \), the curve \( \tilde{\alpha}(u) \) is a geodesic, where
\[
\tilde{\alpha}(u)(t) = \alpha(u, t), \quad t \in [0, 1].
\]

Note that the geodesics \( \tilde{\alpha}(u) \) do not necessarily begin at \( p \) and end at \( q \), and so a geodesic variation is not a “fixed endpoints” variation. See Figure 13.3.

![Figure 13.3: A geodesic variation for \( S^2 \) with its associated Jacobi field \( W(t) \)](image)

**Proposition 13.13.** If \( \alpha: (-\epsilon, \epsilon) \times [0, 1] \rightarrow M \) is a geodesic variation of \( \gamma \in \Omega(p, q) \), then the vector field \( W(t) = \frac{\partial \alpha}{\partial u}(0, t) \) is a Jacobi field along \( \gamma \).

**Proof.** As \( \alpha \) is a geodesic variation, we have
\[
\frac{D}{dt} \frac{\partial \alpha}{\partial t} = 0.
\]

Hence, using Proposition 13.3, we have
\[
0 = \frac{D}{\partial t} \left( \frac{D}{\partial u} \frac{\partial \alpha}{\partial t} \right) + R \left( \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u} \right) \frac{\partial \alpha}{\partial t} = \frac{D^2}{\partial t \partial u} \frac{\partial \alpha}{\partial u} + R \left( \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u} \right) \frac{\partial \alpha}{\partial t},
\]

where we used the fact (already used before) that
\[
\frac{D}{\partial t} \frac{\partial \alpha}{\partial u} = \frac{D}{\partial u} \frac{\partial \alpha}{\partial t},
\]
as the Levi-Civita connection is torsion-free. \( \square \)
For example, on the sphere $S^n$, for any two antipodal points $p$ and $q$, rotating the sphere keeping $p$ and $q$ fixed, the variation field along a geodesic $\gamma$ through $p$ and $q$ (a great circle) is a Jacobi field vanishing at $p$ and $q$. Rotating in $n - 1$ different directions one obtains $n - 1$ linearly independent Jacobi fields and thus, $p$ and $q$ are conjugate along $\gamma$ with multiplicity $n - 1$.

Interestingly, the converse of Proposition 13.13 holds.

**Proposition 13.14.** For every Jacobi field $W(t)$ along a geodesic $\gamma \in \Omega(p,q)$, there is some geodesic variation $\alpha: (-\epsilon, \epsilon) \times [0,1] \to M$ of $\gamma$ such that $W(t) = \frac{\partial \alpha}{\partial u}(0,t)$. Furthermore, for every point $\gamma(a)$, there is an open subset $U$ containing $\gamma(a)$ such that the Jacobi fields along a geodesic segment in $U$ are uniquely determined by their values at the endpoints of the geodesic.

**Proof.** (After Milnor, see [125], Chapter III, Lemma 14.4.) We begin by proving the second assertion. By Proposition 12.5 (1), there is an open subset $U$ with $\gamma(0) \in U$, so that any two points of $U$ are joined by a unique minimal geodesic which depends differentially on the endpoints. Suppose that $\gamma(t) \in U$ for $t \in [0,\delta]$. We will construct a Jacobi field $W$ along $\gamma \upharpoonright [0,\delta]$ with arbitrarily prescribed values $u$ at $t = 0$ and $v$ at $t = \delta$. Choose some curve $c_0: (-\epsilon, \epsilon) \to U$ so that $c_0(0) = \gamma(0)$ and $c_0'(0) = u$, and some curve $c_{\delta}': (-\epsilon, \epsilon) \to U$ so that $c_{\delta}(0) = \gamma(\delta)$ and $c_{\delta}'(0) = v$. Now define the map

$$\alpha: (-\epsilon, \epsilon) \times [0,\delta] \to M$$

by letting $\tilde{\alpha}(s): [0,\delta] \to M$ be the unique minimal geodesic from $c_0(s)$ to $c_{\delta}(s)$. It is easily checked that $\alpha$ is a geodesic variation of $\gamma \upharpoonright [0,\delta]$ and that

$$J(t) = \frac{\partial \alpha}{\partial u}(0,t)$$

is a Jacobi field such that $J(0) = u$ and $J(\delta) = v$. See Figure 13.4.

We claim that every Jacobi field along $\gamma \upharpoonright [0,\delta]$ can be obtained uniquely in this way. If $J_{\delta}$ denotes the vector space of all Jacobi fields along $\gamma \upharpoonright [0,\delta]$, the map $J \mapsto (J(0), J(\delta))$ defines a linear map

$$\ell: J_{\delta} \to T_{\gamma(0)} M \times T_{\gamma(\delta)} M.$$ 

The above argument shows that $\ell$ is onto. However, both vector spaces have the same dimension $2n$, so $\ell$ is an isomorphism. Therefore, every Jacobi field in $J_{\delta}$ is determined by its values at $\gamma(0)$ and $\gamma(\delta)$.

Now the above argument can be repeated for every point $\gamma(a)$ on $\gamma$, so we get an open cover $\{[l_a, r_a]\}$ of $[0,1]$, such that every Jacobi field along $\gamma \upharpoonright [l_a, r_a]$ is uniquely determined by its endpoints. By compactness of $[0,1]$, the above cover possesses some finite subcover, and we get a geodesic variation $\alpha$ defined on the entire interval $[0,1]$ whose variation field is equal to the original Jacobi field, $W$. 

$\square$
13.5. JACOBI FIELDS AND CONJUGATE POINTS

Figure 13.4: The local geodesic variation $\alpha$ with its Jacobi field such that $J(0) = u$ and $J(\delta) = v$

Remark: The proof of Proposition 13.14 also shows that there is some open interval $(-\delta, \delta)$ such that if $t \in (-\delta, \delta)$, then $\gamma(t)$ is not conjugate to $\gamma(0)$ along $\gamma$. In fact, the Morse Index Theorem implies that for any geodesic segment, $\gamma: [0, 1] \to M$, there are only finitely many points which are conjugate to $\gamma(0)$ along $\gamma$ (see Milnor [125], Part III, Corollary 15.2).

There is also an intimate connection between Jacobi fields and the differential of the exponential map, and between conjugate points and critical points of the exponential map.

Recall that if $f: M \to N$ is a smooth map between manifolds, a point $p \in M$ is a critical point of $f$ iff the tangent map at $p$

$$df_p: T_p M \to T_{f(p)} N$$

is not surjective. If $M$ and $N$ have the same dimension, which will be the case for the rest of this section, $df_p$ is not surjective iff it is not injective, so $p$ is a critical point of $f$ iff there is some nonzero vector $u \in T_p M$ such that $df_p(u) = 0$.

If $\exp_p: T_p M \to M$ is the exponential map, for any $v \in T_p M$ where $\exp_p(v)$ is defined, we have the derivative of $\exp_p$ at $v$:

$$(d \exp_p)_v: T_v(T_p M) \to T_p M.$$  

Since $T_p M$ is a finite-dimensional vector space, $T_v(T_p M)$ is isomorphic to $T_p M$, so we identify $T_v(T_p M)$ with $T_p M$.

Jacobi fields can be used to compute the derivative of the exponential.

**Proposition 13.15.** Given any point $p \in M$, for any vectors $u, v \in T_p M$, if $\exp_p v$ is defined, then

$$J(t) = (d \exp_p)_{tv}(tu), \quad 0 \leq t \leq 1,$$

is the unique Jacobi field such that $J(0) = 0$ and $\frac{DJ}{dt}(0) = u$. 
Proof. We follow the proof in Gallot, Hulin and Lafontaine [73] (Chapter 3, Corollary 3.46). Another proof can be found in Do Carmo [60] (Chapter 5, Proposition 2.7). Let \( \gamma \) be the geodesic given by \( \gamma(t) = \exp_p(tv) \). In \( T_pM \) equipped with the inner product \( g_p \), the Jacobi field \( X \) along the geodesic \( t \mapsto tv \) such that \( X(0) = 0 \) and \( (DX/dt)(0) = u \) is just \( X(t) = tu \). This Jacobi field is generated by the variation \( H(s,t) = t(v + su) \) since \( \partial H/\partial s \) \( H(0,t) = tu \). Because all the curves in this variation are radial geodesics, the variation \( \alpha(s,t) = \exp_p H(s,t) \) of \( \gamma \) (in \( M \)) is also a geodesic variation, and by Proposition 13.13, the vector field \( J(t) = \partial \alpha/\partial s(0,t) \) is a Jacobi vector field. See Figure 13.5.

Figure 13.5: The radial geodesic variation and its image under \( \exp_p \). Note that \( J(t) \) is the dark pink vector field.

By the chain rule we have \( J(t) = (d \exp_p)_{tu}(tu) \), and since \( J(0) = 0 \) and \( (DJ/dt)(0) = u \), we conclude that \( J(t) = (d \exp_p)_{tu}(tu) \) is the unique Jacobi field such that \( J(0) = 0 \) and \( (DJ/dt)(0) = u \).

Proposition 13.16. Let \( \gamma \in \Omega(p,q) \) be a geodesic. The point \( r = \gamma(t) \), with \( t \in (0,1] \), is conjugate to \( p \) along \( \gamma \) iff \( v = t\gamma'(0) \) is a critical point of \( \exp_p \). Furthermore, the multiplicity of \( p \) and \( r \) as conjugate points is equal to the dimension of the kernel of \( (d \exp_p)_v \).
13.5. JACOBI FIELDS AND CONJUGATE POINTS

Proof. We follow the proof in Do Carmo [60] (Chapter 5, Proposition 3.5). Other proofs can be found in O’Neill [138] (Chapter 10, Proposition 10), or Milnor [125] (Part III, Theorem 18.1). The point \( r = \gamma(t) \) is conjugate to \( p \) along \( \gamma \) if there is a non-zero Jacobi field \( J \) along \( \gamma \) such that \( J(0) = 0 \) and \( J(t) = 0 \). Let \( v = \gamma'(0) \) and \( w = (DJ/dt)(0) \). From Proposition 13.15, we have

\[
J(t) = (d\exp_p)_{tv}(tw), \quad 0 \leq t \leq 1.
\]

Observe that \( J \) is non-zero iff \( (DJ/dt)(0) = w \neq 0 \). Therefore, \( r = \gamma(t) \) is conjugate to \( p \) along \( \gamma \) iff

\[
0 = J(t) = (d\exp_p)_{tv}\left(t\frac{DJ}{dt}(0)\right), \quad \frac{DJ}{dt}(0) \neq 0;
\]

that is, iff \( tv \) is a critical point of \( \exp_p \).

The multiplicity of \( p \) and \( r \) as conjugate points is equal to the number of linearly independent Jacobi fields \( J_1, \ldots, J_k \) such that \( J_i(0) = J_i(t) = 0 \) for \( i = 1, \ldots, k \). It is easy to check that \( J_1, \ldots, J_k \) are linearly independent iff \( (DJ_1/dt)(0), \ldots, (DJ_k/dt)(0) \) are linearly independent in \( T_pM \). Indeed, if \( (DJ_1/dt)(0), \ldots, (DJ_k/dt)(0) \) are linearly independent, then \( J_1, \ldots, J_k \) must be linearly independent since otherwise we would have

\[
\lambda_1 J_1 + \cdots + \lambda_k J_k = 0
\]

with some \( \lambda_i \neq 0 \), and by taking the derivative we would obtain a nontrivial dependency among \( (DJ_1/dt)(0), \ldots, (DJ_k/dt)(0) \). Conversely, if \( J_1, \ldots, J_k \) are linearly independent, then if we could express some \( (DJ_i/dt)(0) \) as

\[
\frac{DJ_i}{dt}(0) = \sum_{h \neq i} \lambda_h \frac{DJ_h}{dt}(0)
\]

with some \( \lambda_h \neq 0 \), then the Jacobi field

\[
J(t) = \sum_{h \neq i} \lambda_h J_h(t)
\]

is such that \( J(0) = 0 \) and \( (DJ/dt)(0) = (DJ_i/dt)(0) \), so by uniqueness \( J = J_i \), and \( J_i \) is a nontrivial combination of the other \( J_h \), a contradiction. Since

\[
J_i(t) = (d\exp_p)_{tv}\left(t\frac{DJ_i}{dt}(0)\right),
\]

we have \( J_i(t) = 0 \) iff \( (DJ_i/dt)(0) \in \text{Ker}(d\exp_p)_{tv} \), so the multiplicity of \( p \) and \( r \) is equal to the dimension of \( \text{Ker}(d\exp_p)_{tv} \).

Using Proposition 13.14 it is easy to characterize conjugate points in terms of geodesic variations; see O’Neill [138] (Chapter 10, Proposition 10).
Proposition 13.17. If $\gamma \in \Omega(p, q)$ is a geodesic, then $q$ is conjugate to $p$ iff there is a geodesic variation $\alpha$ of $\gamma$ such that every geodesic $\tilde{\alpha}(u)$ starts from $p$, the Jacobi field $J(t) = \frac{\partial \alpha}{\partial u}(0, t)$ does not vanish identically, and $J(1) = 0$.

Jacobi fields can be used to compute the sectional curvature of the sphere $S^n$ and the sectional curvature of hyperbolic space $H^n = H^+_n(1)$, both equipped with the canonical metric. This requires knowing the geodesics in $S^n$ and $H^n$. This is done in Section 19.4 for the sphere. The hyperbolic space $H^n = H^+_n(1)$ is shown to be a symmetric space in Section 19.8, and it would be easy to derive its geodesics by analogy with what we did for the sphere. For the sake of brevity, we will assume without proof that we know these geodesics. The reader may consult Gallot, Hulin and Lafontaine [73] or O’Neill [138] for details.

First, we consider the sphere $S^n$. For any $p \in S^n$, the geodesic from $p$ with initial velocity a unit vector $v$ is

$$\gamma(t) = (\cos t)p + (\sin t)v.$$ 

Pick some unit vector $u \in T_p M$ orthogonal to $v$. The variation

$$\alpha(s, t) = (\cos t)p + (\sin t)((\cos s)v + (\sin s)u)$$

is a geodesic variation. We obtain the Jacobi vector field

$$Y(t) = \frac{\partial \alpha}{\partial s}(0, t) = (\sin t)u.$$ 

Since $Y$ satisfies the Jacobi differential equation, we have

$$Y'' + R(\gamma', Y)\gamma' = 0.$$ 

But, as $Y(t) = (\sin t)u$, we have

$$Y + Y'' = 0,$$

so

$$R(\gamma', Y)\gamma' = Y,$$

which yields

$$1 = \langle u, u \rangle = \langle R(\gamma', u)\gamma', u \rangle = R(\gamma', u, \gamma', u)$$

since $\langle Y, Y \rangle = (\sin t)^2$ and $R(\gamma', Y, \gamma', Y) = (\sin t)^2R(\gamma', u, \gamma', u)$. Since $\gamma'(0) = v$, it follows that $R(v, u, v, u) = 1$, which means that the sectional curvature of $S^n$ is constant and equal to 1.

Let us now consider the hyperbolic space $H^n$. This time, the geodesic from $p$ with initial velocity a unit vector $v$ is

$$\gamma(t) = (\cosh t)p + (\sinh t)v.$$ 

Pick some unit vector $u \in T_p M$ orthogonal to $v$. The variation

$$\alpha(s, t) = (\cosh t)p + (\sinh t)((\cosh s)v + (\sinh s)u)$$
is a geodesic variation and we obtain the Jacobi vector field
\[ Y(t) = \frac{\partial \alpha}{\partial s}(0, t) = (\sinh t)u. \]
This time,
\[ Y'' - Y = 0, \]
so the Jacobi equation becomes
\[ R(\gamma', Y)\gamma' = -Y. \]
It follows that
\[ -1 = -\langle u, u \rangle = \langle R(\gamma', u)\gamma', u \rangle = R(\gamma', u, \gamma', u) \]
and since \( \gamma'(0) = v \), we get \( R(v, u, v, u) = -1 \), which means that the sectional curvature of \( H^n \) is constant and equal to \(-1\).

Using the covering map of \( \mathbb{RP}^n \) by \( S^n \), it can be shown that \( \mathbb{RP}^n \) with the canonical metric also has constant sectional curvature equal to \(+1\); see Gallot, Hulin and Lafontaine [73] (Chapter III, section 3.49).

**Remark:** If \( u, v \in T_p M \) are orthogonal unit vectors, then \( R(u, v, u, v) = K(u, v) \), the sectional curvature of the plane spanned by \( u \) and \( v \) in \( T_p M \), and for \( t \) small enough, we have
\[ \| J(t) \| = t - \frac{1}{6} K(u, v)t^3 + o(t^3). \]
(Here, \( o(t^3) \) stands for an expression of the form \( t^4 R(t) \), such that \( \lim_{t \to 0} R(t) = 0 \).) Intuitively, this formula tells us how fast the geodesics that start from \( p \) and are tangent to the plane spanned by \( u \) and \( v \) spread apart. Locally, for \( K(u, v) > 0 \) the radial geodesics spread apart less than the rays in \( T_p M \), and for \( K(u, v) < 0 \) they spread apart more than the rays in \( T_p M \). For more details, see Do Carmo [60] (Chapter 5, Proposition 2.7, Corollary 2.10 and the remark that follows.).

There is also another version of “Gauss lemma” whose proof uses Jacobi fields (see Gallot, Hulin and Lafontaine [73], Chapter 3, Lemma 3.70):

**Proposition 13.18. (Gauss Lemma)** Given any point \( p \in M \), for any vectors \( u, v \in T_p M \), if \( \exp_p v \) is defined, then
\[ \langle d(\exp_p)_{tv}(u), d(\exp_p)_{tv}(v) \rangle = \langle u, v \rangle, \quad 0 \leq t \leq 1. \]

As our (connected) Riemannian manifold \( M \) is a metric space, the path space \( \Omega(p, q) \) is also a metric space if we use the metric \( d^* \) given by
\[ d^*(\omega_1, \omega_2) = \max_t (d(\omega_1(t), \omega_2(t))), \]
where \( d \) is the metric on \( M \) induced by the Riemannian metric.

**Remark:** The topology induced by \( d^* \) turns out to be the compact open topology on \( \Omega(p, q) \).
**Theorem 13.19.** Let $\gamma \in \Omega(p, q)$ be a geodesic. Then, the following properties hold:

1. If there are no conjugate points to $p$ along $\gamma$, then there is some open subset $V$ of $\Omega(p, q)$, with $\gamma \in V$, such that

$$L(\omega) \geq L(\gamma) \quad \text{and} \quad E(\omega) \geq E(\gamma),$$

for all $\omega \in V$, with strict inequality when $\omega([0, 1]) \neq \gamma([0, 1])$. We say that $\gamma$ is a local minimum.

2. If there is some $t \in (0, 1)$ such that $p$ and $\gamma(t)$ are conjugate along $\gamma$, then there is a fixed endpoints variation $\alpha$, such that

$$L(\tilde{\alpha}(u)) < L(\gamma) \quad \text{and} \quad E(\tilde{\alpha}(u)) < E(\gamma),$$

for $u$ small enough.

A proof of Theorem 13.19 can be found in Gallot, Hulin and Lafontaine [73] (Chapter 3, Theorem 3.73) or in O’Neill [138] (Chapter 10, Theorem 17 and Remark 18).

### 13.6 Jacobi Field Applications in Topology and Curvature

As before, all our manifolds are Riemannian manifolds equipped with the Levi-Civita connection. Jacobi fields and conjugate points are basic tools that can be used to prove many global results of Riemannian geometry. The flavor of these results is that certain constraints on curvature (sectional, Ricci, sectional) have a significant impact on the topology. One may want consider the effect of non-positive curvature, constant curvature, curvature bounded from below by a positive constant, etc. This is a vast subject and we highly recommend Berger’s Panorama of Riemannian Geometry [19] for a masterly survey. We will content ourselves with three results:

1. Hadamard and Cartan’s Theorem about complete manifolds of non-positive sectional curvature.

2. Myers’ Theorem about complete manifolds of Ricci curvature bounded from below by a positive number.

3. The Morse Index Theorem.

First, on the way to Hadamard and Cartan, we begin with a proposition.

**Proposition 13.20.** Let $M$ be a complete Riemannian manifold with non-positive curvature $K \leq 0$. Then, for every geodesic $\gamma \in \Omega(p, q)$, there are no conjugate points to $p$ along $\gamma$. Consequently, the exponential map $\exp_p : T_p M \to M$ is a local diffeomorphism for all $p \in M$. 
Proof. Let $J$ be a Jacobi field along $\gamma$. Then,

$$\frac{D^2 J}{dt^2} + R(\gamma', J)\gamma' = 0,$$

so that by the definition of the sectional curvature,

$$\left\langle \frac{D^2 J}{dt^2}, J \right\rangle = -\langle R(\gamma', J)\gamma', J \rangle = -R(\gamma', J, \gamma', J) \geq 0.$$

It follows that

$$\frac{d}{dt} \left\langle \frac{DJ}{dt}, J \right\rangle = \left\langle \frac{D^2 J}{dt^2}, J \right\rangle + \left\| \frac{DJ}{dt} \right\|^2 \geq 0.$$

Thus, the function $t \mapsto \left\langle \frac{DJ}{dt}, J \right\rangle$ is monotonic increasing, and strictly so if $\frac{DJ}{dt} \neq 0$. If $J$ vanishes at both 0 and $t$, for any given $t \in (0, 1]$, then so does $\langle \frac{DJ}{dt}, J \rangle$, and hence $\langle \frac{DJ}{dt}, J \rangle$ must vanish throughout the interval $[0, t]$. This implies

$$J(0) = \frac{DJ}{dt}(0) = 0,$$

so that $J$ is identically zero. Therefore, $t$ is not conjugate to 0 along $\gamma$. By Proposition 13.16, $d\exp_p$ is nonsingular for all $p \in M$, which implies that $\exp_p$ is a local diffeomorphism. \qed

**Theorem 13.21.** *(Hadamard–Cartan)* Let $M$ be a complete Riemannian manifold. If $M$ has non-positive sectional curvature $K \leq 0$, then the following hold:

1. For every $p \in M$, the map $\exp_p : T_p M \to M$ is a Riemannian covering, i.e. $\exp_p$ is a smooth covering and a local isometry.

2. If $M$ is simply connected then $M$ is diffeomorphic to $\mathbb{R}^n$, where $n = \dim(M)$; more precisely, $\exp_p : T_p M \to M$ is a diffeomorphism for all $p \in M$. Furthermore, any two points on $M$ are joined by a unique minimal geodesic.

Proof. We follow the proof in Sakai [150] (Chapter V, Theorem 4.1).

1. By Proposition 13.20, the exponential map $\exp_p : T_p M \to M$ is a local diffeomorphism for all $p \in M$. Let $\tilde{g}$ be the pullback metric $\tilde{g} = (\exp_p)^* g$ on $T_p M$ (where $g$ denotes the metric on $M$). We claim that $(T_p M, \tilde{g})$ is complete.

   This is because, for every nonzero $u \in T_p M$, the line $t \mapsto tu$ is mapped to the geodesic $t \mapsto \exp_p(tu)$ in $M$, which is defined for all $t \in \mathbb{R}$ since $M$ is complete, and thus this line is a geodesic in $(T_p M, \tilde{g})$. Since this holds for all $u \in T_p M$, $(T_p M, \tilde{g})$ is geodesically complete at 0, so by Hopf-Rinow, it is complete. But now, $\exp_p : T_p M \to M$ is a local isometry, and by Proposition 14.6, it is a Riemannian covering map.

2. If $M$ is simply connected, then by Proposition 9.16, the covering map $\exp_p : T_p M \to M$ is a diffeomorphism ($T_p M$ is connected). Therefore, $\exp_p : T_p M \to M$ is a diffeomorphism for all $p \in M$. \qed
Other proofs of Theorem 13.21 can be found in Do Carmo [60] (Chapter 7, Theorem 3.1), Gallot, Hulin and Lafontaine [73] (Chapter 3, Theorem 3.87), Kobayashi and Nomizu [107] (Chapter VIII, Theorem 8.1) and Milnor [125] (Part III, Theorem 19.2).

**Remark:** A version of Theorem 13.21 was first proved by Hadamard and then extended by Cartan.

Theorem 13.21 was generalized by Kobayashi, see Kobayashi and Nomizu [107] (Chapter VIII, Remark 2 after Corollary 8.2). Also, it is shown in Milnor [125] that if $M$ is complete, assuming non-positive sectional curvature, then all homotopy groups $\pi_i(M)$ vanish for $i > 1$, and that $\pi_1(M)$ has no element of finite order except the identity. Finally, non-positive sectional curvature implies that the exponential map does not decrease distance (Kobayashi and Nomizu [107], Chapter VIII, Section 8, Lemma 3).

We now turn to manifolds with strictly positive curvature bounded away from zero and to Myers’ Theorem. The first version of such a theorem was first proved by Bonnet for surfaces with positive sectional curvature bounded away from zero. It was then generalized by Myers in 1941. For these reasons, this theorem is sometimes called the *Bonnet-Myers’ Theorem*. The proof of Myers Theorem involves a beautiful “trick.”

Given any metric space $X$, recall that the *diameter* of $X$ is defined by

$$\text{diam}(X) = \sup \{d(p,q) \mid p, q \in X\}.$$ 

The diameter of $X$ may be infinite.

**Theorem 13.22. (Myers)** Let $M$ be a complete Riemannian manifold of dimension $n$ and assume that

$$\text{Ric}(u,u) \geq (n-1)/r^2, \quad \text{for all unit vectors, } u \in T_pM, \text{ and for all } p \in M,$$

with $r > 0$. Then,

1. The diameter of $M$ is bounded by $\pi r$ and $M$ is compact.
2. The fundamental group of $M$ is finite.

**Proof.** (1) Pick any two points $p, q \in M$ and let $d(p,q) = L$. As $M$ is complete, by Hopf and Rinow’s Theorem, there is a minimal geodesic $\gamma$ joining $p$ and $q$, and by Proposition 13.9, the bilinear index form $I$ associated with $\gamma$ is positive semi-definite, which means that $I(W,W) \geq 0$ for all vector fields $W \in T_{p}\Omega(p,q)$. Pick an orthonormal basis $(e_1, \ldots, e_n)$ of $T_pM$, with $e_1 = \gamma'(0)/L$. Using parallel transport, we get a field of orthonormal frames $(X_1, \ldots, X_n)$ along $\gamma$, with $X_1(t) = \gamma'(t)/L$. Now comes Myers’ beautiful trick. Define new vector fields $Y_i$ along $\gamma$, by

$$W_i(t) = \sin(\pi t)X_i(t), \quad 2 \leq i \leq n.$$
We have
\[ \gamma'(t) = LX_1 \quad \text{and} \quad \frac{DX_i}{dt} = 0. \]
Furthermore, observe that
\[ \frac{DW_i}{dt} = \pi \cos(\pi t)X_i, \quad \frac{D^2 W_i}{dt^2} = -\pi^2 \sin(\pi t)X_i. \]
Then by the second variation formula,
\[ \frac{1}{2} I(W_i, W_i) = -\int_0^1 \langle W_i, \frac{D^2 W_i}{dt^2} + R(\gamma', W_i)\gamma' \rangle dt \]
\[ = -\int_0^1 \langle \sin(\pi t)X_i, -\pi^2 \sin(\pi t)X_i + R(LX_1, \sin(\pi t)X_i)LX_1 \rangle dt \]
\[ = -\int_0^1 \langle \sin(\pi t)X_i, -\pi^2 \sin(\pi t)X_i + L^2 \sin(\pi t)R(X_1, X_i)X_1 \rangle dt \]
\[ = \int_0^1 (\sin(\pi t))^2 (\pi^2 - L^2 \langle R(X_1, X_i)X_1, X_i \rangle) dt, \]
for \( i = 2, \ldots, n \). Adding up these equations and using the fact that
\[ \text{Ric}(X_1(t), X_1(t)) = \sum_{i=2}^n \langle R(X_1(t), X_i(t))X_1(t), X_i(t) \rangle, \]
we get
\[ \frac{1}{2} \sum_{i=2}^n I(W_i, W_i) = \int_0^1 (\sin(\pi t))^2 [(n - 1)\pi^2 - L^2 \text{Ric}(X_1(t), X_1(t))] dt. \]
Now by hypothesis,
\[ \text{Ric}(X_1(t), X_1(t)) \geq (n - 1)/r^2, \]
so
\[ 0 \leq \frac{1}{2} \sum_{i=2}^n I(W_i, W_i) \leq \int_0^1 (\sin(\pi t))^2 \left[ (n - 1)\pi^2 - (n - 1)\frac{L^2}{r^2} \right] dt, \]
which implies \( \frac{L^2}{r^2} \leq \pi^2 \), that is
\[ d(p, q) = L \leq \pi r. \]
As the above holds for every pair of points \( p, q \in M \), we conclude that
\[ \text{diam}(M) \leq \pi r. \]
Since closed and bounded subsets in a complete manifold are compact, \( M \) itself must be compact.

(2) Since the universal covering space \( \tilde{M} \) of \( M \) has the pullback of the metric on \( M \), this metric satisfies the same assumption on its Ricci curvature as that of \( M \). Therefore, \( \tilde{M} \) is also compact, which implies that the fundamental group \( \pi_1(M) \) is finite (see the discussion at the end of Section 9.2). \( \square \)
 Remarks:

(1) The condition on the Ricci curvature cannot be weakened to $\text{Ric}(u, u) > 0$ for all unit vectors. Indeed, the paraboloid of revolution $z = x^2 + y^2$ satisfies the above condition, yet it is not compact.

(2) Theorem 13.22 also holds under the stronger condition that the sectional curvature $K(u, v)$ satisfies

$$K(u, v) \geq \frac{(n - 1)}{r^2},$$

for all orthonormal vectors, $u, v$. In this form, it is due to Bonnet (for surfaces).

It would be a pity not to include in this section a beautiful theorem due to Morse.

**Theorem 13.23.** (Morse Index Theorem) Given a geodesic $\gamma \in \Omega(p, q)$, the index $\lambda$ of the index form $I: T_\gamma \Omega(p, q) \times T_\gamma \Omega(p, q) \to \mathbb{R}$ is equal to the number of points $\gamma(t)$, with $0 \leq t \leq 1$, such that $\gamma(t)$ is conjugate to $p = \gamma(0)$ along $\gamma$, each such conjugate point counted with its multiplicity. The index $\lambda$ is always finite.

As a corollary of Theorem 13.23, we see that there are only finitely many points which are conjugate to $p = \gamma(0)$ along $\gamma$.

A proof of Theorem 13.23 can be found in Milnor [125] (Part III, Section 15) and also in Do Carmo [60] (Chapter 11) or Kobayashi and Nomizu [107] (Chapter VIII, Section 6).

A key ingredient of the proof is that the vector space $T_\gamma \Omega(p, q)$ can be split into a direct sum of subspaces mutually orthogonal with respect to $I$, on one of which (denoted $T'$) $I$ is positive definite. Furthermore, the subspace orthogonal to $T'$ is finite-dimensional. This space is obtained as follows: Since for every point $\gamma(t)$ on $\gamma$, there is some open subset $U_t$ containing $\gamma(t)$ such that any two points in $U_t$ are joined by a unique minimal geodesic, by compactness of $[0, 1]$, there is a subdivision $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0, 1]$ so that $\gamma \restriction [t_i, t_{i+1}]$ lies within an open where it is a minimal geodesic.

Let $T_\gamma \Omega(t_0, \ldots, t_k) \subseteq T_\gamma \Omega(p, q)$ be the vector space consisting of all vector fields $W$ along $\gamma$ such that

1. $W \restriction [t_i, t_{i+1}]$ is a Jacobi field along $\gamma \restriction [t_i, t_{i+1}]$, for $i = 0, \ldots, k - 1$.
2. $W(0) = W(1) = 0$.

The space $T_\gamma \Omega(t_0, \ldots, t_k) \subseteq T_\gamma \Omega(p, q)$ is a finite-dimensional vector space consisting of broken Jacobi fields. Let $T' \subseteq T_\gamma \Omega(p, q)$ be the vector space consisting of all vector fields $W \in T_\gamma \Omega(p, q)$ for which

$$W(t_i) = 0, \quad 0 \leq i \leq k.$$

It is not hard to prove that

$$T_\gamma \Omega(p, q) = T_\gamma \Omega(t_0, \ldots, t_k) \oplus T',$$
that $T_\gamma \Omega(t_0, \ldots, t_k)$ and $T'$ are orthogonal w.r.t $I$, and that $I \mid T'$ is positive definite. The reason why $I(W, W) \geq 0$ for $W \in T'$ is that each segment $\gamma \mid [t_i, t_{i+1}]$ is a minimal geodesic, which has smaller energy than any other path between its endpoints.

As a consequence, the index (or nullity) of $I$ is equal to the index (or nullity) of $I$ restricted to the finite dimensional vector space $T_\gamma \Omega(t_0, \ldots, t_k)$. This shows that the index is always finite.

In the next section we will use conjugate points to give a more precise characterization of the cut locus.

13.7 Cut Locus and Injectivity Radius: Some Properties

As usual, all our manifolds are Riemannian manifolds equipped with the Levi-Civita connection. We begin by reviewing the definition of the cut locus from a slightly different point of view. Let $M$ be a complete Riemannian manifold of dimension $n$. There is a bundle $UM$, called the unit tangent bundle, such that the fibre at any $p \in M$ is the unit sphere $S^{n-1} \subseteq T_pM$ (check the details). As usual, we let $\pi: UM \to M$ denote the projection map which sends every point in the fibre over $p$ to $p$. Then, we have the function

$$\rho: UM \to \mathbb{R},$$

defined so that for all $p \in M$, for all $v \in S^{n-1} \subseteq T_pM$,

$$\rho(v) = \sup_{t \in \mathbb{R} \cup \{\infty\}} d(\pi(v), \exp_p(tv)) = t$$

$$= \sup\{t \in \mathbb{R} \cup \{\infty\} \mid \text{the geodesic } t \mapsto \exp_p(tv) \text{ is minimal on } [0, t]\}.$$

The number $\rho(v)$ is called the cut value of $v$. It can be shown that $\rho$ is continuous, and for every $p \in M$, we let

$$\widetilde{\text{Cut}}(p) = \{\rho(v)v \in T_pM \mid v \in UM \cap T_pM, \rho(v) \text{ is finite}\}$$

be the tangential cut locus of $p$, and

$$\text{Cut}(p) = \exp_p(\widetilde{\text{Cut}}(p))$$

be the cut locus of $p$. The point $\exp_p(\rho(v)v)$ in $M$ is called the cut point of the geodesic $t \mapsto \exp_p(tv)$, and so the cut locus of $p$ is the set of cut points of all the geodesics emanating from $p$.

Also recall from Definition 12.8 that

$$U_p = \{v \in T_pM \mid \rho(v) > 1\},$$
and that $\mathcal{U}_p$ is open and star-shaped. It can be shown that

$$\tilde{\text{Cut}}(p) = \partial \mathcal{U}_p,$$

and the following property holds:

**Theorem 13.24.** If $M$ is a complete Riemannian manifold, then for every $p \in M$, the exponential map $\exp_p$ is a diffeomorphism between $\mathcal{U}_p$ and its image $\exp_p(\mathcal{U}_p) = M - \text{Cut}(p)$ in $M$.

**Proof.** The fact that $\exp_p$ is injective on $\mathcal{U}_p$ was shown in Proposition 12.17. Now, for any $v \in \mathcal{U}$, as $t \mapsto \exp_p(tv)$ is a minimal geodesic for $t \in [0,1]$, by Theorem 13.19 (2), the point $\exp_p v$ is not conjugate to $p$, so $d(\exp_p)_v$ is bijective, which implies that $\exp_p$ is a local diffeomorphism. As $\exp_p$ is also injective, it is a diffeomorphism. \qed

Theorem 13.24 implies that the cut locus is closed.

**Remark:** In fact, $M - \text{Cut}(p)$ can be retracted homeomorphically onto a ball around $p$, and $\text{Cut}(p)$ is a deformation retract of $M - \{p\}$.

The following Proposition gives a rather nice characterization of the cut locus in terms of minimizing geodesics and conjugate points:

**Proposition 13.25.** Let $M$ be a complete Riemannian manifold. For every pair of points $p, q \in M$, the point $q$ belongs to the cut locus of $p$ iff one of the two (not mutually exclusive from each other) properties hold:

(a) There exist two distinct minimizing geodesics from $p$ to $q$.

(b) There is a minimizing geodesic $\gamma$ from $p$ to $q$, and $q$ is the first conjugate point to $p$ along $\gamma$.

A proof of Proposition 13.25 can be found in Do Carmo [60] (Chapter 13, Proposition 2.2) Kobayashi and Nomizu [107] (Chapter VIII, Theorem 7.1) or Klingenberg [104] (Chapter 2, Lemma 2.1.11).

Observe that Proposition 13.25 implies the following symmetry property of the cut locus: $q \in \text{Cut}(p)$ iff $p \in \text{Cut}(q)$. Furthermore, if $M$ is compact, we have

$$p = \bigcap_{q \in \text{Cut}(p)} \text{Cut}(q).$$

Proposition 13.25 admits the following sharpening:

**Proposition 13.26.** Let $M$ be a complete Riemannian manifold. For all $p, q \in M$, if $q \in \text{Cut}(p)$, then:
(a) If among the minimizing geodesics from \( p \) to \( q \), there is one, say \( \gamma \), such that \( q \) is not conjugate to \( p \) along \( \gamma \), then there is another minimizing geodesic \( \omega \neq \gamma \) from \( p \) to \( q \).

(b) Suppose \( q \in \text{Cut}(p) \) realizes the distance from \( p \) to \( \text{Cut}(p) \) (i.e. \( d(p, q) = d(p, \text{Cut}(p)) \)). If there are no minimal geodesics from \( p \) to \( q \) such that \( q \) is conjugate to \( p \) along this geodesic, then there are exactly two minimizing geodesics \( \gamma_1 \) and \( \gamma_2 \) from \( p \) to \( q \), with \( \gamma_2'(1) = -\gamma_1'(1) \). Moreover, if \( d(p, q) = i(M) \) (the injectivity radius), then \( \gamma_1 \) and \( \gamma_2 \) together form a closed geodesic.

Except for the last statement, Proposition 13.26 is proved in Do Carmo [60] (Chapter 13, Proposition 2.12). The last statement is from Klingenberg [104] (Chapter 2, Lemma 2.1.11).

We also have the following characterization of \( \widetilde{\text{Cut}}(p) \):

**Proposition 13.27.** Let \( M \) be a complete Riemannian manifold. For any \( p \in M \), the set of vectors \( u \in \widetilde{\text{Cut}}(p) \) such that is some \( v \in \widetilde{\text{Cut}}(p) \) with \( v \neq u \) and \( \exp_p(u) = \exp_p(v) \) is dense in \( \widetilde{\text{Cut}}(p) \).

Proposition 13.27 is proved in Klingenberg [104] (Chapter 2, Theorem 2.1.14).

We conclude this section by stating a classical theorem of Klingenberg about the injectivity radius of a manifold of bounded positive sectional curvature.

**Theorem 13.28.** (Klingenberg) Let \( M \) be a complete Riemannian manifold and assume that there are some positive constants \( K_{\min}, K_{\max} \), such that the sectional curvature of \( K \) satisfies

\[
0 < K_{\min} \leq K \leq K_{\max}.
\]

Then, \( M \) is compact, and either

(a) \( i(M) \geq \pi / \sqrt{K_{\max}} \), or

(b) There is a closed geodesic \( \gamma \) of minimal length among all closed geodesics in \( M \) and such that

\[
i(M) = \frac{1}{2} L(\gamma).
\]

The proof of Theorem 13.28 is quite hard. A proof using Rauch’s comparison Theorem can be found in Do Carmo [60] (Chapter 13, Proposition 2.13).
Chapter 14

Isometries, Local Isometries, Riemannian Coverings and Submersions, Killing Vector Fields

The goal of this chapter is to understand the behavior of isometries and local isometries, in particular their action on geodesics. In Section 14.1 we show that isometries preserve the Levi-Civita connection. Local isometries preserve all concepts that are local in nature, such as geodesics, the exponential map, sectional, Ricci, and scalar curvature. In Section 14.2 we define Riemannian covering maps. These are smooth covering maps \( \pi: M \rightarrow N \) that are also local isometries. There is a nice correspondence between the geodesics in \( M \) and the geodesics in \( N \). We prove that if \( M \) is complete, \( N \) is connected, and \( \pi: M \rightarrow N \) is a local isometry, then \( \pi \) is a Riemannian covering. In Section 14.3 we introduce Riemannian submersions. Given a submersion \( \pi: M \rightarrow B \) between two Riemannian manifolds \((M, g)\) and \((B, h)\), for every \( b \in B \) in the image of \( \pi \), the fibre \( \pi^{-1}(b) \) is a Riemannian submanifold of \( M \), and for every \( p \in \pi^{-1}(b) \), the tangent space \( T_pM \) to \( M \) at \( p \) splits into the two components

\[
T_pM = \text{Ker} \, d\pi_p \oplus (\text{Ker} \, d\pi_p)^\perp,
\]

where \( \text{V}_p = \text{Ker} \, d\pi_p \) is the vertical subspace of \( T_pM \) and \( \text{H}_p = (\text{Ker} \, d\pi_p)^\perp \) (the orthogonal complement of \( \text{V}_p \) with respect to the metric \( g_p \) on \( T_pM \)) is the horizontal subspace of \( T_pM \). If the map \( d\pi_p \) is an isometry between the horizontal subspace \( \text{H}_p \) of \( T_pM \) and \( T_{\pi(p)}B \) for every \( p \), then \( \pi \) is a Riemannian submersion. In this case most of the differential geometry of \( B \) can be studied by “lifting” from \( B \) to \( M \), and then projecting down to \( B \) again. In Section 14.4 we define Killing vector fields. A Killing vector field \( X \) satisfies the condition

\[
X(\langle Y, Z \rangle) = \langle [X, Y], Z \rangle + \langle Y, [X, Z] \rangle,
\]

for all \( Y, Z \in \mathfrak{X}(M) \). A vector field \( X \) is a Killing vector field iff the diffeomorphisms \( \Phi_t \) induced by the flow \( \Phi \) of \( X \) are isometries (on their domain). Killing vector fields play an important role in the study of reductive homogeneous spaces; see Section 19.4.
14.1 Isometries and Local Isometries

Recall that a local isometry between two Riemannian manifolds $M$ and $N$ is a smooth map $\varphi: M \to N$ so that

$$<(d\varphi)_p(u), (d\varphi)_p(v)>_{\varphi(p)} = <u, v>_p,$$

for all $p \in M$ and all $u, v \in T_p M$. See Definition 10.4. An isometry is a local isometry and a diffeomorphism.

By the inverse function theorem, if $\varphi: M \to N$ is a local isometry, then for every $p \in M$, there is some open subset $U \subseteq M$ with $p \in U$ so that $\varphi \mid U$ is an isometry between $U$ and $\varphi(U)$.

Also recall by Definition 8.4 that if $\varphi: M \to N$ is a diffeomorphism, then for any vector field $X$ on $M$, the vector field $\varphi_* X$ on $N$ (called the push-forward of $X$) is given by

$$(\varphi_* X)_q = d\varphi \varphi^{-1}(q)X(\varphi^{-1}(q)), \quad \text{for all } q \in N,$$

or equivalently, by

$$(\varphi_* X)_{\varphi(p)} = d\varphi_p X(p), \quad \text{for all } p \in M.$$

For any smooth function $h: N \to \mathbb{R}$, for any $q \in N$, we have

$$(\varphi_* X)(h)_q = dh_q((\varphi_* X)(q))$$

$$= dh_q (d\varphi \varphi^{-1}(q)X(\varphi^{-1}(q)))$$

$$= d(h \circ \varphi) \varphi^{-1}(q)X(\varphi^{-1}(q))$$

$$= X(h \circ \varphi) \varphi^{-1}(q),$$

See Figure 14.1.

In other words, we have shown that

$$(\varphi_* X)(h)_q = X(h \circ \varphi) \varphi^{-1}(q),$$

or

$$(\varphi_* X)(h)_{\varphi(p)} = X(h \circ \varphi)_p. \quad (*)$$

It is natural to expect that isometries preserve all “natural” Riemannian concepts and this is indeed the case. We begin with the Levi-Civita connection.

**Proposition 14.1.** If $\varphi: M \to N$ is an isometry, then

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_* X} (\varphi_* Y), \quad \text{for all } X, Y \in \mathfrak{X}(M),$$

where $\nabla_X Y$ is the Levi-Civita connection induced by the metric on $M$ and similarly on $N$. 

14.1. ISOMETRIES AND LOCAL ISOMETRIES

Proof. Let \( X, Y, Z \in \mathfrak{X}(M) \). A proof can be found in O’Neill [138] (Chapter 3, Proposition 59), but we find it instructive to give a proof using the Koszul formula (Proposition 11.8),

\[
2 \langle \nabla X Y, Z \rangle = X(\langle Y, Z \rangle) + Y(\langle X, Z \rangle) - Z(\langle X, Y \rangle) - \langle Y, [X, Z] \rangle - \langle X, [Y, Z] \rangle - \langle Z, [Y, X] \rangle.
\]

We have

\[
(\varphi_*(\nabla X Y))_{\varphi(p)} = d\varphi_p(\nabla X Y)_p,
\]

and as \( \varphi \) is an isometry,

\[
\langle d\varphi_p(\nabla X Y)_p, d\varphi_p Z_p \rangle_{\varphi(p)} = \langle (\nabla X Y)_p, Z_p \rangle_p,
\]

(**)

so Koszul yields

\[
2 \langle \varphi_*(\nabla X Y), \varphi_* Z \rangle_{\varphi(p)} = 2 \langle d\varphi_p(\nabla X Y)_p, d\varphi_p Z_p \rangle_p = 2 \langle (\nabla X Y)_p, Z_p \rangle_p
\]

\[
= X(\langle Y, Z \rangle)_p + Y(\langle X, Z \rangle)_p - Z(\langle X, Y \rangle)_p - \langle Y, [X, Z] \rangle_p - \langle X, [Y, Z] \rangle_p - \langle Z, [Y, X] \rangle_p.
\]

Next, we need to compute

\[
\langle \nabla_{\varphi_*(X)}(\varphi_* Y), \varphi_* Z \rangle_{\varphi(p)}.
\]
When we plug $\varphi_* X$, $\varphi_* Y$ and $\varphi_* Z$ into the Koszul formula, as $\varphi$ is an isometry, for the fourth term on the right-hand side we get

$$\langle \varphi_* Y, [\varphi_* X, \varphi_* Z] \rangle_{\varphi(p)} = \langle d\varphi_p Y_p, [d\varphi_p X_p, d\varphi_p Z_p] \rangle_{\varphi(p)}$$

$$= \langle d\varphi_p Y_p, d\varphi_p [X_p, Z_p] \rangle_{\varphi(p)}, \quad \text{by Proposition 8.5}$$

$$= \langle Y_p, [X_p, Z_p] \rangle_p, \quad \text{by (**) }$$

and similarly for the fifth and sixth term on the right-hand side. For the first term on the right-hand side, we get

$$(\varphi_* X)(\langle \varphi_* Y, \varphi_* Z \rangle)_{\varphi(p)} = (\varphi_* X)(\langle d\varphi_p Y_p, d\varphi_p Z_p \rangle)_{\varphi(p)}$$

$$= (\varphi_* X)(\langle Y_p, Z_p \rangle_{\varphi^{-1}(\varphi(p))})_{\varphi(p)}, \quad \text{by (**) }$$

$$= (\varphi_* X)(\langle Y, Z \rangle \circ \varphi^{-1})_{\varphi(p)}$$

$$= X((Y, Z) \circ \varphi^{-1} \circ \varphi)_p, \quad \text{by (*) }$$

$$= X((Y, Z))_p,$$

and similarly for the second and third term. Consequently, we get

$$2\langle \nabla_{\varphi_* X} (\varphi_* Y), \varphi_* Z \rangle_{\varphi(p)} = X((Y, Z)_p) + Y((X, Z)_p) - Z((X, Y)_p)$$

$$- \langle Y, [X, Z] \rangle_p - \langle X, [Y, Z] \rangle_p - \langle Z, [Y, X] \rangle_p.$$ 

By comparing right-hand sides, we get

$$2\langle \varphi_*(\nabla_X Y), \varphi_* Z \rangle_{\varphi(p)} = 2\langle \nabla_{\varphi_* X} (\varphi_* Y), \varphi_* Z \rangle_{\varphi(p)}$$

for all $X, Y, Z$, and as $\varphi$ is a diffeomorphism, this implies

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_* X} (\varphi_* Y),$$

as claimed. \qed

As a corollary of Proposition 14.1, the curvature induced by the connection is preserved; that is

$$\varphi_* R(X, Y) Z = R(\varphi_* X, \varphi_* Y) \varphi_* Z,$$

as well as the parallel transport, the covariant derivative of a vector field along a curve, the exponential map, sectional curvature, Ricci curvature and geodesics.

Actually, all concepts that are local in nature are preserved by local isometries! So, except for the Levi-Civita connection and the Riemann tensor on vector fields, all the above concepts are preserved under local isometries. For the record we state:

**Proposition 14.2.** If $\varphi: M \to N$ is a local isometry between two Riemannian manifolds equipped with the Levi-Civita connection, then the following concepts are preserved:
(1) The covariant derivative of vector fields along a curve $\gamma$; that is
\[ d\varphi_{\gamma(t)} \frac{DX}{dt} = D\varphi_\ast X \frac{dt}{dt}, \]
for any vector field $X$ along $\gamma$, with $(\varphi_\ast X)(t) = d\varphi_{\gamma(t)} Y(t)$, for all $t$.

(2) Parallel translation along a curve. If $P_\gamma$ denotes parallel transport along the curve $\gamma$ and if $P_{\varphi \circ \gamma}$ denotes parallel transport along the curve $\varphi \circ \gamma$, then
\[ d\varphi_{\gamma(t)} \circ P_\gamma = P_{\varphi \circ \gamma} \circ d\varphi_{\gamma(0)}. \]

(3) Geodesics. If $\gamma$ is a geodesic in $M$, then $\varphi \circ \gamma$ is a geodesic in $N$. Thus, if $\gamma_v$ is the unique geodesic with $\gamma(0) = p$ and $\gamma_v'(0) = v$, then
\[ \varphi \circ \gamma_v = \gamma_{d\varphi_p v}, \]
wherever both sides are defined. Note that the domain of $\gamma_{d\varphi_p v}$ may be strictly larger than the domain of $\gamma_v$. For example, consider the inclusion of an open disc into $\mathbb{R}^2$.

(4) Exponential maps. We have
\[ \varphi \circ \exp_p = \exp_{\varphi(p)} \circ d\varphi_p, \]
wherever both sides are defined. See Figure 14.2.

(5) Riemannian curvature tensor. We have
\[ d\varphi_p R(x, y)z = R(d\varphi_p x, d\varphi_p y)d\varphi_p z, \quad \text{for all } x, y, z \in T_p M. \]

(6) Sectional, Ricci, and Scalar curvature. We have
\[ K(d\varphi_p x, d\varphi_p y) = K(x, y)_p, \]
for all linearly independent vectors $x, y \in T_p M$;
\[ \text{Ric}(d\varphi_p x, d\varphi_p y) = \text{Ric}(x, y)_p \]
for all $x, y \in T_p M$;
\[ S_M = S_N \circ \varphi. \]
where $S_M$ is the scalar curvature on $M$ and $S_N$ is the scalar curvature on $N$.

A useful property of local isometries is stated below. For a proof, see O’Neill [138] (Chapter 3, Proposition 62):

**Proposition 14.3.** Let $\varphi, \psi : M \to N$ be two local isometries. If $M$ is connected and if $\varphi(p) = \psi(p)$ and $d\varphi_p = d\psi_p$ for some $p \in M$, then $\varphi = \psi$.

The idea is to prove that
\[ \{ p \in M \mid d\varphi_p = d\psi_p \} \]
is both open and closed, and for this, to use the preservation of the exponential under local diffeomorphisms.
14.2 Riemannian Covering Maps

The notion of covering map discussed in Section 9.2 (see Definition 9.5) can be extended to Riemannian manifolds.

**Definition 14.1.** If $M$ and $N$ are two Riemannian manifold, then a map $\pi: M \to N$ is a Riemannian covering iff the following conditions hold:

1. The map $\pi$ is a smooth covering map.
2. The map $\pi$ is a local isometry.

Recall from Section 9.2 that a covering map is a local diffeomorphism. A way to obtain a metric on a manifold $M$ is to pull-back the metric $g$ on a manifold $N$ along a local diffeomorphism $\varphi: M \to N$ (see Section 10.2). If $\varphi$ is a covering map, then it becomes a Riemannian covering map.

**Proposition 14.4.** Let $\pi: M \to N$ be a smooth covering map. For any Riemannian metric $g$ on $N$, there is a unique metric $\pi^*g$ on $M$, so that $\pi$ is a Riemannian covering.
Proof. We define the pull-back metric $\pi^*g$ on $M$ induced by $g$ as follows: For all $p \in M$, for all $u, v \in T_pM$,

$$(\pi^*g)_p(u, v) = g(d\pi_p(u), d\pi_p(v)).$$

We need to check that $(\pi^*g)_p$ is an inner product, which is very easy since $d\pi_p$ is a linear isomorphism. Our map $\pi$ between the two Riemannian manifolds $(M, \pi^*g)$ and $(N, g)$ becomes a local isometry. Now, every metric on $M$ making $\pi$ a local isometry has to satisfy the equation defining $\pi^*g$, so this metric is unique. 

In general, if $\pi: M \to N$ is a smooth covering map, a metric on $M$ does not induce a metric on $N$ such that $\pi$ is a Riemannian covering. However, if $N$ is obtained from $M$ as a quotient by some suitable group action (by a group $G$) on $M$, then the projection $\pi: M \to M/G$ is a Riemannian covering.

In the rest of this section, we assume that our Riemannian manifolds are equipped with the Levi-Civita connection. Because a Riemannian covering map is a local isometry, we have the following useful result.

**Proposition 14.5.** Let $\pi: M \to N$ be a Riemannian covering. Then, the geodesics of $(N, h)$ are the projections of the geodesics of $(M, g)$ (curves of the form $\pi \circ \gamma$, where $\gamma$ is a geodesic in $M$), and the geodesics of $(M, g)$ are the liftings of the geodesics of $(N, h)$ (curves $\gamma$ in $M$ such that $\pi \circ \gamma$ is a geodesic of $(N, h)$).

As a corollary of Proposition 14.4 and Theorem 9.13, every connected Riemannian manifold $M$ has a simply connected covering map $\tilde{\pi}: \tilde{M} \to M$, where $\pi$ is a Riemannian covering. Furthermore, if $\pi: M \to N$ is a Riemannian covering and $\varphi: P \to N$ is a local isometry, it is easy to see that its lift $\tilde{\varphi}: P \to M$ is also a local isometry. In particular, the deck-transformations of a Riemannian covering are isometries.

In general a local isometry is not a Riemannian covering. However, this is the case when the source space is complete.

**Proposition 14.6.** Let $\pi: M \to N$ be a local isometry with $N$ connected. If $M$ is a complete manifold, then $\pi$ is a Riemannian covering map.

Proof. We follow the proof in Sakai [150] (Chapter III, Theorem 5.4). Because $\pi$ is a local isometry, geodesics in $M$ can be projected onto geodesics in $N$ and geodesics in $N$ can be lifted back to $M$. The proof makes heavy use of these facts.

First, we prove that $N$ is complete. Pick any $p \in M$ and let $q = \pi(p)$. For any geodesic $\gamma_v$ of $N$ with initial point $q \in N$ and initial direction the unit vector $v \in T_qN$, consider the geodesic $\tilde{\gamma}_u$ of $M$ with initial point $p$, and with $u = d\pi_q^{-1}(v) \in T_pM$. As $\pi$ is a local isometry, it preserves geodesics, so

$$\gamma_v = \pi \circ \tilde{\gamma}_u,$$

and since $\tilde{\gamma}_u$ is defined on $\mathbb{R}$ because $M$ is complete, so is $\gamma_v$. As $\exp_q$ is defined on the whole of $T_qN$, by Hopf-Rinow, $N$ is complete. See Figure 14.3.
Next, we prove that $\pi$ is surjective. As $N$ is complete, for any $q_1 \in N$, there is a minimal geodesic $\gamma: [0, b] \to N$ joining $q$ to $q_1$ and for the geodesic $\tilde{\gamma}$ in $M$ emanating from $p$ and with initial direction $d\pi^{-1}_q(\gamma'(0))$, we have $\pi(\tilde{\gamma}(b)) = \gamma(b) = q_1$, establishing surjectivity.

For any $q \in N$, pick $r > 0$ with $r < i(q)$, where $i(q)$ denotes the injectivity radius of $N$ at $q$ and consider the open metric ball $B_r(q) = \exp_q(B(0_q, r))$ (where $B(0_q, r)$ is the open ball of radius $r$ in $T_q N$). Let

$$\pi^{-1}(q) = \{p_i\}_{i \in I} \subseteq M.$$  

We claim that the following properties hold:

1. If we write $B_r(p_i) = \exp_{p_i}(B(0_{p_i}, r))$, then each map $\pi \upharpoonright B_r(p_i): B_r(p_i) \to B_r(q)$ is a diffeomorphism, in fact an isometry.

2. $\pi^{-1}(B_r(q)) = \bigcup_{i \in I} B_r(p_i)$.

3. $B_r(p_i) \cap B_r(p_j) = \emptyset$ whenever $i \neq j$.

It follows from (1), (2) and (3) that $B_r(q)$ is evenly covered by the family of open sets $\{B_r(p_i)\}_{i \in I}$, so $\pi$ is a covering map.

1. Since $\pi$ is a local isometry, it maps geodesics emanating from $p_i$ to geodesics emanating...
from \(q\), so the following diagram commutes:

\[
\begin{array}{ccc}
B(0_{p_i}, r) & \xrightarrow{d\pi_{p_i}} & B(0_q, r) \\
\exp_{p_i} & & \exp_q \\
B_r(p_i) & \xrightarrow{\pi} & B_r(q).
\end{array}
\]

See Figure 14.3. Since \(\exp_q \circ d\pi_{p_i}\) is a diffeomorphism, \(\pi \upharpoonright B_r(p_i)\) must be injective, and since \(\exp_{p_i}\) is surjective, so is \(\pi \upharpoonright B_r(p_i)\). Then, \(\pi \upharpoonright B_r(p_i)\) is a bijection, and as \(\pi\) is a local diffeomorphism, \(\pi \upharpoonright B_r(p_i)\) is a diffeomorphism.

(2) Obviously, \(\bigcup_{i \in I} B_r(p_i) \subseteq \pi^{-1}(B_r(q))\), by (1). Conversely, pick \(p_1 \in \pi^{-1}(B_r(q))\). For \(q_1 = \pi(p_1)\), we can write \(q_1 = \exp_q v\), for some \(v \in B(0_q, r)\), and the map \(\gamma(t) = \exp_q (1-t)v\), for \(t \in [0,1]\), is a geodesic in \(N\) joining \(q_1\) to \(q\). Then, we have the geodesic \(\tilde{\gamma}\) emanating from \(p_1\) with initial direction \(d\pi_{q_1}^{-1}(\gamma'(0))\), and as \(\pi \circ \tilde{\gamma}(1) = \gamma(1) = q\), we have \(\tilde{\gamma}(1) = p_i\) for some \(\alpha\). Since \(\gamma\) has length less than \(r\), we get \(p_1 \in B_i(p_i)\).

(3) Suppose \(p_1 \in B_r(p_i) \cap B_i(p_j)\). We can pick a minimal geodesic \(\tilde{\gamma}\), in \(B_r(p_i)\) (resp. \(\tilde{\omega}\) in \(B_i(p_j)\)) joining \(p_i\) to \(p\) (resp. joining \(p_j\) to \(p\)). Then, the geodesics \(\pi \circ \tilde{\gamma}\) and \(\pi \circ \tilde{\omega}\) are geodesics in \(B_r(q)\) from \(q\) to \(\pi(p_1)\), and their length is less than \(r\). Since \(r < i(q)\), these geodesics are minimal so they must coincide. Therefore, \(\gamma = \omega\), which implies \(i = j\). \(\square\)

### 14.3 Riemannian Submersions

Let \(\pi : M \to B\) be a submersion between two Riemannian manifolds \((M,g)\) and \((B,h)\). For every \(b \in B\) in the image of \(\pi\), the fibre \(\pi^{-1}(b)\) is a Riemannian submanifold of \(M\), and for every \(p \in \pi^{-1}(b)\), the tangent space \(T_p \pi^{-1}(b)\) to \(\pi^{-1}(b)\) at \(p\) is \(\text{Ker} \, d\pi_p\). The tangent space \(T_p M\) to \(M\) at \(p\) splits into the two components

\[
T_p M = \text{Ker} \, d\pi_p \oplus (\text{Ker} \, d\pi_p)^\perp,
\]

where \(\mathcal{V}_p = \text{Ker} \, d\pi_p\) is the vertical subspace of \(T_p M\) and \(\mathcal{H}_p = (\text{Ker} \, d\pi_p)^\perp\) (the orthogonal complement of \(\mathcal{V}_p\) with respect to the metric \(g_p\) on \(T_p M\)) is the horizontal subspace of \(T_p M\).

Any tangent vector \(u \in T_p M\) can be written uniquely as

\[
u = u_H + u_V,
\]

with \(u_H \in \mathcal{H}_p\), called the horizontal component of \(u\), and \(u_V \in \mathcal{V}_p\), called the vertical component of \(u\); see Figure 14.4. A tangent vector \(u \in T_p M\) is said to be horizontal iff \(u \in \mathcal{H}_p\) (equivalently iff \(u_V = 0\)).

Because \(\pi\) is a submersion, \(d\pi_p\) gives a linear isomorphism between \(\mathcal{H}_p\) and \(T_{\pi(p)}B\). If \(d\pi_p\) is an isometry, then most of the differential geometry of \(B\) can be studied by “lifting” from \(B\) to \(M\).
Definition 14.2. A map $\pi : M \to B$ between two Riemannian manifolds $(M, g)$ and $(B, h)$ is a Riemannian submersion if the following properties hold:

1. The map $\pi$ is a smooth submersion.

2. For every $p \in M$, the map $d\pi_p$ is an isometry between the horizontal subspace $\mathcal{H}_p$ of $T_pM$ and $T_{\pi(p)}B$.

We will see later that Riemannian submersions arise when $B$ is a reductive homogeneous space, or when $B$ is obtained from a free and proper action of a Lie group acting by isometries on $B$.

If $\pi : M \to B$ is a Riemannian submersion which is surjective onto $B$, then every vector field $X$ on $B$ has a unique horizontal lift $\overline{X}$ on $M$, defined such that for every $b \in B$ and every $p \in \pi^{-1}(b)$,

$$\overline{X}(p) = (d\pi_p)^{-1}X(b).$$

Since $d\pi_p$ is an isomorphism between $\mathcal{H}_p$ and $T_bB$, the above condition can be written

$$d\pi \circ \overline{X} = X \circ \pi,$$

which means that $\overline{X}$ and $X$ are $\pi$-related (see Definition 8.5).

The following proposition is proved in O’Neill [138] (Chapter 7, Lemma 45) and Gallot, Hulin, Lafontaine [73] (Chapter 2, Proposition 2.109).
Proposition 14.7. Let \( \pi: M \to B \) be a Riemannian submersion between two Riemannian manifolds \((M,g)\) and \((B,h)\) equipped with the Levi-Civita connection.

1. If \( \gamma \) is a geodesic in \( M \) such that \( \gamma'(0) \) is a horizontal vector, then \( \gamma \) is horizontal geodesic in \( M \) (which means that \( \gamma'(t) \) is a horizontal vector for all \( t \)), and \( c = \pi \circ \gamma \) is a geodesic in \( B \) of the same length than \( \gamma \). See Figure 14.5.

2. For every \( p \in M \), if \( c \) is a geodesic in \( B \) such that \( c(0) = \pi(p) \), then for some \( \epsilon \) small enough, there is a unique horizontal lift \( \gamma \) of the restriction of \( c \) to \([-\epsilon, \epsilon]\), and \( \gamma \) is a geodesic of \( M \).

Furthermore, if \( \pi: M \to B \) is surjective, then:

3. For any two vector fields \( X, Y \in \mathfrak{X}(B) \), we have
   
   (a) \( \langle X, Y \rangle = \langle X, Y \rangle \circ \pi \).
   (b) \( [X, Y]_H = [X, Y] \).
   (c) \( (\nabla_{\pi Y} X)_H = \nabla_{X Y}, \) where \( \nabla \) is the Levi–Civita connection on \( M \).

4. If \( M \) is complete, then \( B \) is also complete.

Proof. We prove (1) and (2), following Gallot, Hulin, Lafontaine [73] (Proposition 2.109). We begin with (2). We claim that a Riemannian submersion shortens distance. More precisely, given any two points \( p_1, p_2 \in M \),

\[
d_B(\pi(p_1), \pi(p_2)) \leq d_M(p_1, p_2),
\]

where \( d_M \) is the Riemannian distance on \( M \) and \( d_B \) is the Riemannian distance on \( B \). It suffices to prove that if \( \gamma \) is a curve of \( M \), then \( L(\gamma) \geq L(\pi \circ \gamma) \). For any \( p \in M \), every tangent vector \( u \in T_p M \) can be written uniquely as an orthogonal sum \( u = u_H + u_V \), and since \( d\pi_p \) is an isometry between \( \mathcal{H}_p \) and \( T_{\pi(p)}B \), we have

\[
\|u\|^2 = \|u_H\|^2 + \|u_V\|^2 \geq \|u_H\|^2 = \|d\pi_p(u_H)\|^2 = \|d\pi_p(u)\|^2.
\]

This implies that

\[
L(\gamma) = \int_0^1 \|\gamma'(t)\| \, dt \geq \int_0^1 \|\pi \circ \gamma'(t)\| \, dt = L(\pi \circ \gamma),
\]

as claimed.

For any \( p \in M \), let \( c \) be a geodesic through \( b = \pi(p) \) for \( t = 0 \). For \( \epsilon \) small enough, the exponential map \( \exp_b \) is a diffeomorphism, so \( W = c((-\epsilon, \epsilon)) \) is a one-dimensional submanifold of \( B \). Since \( \pi \) is a submersion, \( V = \pi^{-1}(W) \) is a submanifold of \( M \). Define a horizontal vector field \( X \) on \( V \) by

\[
X(q) = (d\pi_q)^{-1}(c'(\pi(q))), \quad q \in V,
\]
Figure 14.5: An illustration of Part (1), Proposition 14.7. Both $\gamma$ and $c$ are equal length geodesics in $M$ and $B$ respectively. All the tangent vectors to $\gamma$ lie in horizontal subspace.

where $d\pi_q$ is the isomorphism between $H_q$ and $T_{\pi(q)}B$. For any $q \in V$, there is a unique integral curve $\gamma_q$ through $q$. In particular, $p \in V$, so the curve $\gamma_p$ is defined near 0. We claim that it is a geodesic. This is because, first $\|\gamma'(t)\| = \|c'(t)\|$ is a constant, and second, for $s$ small enough, the curve $\gamma$ is locally minimal, that is

$$L(\gamma) \left|_{[t,t+s]} \right. = L(c) \left|_{[t,t+s]} \right. = d(c(t), c(t+s)) \leq d(\gamma(t), \gamma(t+s)).$$

See Figure 14.6.

We can now prove (1). Let $\gamma$ be a geodesic through $p = \gamma(0)$ such that $\gamma'(0)$ is a horizontal vector, and write $b = \pi(p)$ and $u = d\pi_p(\gamma'(0))$. Let $c$ be the unique geodesic of $B$ such that $c(0) = b$ and $c'(0) = u$. By (2) we have a horizontal lift $\tilde{\gamma}$ of $c$ starting at $p$, and we know it is a geodesic. By construction, $\tilde{\gamma}'(0) = \gamma'(0)$, so by uniqueness $\gamma$ and $\tilde{\gamma}$ coincide on their common domain of definition. It follows that the set of parameters where the geodesic $\gamma$ is
horizontal, and where it is a lift of $c$ is an open subset containing 0. These two conditions being also closed, they must be satisfied on the maximal interval of definition of $\gamma$. It is now obvious that $c = \pi \circ \gamma$, a geodesic in $B$ of the same length as $\gamma$. \qed

In (2), we can’t expect in general that the whole geodesic $c$ in $B$ can be lifted to $M$. This is because the manifold $(B, h)$ may be complete but $(M, g)$ may not be. For example, consider the inclusion map $\pi: (\mathbb{R}^2 - \{0\}) \to \mathbb{R}^2$, with the canonical Euclidean metrics.

An example of a Riemannian submersion is $\pi: S^{2n+1} \to \mathbb{CP}^n$, where $S^{2n+1}$ has the canonical metric and $\mathbb{CP}^n$ has the Fubini–Study metric.

Remark: It shown in Petersen [140] (Chapter 3, Section 5), that the connection $\nabla_X Y$ on $M$ is given by

$$\nabla_X Y = \nabla_X Y + \frac{1}{2}[X, Y]_\nabla.$$ 

14.4 Isometries and Killing Vector Fields

If $X$ is a vector field on a manifold $M$, then we saw that we can define the notion of Lie derivative for vector fields ($L_X Y = [X, Y]$) and for functions ($L_X f = X(f)$). It is possible
to generalize the notion of Lie derivative to an arbitrary tensor field \( S \) (see Section 23.3). In particular, if \( S = g \) (the metric tensor), the Lie derivative \( L_X g \) is defined by

\[
L_X g(Y, Z) = X(\langle Y, Z \rangle) - \langle [X, Y], Z \rangle - \langle Y, [X, Z] \rangle,
\]

with \( X, Y, Z \in \mathfrak{X}(M) \), and where we write \( \langle X, Y \rangle \) and \( g(X, Y) \) interchangeably. If \( \Phi_t \) is an isometry (on its domain), where \( \Phi \) is the global flow associated with the vector field \( X \), then \( \Phi_t^* (g) = g \), and it can be shown that this implies that \( L_X g = 0 \). In fact, we have the following result proved in O’Neill [138] (Chapter 9, Proposition 23).

**Proposition 14.8.** For any vector field \( X \) on a Riemannian manifold \((M, g)\), the diffeomorphisms \( \Phi_t \) induced by the flow \( \Phi \) of \( X \) are isometries (on their domain) iff \( L_X g = 0 \).

Informally, Proposition 14.8 says that \( L_X g \) measures how much the vector field \( X \) changes the metric \( g \).

**Definition 14.3.** Given a Riemannian manifold \((M, g)\), a vector field \( X \) is a *Killing vector field* iff the Lie derivative of the metric vanishes; that is, \( L_X g = 0 \).

Killing vector fields play an important role in the study of reductive homogeneous spaces; see Section 19.4. They also interact with the Ricci curvature and play a crucial role in the Bochner technique; see Petersen [140] (Chapter 7).

As the notion of Lie derivative, the notion of covariant derivative \( \nabla_X Y \) of a vector field \( Y \) in the direction \( X \) can be generalized to tensor fields (see Section 29.1). In particular, the covariant derivative \( \nabla_X g \) of the Riemannian metric \( g \) on a manifold \( M \) turns out to be given by

\[
\nabla_X (g)(Y, Z) = X(\langle Y, Z \rangle) - \langle \nabla_X Y, Z \rangle - \langle Y, \nabla_X Z \rangle,
\]

for all \( X, Y, Z \in \mathfrak{X}(M) \) (see Proposition 29.5). In this section, we adopt the above formula as the definition of \( \nabla_X (g)(Y, Z) \). Then, observe that the connection \( \nabla \) on \( M \) is compatible with \( g \) iff \( \nabla_X (g) = 0 \) for all \( X \). We define the covariant derivative \( \nabla X \) of a vector field \( X \) as the \((1,1)\)-tensor defined so that

\[
(\nabla X)(Y) = \nabla_Y X
\]

for all \( X, Y \in \mathfrak{X}(M) \). The above facts imply the following Proposition.

**Proposition 14.9.** Let \((M, g)\) be a Riemannian manifold and let \( \nabla \) be the Levi–Civita connection on \( M \) induced by \( g \). For every vector field \( X \) on \( M \), the following conditions are equivalent:

1. \( X \) is a Killing vector field; that is, \( L_X g = 0 \).
2. \( X(\langle Y, Z \rangle) = \langle [X, Y], Z \rangle + \langle Y, [X, Z] \rangle \) for all \( Y, Z \in \mathfrak{X}(M) \).
(3) $\langle \nabla_Y X, Z \rangle + \langle \nabla_Z X, Y \rangle = 0$ for all $Y, Z \in \mathfrak{X}(M)$; that is, $\nabla X$ is skew-adjoint relative to $g$.

Proof. Since

$$L_X g(Y, Z) = X(\langle Y, Z \rangle) - \langle [X, Y], Z \rangle - \langle Y, [X, Z] \rangle,$$

the equivalence of (1) and (2) is clear.

Since $\nabla$ is the Levi–Civita connection, we have $\nabla_X g = 0$, so

$$X(\langle Y, Z \rangle) - \langle \nabla_X Y, Z \rangle - \langle Y, \nabla_X Z \rangle = 0,$$

which yields

$$\langle [X, Y], Z \rangle + \langle Y, [X, Z] \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle.$$

Since $\nabla$ is also torsion-free we have

$$\nabla_X Y - \nabla_Y X = [X, Y]$$

$$\nabla_X Z - \nabla_Z X = [X, Z],$$

so we get

$$\langle [X, Y], Z \rangle + \langle Y, [X, Z] \rangle = \langle \nabla_X Y, Z \rangle + \langle Y, \nabla_X Z \rangle$$

$$= \langle \nabla_Y X, Z \rangle + \langle Y, \nabla_Z X \rangle + \langle [X, Y], Z \rangle + \langle Y, [X, Z] \rangle,$$

that is,

$$\langle \nabla_Y X, Z \rangle + \langle \nabla_Z X, Y \rangle = 0.$$

This proves that (2) and (3) are equivalent.

Condition (3) shows that any parallel vector field is a Killing vector field.

Remark: It can be shown that if $\gamma$ is any geodesic in $M$, then the restriction $X_\gamma$ of any Killing vector field $X$ to $\gamma$ is a Jacobi field (see Section 13.5), and that $\langle X, \gamma' \rangle$ is constant along $\gamma$ (see O'Neill [138], Chapter 9, Lemma 26).
Chapter 15

Lie Groups, Lie Algebras, and the Exponential Map

In Chapter 4, we defined the notion of a Lie group as a certain type of manifold embedded in $\mathbb{R}^N$, for some $N \geq 1$. Now that we have the general concept of a manifold, we can define Lie groups in more generality. Now, if every Lie group was a linear group (a group of matrices), then there would be no need for a more general definition. However, there are Lie groups that are not matrix groups, although it is not a trivial task to exhibit such groups and to prove that they are not matrix groups.

An example of a Lie group which is not a matrix group described in Hall [84] is $G = \mathbb{R} \times \mathbb{R} \times S^1$, with the multiplication given by

$$(x_1, y_1, u_1) \cdot (x_2, y_2, u_2) = (x_1 + x_2, y_1 + y_2, e^{ix_1 y_2} u_1 u_2).$$

If we define the group $H$ (the Heisenberg group) as the group of $3 \times 3$ upper triangular matrices given by

$$H = \left\{ \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \mid a, b, c \in \mathbb{R} \right\},$$

then it is easy to show that the map $\varphi: H \to G$ given by

$$\varphi \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} = (a, c, e^{ib})$$

is a surjective group homomorphism. It is easy to check that the kernel of $\varphi$ is the discrete group

$$N = \left\{ \begin{pmatrix} 1 & 0 & k2\pi \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mid k \in \mathbb{Z} \right\}.$$
Both groups $H$ and $N$ are matrix groups, yet $G = H/N$ is a Lie group and it can be shown using some representation theory that $G$ is not a matrix group (see Hall [84], Appendix C.3).

Other examples of Lie groups that are not matrix groups are obtained by considering the universal cover $\tilde{\text{SL}}(2, \mathbb{R})$ of $\text{SL}(2, \mathbb{R})$ for $n \geq 2$. The group $\text{SL}(2, \mathbb{R})$ is a matrix group which is not simply connected for $n \geq 2$, and its universal cover $\tilde{\text{SL}}(2, \mathbb{R})$ is a Lie groups which is not a matrix group; see Hall [84] (Appendix C.3) or Ziller [180] (Example 2.22).

Given a Lie group $G$ (not necessarily a matrix group) we begin by defining the Lie bracket on the tangent space $\mathfrak{g} = T_1 G$ at the identity in terms of the adjoint representation of $G$

$$\text{Ad}: G \to \text{GL}(\mathfrak{g}),$$

and its derivative at 1, the adjoint representation of $\mathfrak{g}$,

$$\text{ad}: \mathfrak{g} \to \text{gl}(\mathfrak{g});$$

namely, $[u, v] = \text{ad}(u)(v)$.

In Section 15.2, we define left and right invariant vector fields on a Lie group. The map $X \mapsto X(1)$ establishes an isomorphism between the space of left-invariant (resp. right-invariant) vector fields on $G$ and $\mathfrak{g}$. Then, by considering integral curves of left-invariant vector fields, we define the generalization of the exponential map $\exp: \mathfrak{g} \to G$ to arbitrary Lie groups that are not necessarily matrix groups. We prove some fundamental properties of the exponential map.

In Section 15.3, we revisit homomorphisms of Lie groups and Lie algebras and generalize certain results shown for matrix groups to arbitrary Lie groups. We also define immersed Lie subgroups and (closed) Lie subgroups.

In Section 15.4, we explore the correspondence between Lie groups and Lie algebras and state some of the Lie theorems.

Section 15.5 is devoted to semidirect products of Lie algebras and Lie Groups. These are constructions that generalize the notion of direct sum (for Lie algebra) and direct products (for Lie groups). For example, the Lie algebra $\mathfrak{se}(n)$ is the semidirect product of $\mathbb{R}^n$ and $\mathfrak{so}(n)$, and the Lie group $\text{SE}(n)$ is the semidirect product of $\mathbb{R}^n$ and $\text{SO}(n)$.

The notion of universal covering group of a Lie group is described in Section 15.6.

In Section 15.7, we show that The Killing vector fields on a Riemannian manifold $M$ form a Lie algebra. We also describe the relationship between the Lie algebra of complete Killing vector fields and the Lie algebra of the isometry group $\text{Isom}(M)$ of the manifold $M$.

Besides classic references on Lie groups and Lie algebras, such as Chevalley [41], Knapp [106], Warner [175], Duistermaat and Kolk [64], Bröcker and tom Dieck [31], Sagle and Walde [149], Helgason [88], Serre [160, 159], Kirillov [102], Fulton and Harris [70], and Bourbaki [28], one should be aware of more introductory sources and surveys such as Tapp [167], Kosmann [108], Hall [84], Sattinger and Weaver [154], Carter, Segal and Macdonald [38], Curtis [46], Baker [16], Rossmann [146], Bryant [32], Mneimné and Testard [130] and Arvanitoyeorgos [11].
15.1 Lie Groups and Lie Algebras

We begin our study of Lie groups by generalizing Definition 4.5.

Definition 15.1. A Lie group is a nonempty subset $G$ satisfying the following conditions:

(a) $G$ is a group (with identity element denoted $e$ or 1).

(b) $G$ is a smooth manifold.

(c) $G$ is a topological group. In particular, the group operation $\cdot : G \times G \to G$ and the inverse map $^{-1}: G \to G$ are smooth.

Remark: The smoothness of inversion follows automatically from the smoothness of multiplication. This can be shown by applying the inverse function theorem to the map $(g, h) \mapsto (g, gh)$, from $G \times G$ to $G \times G$.

We have already met a number of Lie groups: $\text{GL}(n, \mathbb{R})$, $\text{GL}(n, \mathbb{C})$, $\text{SL}(n, \mathbb{R})$, $\text{SL}(n, \mathbb{C})$, $\text{O}(n)$, $\text{SO}(n)$, $\text{U}(n)$, $\text{SU}(n)$, $\text{E}(n, \mathbb{R})$, $\text{SO}(n, 1)$. Also, every linear Lie group of $\text{GL}(n, \mathbb{R})$ (see Definition 4.6) is a Lie group.

We saw in the case of linear Lie groups that the tangent space to $G$ at the identity $g = T_1 G$ plays a very important role. In particular, this vector space is equipped with a (non-associative) multiplication operation, the Lie bracket, that makes $g$ into a Lie algebra. This is again true in this more general setting.

Recall that Lie algebras are defined as follows:

Definition 15.2. A (real) Lie algebra $\mathcal{A}$ is a real vector space together with a bilinear map $[\cdot, \cdot] : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$, called the Lie bracket on $\mathcal{A}$, such that the following two identities hold for all $a, b, c \in \mathcal{A}$:

$$[a, a] = 0,$$

and the so-called Jacobi identity:

$$[a, [b, c]] + [c, [a, b]] + [b, [c, a]] = 0.$$ 

It is immediately verified that $[b, a] = -[a, b]$.

For every $a \in g$, it is customary to define the linear map $\text{ad}(a): g \to g$ by

$$\text{ad}(a)(b) = [a, b], \quad b \in g.$$ 

The map $\text{ad}(a)$ is also denoted $\text{ad}_a$ or $\text{ad} a$. Let us also recall the definition of homomorphisms of Lie groups and Lie algebras.
Definition 15.3. Given two Lie groups \( G_1 \) and \( G_2 \), a homomorphism (or map) of Lie groups is a function \( f: G_1 \to G_2 \) which is a homomorphism of groups, and a smooth map (between the manifolds \( G_1 \) and \( G_2 \)). Given two Lie algebras \( A_1 \) and \( A_2 \), a homomorphism (or map) of Lie algebras is a function \( f: A_1 \to A_2 \) which is a linear map between the vector spaces \( A_1 \) and \( A_2 \), and preserves Lie brackets; that is,

\[
f([A, B]) = [f(A), f(B)]
\]

for all \( A, B \in A_1 \).

An isomorphism of Lie groups is a bijective function \( f \) such that both \( f \) and \( f^{-1} \) are maps of Lie groups, and an isomorphism of Lie algebras is a bijective function \( f \) such that both \( f \) and \( f^{-1} \) are maps of Lie algebras.

The Lie bracket operation on \( g \) can be defined in terms of the so-called adjoint representation.

Given a Lie group \( G \), for every \( a \in G \) we define left translation as the map \( L_a: G \to G \) such that \( L_a(b) = ab \) for all \( b \in G \), and right translation as the map \( R_a: G \to G \) such that \( R_a(b) = ba \) for all \( b \in G \). Because multiplication and the inverse maps are smooth, the maps \( L_a \) and \( R_a \) are diffeomorphisms, and their derivatives play an important role. The inner automorphisms \( R_a^{-1} \circ L_a = L_a \circ R_a^{-1} \), also denoted \( \text{Ad}_a \), play an important role. Note that \( \text{Ad}_a: G \to G \) is defined as

\[
\text{Ad}_a(b) = R_a^{-1}L_a(b) = aba^{-1}.
\]

The derivative

\[
d(\text{Ad}_a)_1: T_1 G \to T_1 G
\]

of \( \text{Ad}_a: G \to G \) at 1 is an isomorphism of Lie algebras, and since \( T_1 G = g \), we get a map denoted

\[
\text{Ad}_a: g \to g,
\]

where \( d(\text{Ad}_a)_1 = \text{Ad}_a \).

The map \( a \mapsto \text{Ad}_a \) is a map of Lie groups

\[
\text{Ad}: G \to \text{GL}(g),
\]

called the adjoint representation of \( G \) (where \( \text{GL}(g) \) denotes the Lie group of all bijective linear maps on \( g \)). In the case of a Lie linear group, we have verified in Section 4.2 that

\[
\text{Ad}(a)(X) = \text{Ad}_a(X) = aXa^{-1}
\]

for all \( a \in G \) and all \( X \in g \).

The derivative

\[
d\text{Ad}_1: g \to gl(g)
\]
of $\text{Ad}: G \to \text{GL}(g)$ at $1$ is map of Lie algebras, denoted by
\[\text{ad}: g \to \mathfrak{gl}(g),\]
called the \textit{adjoint representation} of $g$.

Recall that Theorem 4.8 immediately implies that the Lie algebra $\mathfrak{gl}(g)$ of $\text{GL}(g)$ is the vector space $\text{End}(g, g)$ of all endomorphisms of $g$; that is, the vector space of all linear maps on $g$. In the case of a linear Lie group, we verified in Section 4.2 that
\[\text{ad}(A)(B) = [A, B] = AB - BA,\]
for all $A, B \in g$.

In the case of an abstract Lie group $G$, since $\text{ad}$ is defined, we would like to define the Lie bracket of $g$ in terms of $\text{ad}$. This is the key to the definition of the Lie bracket in the case of a general Lie group (not just a linear Lie group).

**Definition 15.4.** Given a Lie group $G$, the tangent space $g = T_1G$ at the identity with the Lie bracket defined by
\[[u, v] = \text{ad}(u)(v), \quad \text{for all } u, v \in g,\]
is the \textit{Lie algebra of the Lie group} $G$. The Lie algebra $g$ of a Lie group $G$ is also denoted by $\mathfrak{L}(G)$ (for instance, when the notation $g$ is already used for something else).

Actually, we have to justify why $g$ really is a Lie algebra. For this we have

**Proposition 15.1.** Given a Lie group $G$, the Lie bracket $[u, v] = \text{ad}(u)(v)$ of Definition 15.4 satisfies the axioms of a Lie algebra (given in Definition 15.2). Therefore, $g$ with this bracket is a Lie algebra.

**Proof.** The proof requires Proposition 15.9, but we prefer to defer the proof of this proposition until Section 15.3. Since
\[\text{Ad}: G \to \text{GL}(g)\]
is a Lie group homomorphism, by Proposition 15.9, the map $\text{ad} = d\text{Ad}_1$ is a homomorphism of Lie algebras, $\text{ad}: g \to \mathfrak{gl}(g)$, which means that
\[\text{ad}([u, v]) = [\text{ad}(u), \text{ad}(v)] = \text{ad}(u) \circ \text{ad}(v) - \text{ad}(v) \circ \text{ad}(u), \quad \text{for all } u, v \in g,\]
since the bracket in $\mathfrak{gl}(g) = \text{End}(g, g)$, is just the commutator. Applying the above to $z \in g$ gives
\[\text{ad}([u, v])(z) = [[u, v], z]\]
\[= \text{ad}(u) \circ \text{ad}(v)(z) - \text{ad}(v) \circ \text{ad}(u)(z)\]
\[= \text{ad}(u)[v, z] - \text{ad}(v)[u, z] = [u, [v, z]] - [v, [u, z]],\]
which is equivalent to the Jacobi identity. We still have to prove that \([u, u] = 0\), or equivalently, that \([v, u] = -[u, v]\). For this, following Duistermaat and Kolk [64] (Chapter 1, Section 1), consider the map

\[ F: G \times G \longrightarrow G: (a, b) \mapsto aba^{-1}b^{-1}. \]

We claim that the derivative of \(F\) at \((1, 1)\) is the zero map. This follows using the product rule and chain rule from two facts:

1. The derivative of multiplication in a Lie group \(\mu: G \times G \rightarrow G\) is given by

\[ d\mu_{a,b}(u,v) = (dR_{b})_{a}(u) + (dL_{a})_{b}(v), \]

for all \(u \in T_{a}G\) and all \(v \in T_{b}G\). At \((1, 1)\), the above yields

\[ d\mu_{1,1}(u,v) = u + v. \]

2. The derivative of the inverse map \(\iota: G \rightarrow G\) is given by

\[ d\iota_{a}(u) = -(dR_{a^{-1}})_{1} \circ (dL_{a^{-1}})_{a}(u) = -(dL_{a^{-1}})_{1} \circ (dR_{a^{-1}})_{a}(u) \]

for all \(u \in T_{a}G\). At 1, we get

\[ d\iota_{1}(u) = -u. \]

In particular write \(F = F_1 F_2\), where \(F_1: G \times G \rightarrow G\) is \(F_1(a, b) = ab\) and \(F_2: G \times G \rightarrow G\) is \(F_2(a, b) = a^{-1}b^{-1} = (ba)^{-1}\). The product rule and chain rule implies that for all \(u, v \in g \times g\),

\[ dF_{1,1}(u,v) = (dF_{1})_{1,1}(u,v) F_{2}(1,1) + F_{1}(1,1) (dF_{2})_{1,1}(u,v) = (u + v) \cdot 1 + 1 \cdot (-u - v) = 0. \]

Since \(dF_{1,1} = 0\), then \((1, 1)\) is a critical point of \(F\), and we can adapt the standard reasoning, for example found in Milnor [125], page 4-5, to prove that the Hessian \(\text{Hess}(F)\) of \(F\) is well-defined at \((1, 1)\), and is a symmetric bilinear map

\[ \text{Hess}(F)_{(1,1)}: (g \times g) \times (g \times g) \longrightarrow g. \]

Furthermore, for any \((X_1, Y_1)\) and \((X_2, Y_2)\) \(\in g \times g\), the value \(\text{Hess}(F)_{(1,1)}((X_1, Y_1), (X_2, Y_2))\) of the Hessian can be computed by two successive derivatives, either as

\[ (\vec{X}_1, \vec{Y}_1)((\vec{X}_2, \vec{Y}_2)F)_{1,1}, \]

or as

\[ (\vec{X}_2, \vec{Y}_2)((\vec{X}_1, \vec{Y}_1)F)_{1,1}, \]

where \(\vec{X}_i\) and \(\vec{Y}_i\) are smooth vector fields with value \(X_i\) and \(Y_j\) at 1, which exist by Proposition 9.2, and with

\[ (\vec{X}, \vec{Y})\Theta_{(1,1)} = d\Theta_{(1,1)}(X,Y) \]
for any smooth function $\Theta : G \times G \to G$. Note that $(\tilde{X}, \tilde{Y})\Theta_{(1,1)}$ is a slight generalization of the notion the directional derivative of $\Theta$ at $(1,1)$ in the direction $(X, Y)$ ∈ $g \times g$.

We can then compute the differential w.r.t. $b$ at $b = 1$ and evaluate at $(v, v) \in g \times g$, getting $(\text{Ad}_a - \text{id})(v)$. Then, the derivative w.r.t. $a$ at $a = 1$ evaluated at $(u, u) \in g \times g$ is $[u, v]$. On the other hand if we differentiate first w.r.t. $a$ and then w.r.t. $b$, we first get $(\text{id} - \text{Ad}_b)(u)$ and then $-[v, u]$. Since the Hessian is bilinear symmetric, we get $[u, v] = -[v, u]$. 

**Remark:** After proving that $g$ is isomorphic to the vector space of left-invariant vector fields on $G$, we get another proof of Proposition 15.1.

### 15.2 Left and Right Invariant Vector Fields, the Exponential Map

The purpose of this section is to define the exponential map for an arbitrary Lie group in a way that is consistent with our previous definition of the exponential defined for a linear Lie group, namely

$$e^X = I_n + \sum_{p \geq 1} \frac{X^p}{p!} = \sum_{p \geq 0} \frac{X^p}{p!},$$

where $X \in M_n(\mathbb{R})$ or $X \in M_n(\mathbb{C})$. We obtain the desired generalization by recalling Proposition 2.25 which states that for a linear Lie group, the maximal integral curve through initial point $p \in G$ with initial velocity $X$ is given by $\gamma_p(t) = e^{tX}p$; see Sections 2.3 and 8.3. Thus the exponential may be defined in terms of maximal integral curves. Since the notion of maximal integral curves relies on vector fields, we begin our construction of the exponential map for an abstract Lie group $G$ by defining left and right invariant vector fields.

**Definition 15.5.** If $G$ is a Lie group, a vector field $X$ on $G$ is **left-invariant** (resp. **right-invariant**) iff

$$d(L_a)_b(X(b)) = X(L_a(b)) = X(ab), \quad \text{for all } a, b \in G.$$ (resp.

$$d(R_a)_b(X(b)) = X(R_a(b)) = X(ba), \quad \text{for all } a, b \in G.$$)

Equivalently, a vector field $X$ is left-invariant iff the following diagram commutes (and similarly for a right-invariant vector field):

$$\begin{array}{c}
TG \xrightarrow{d(L_a)} TG \\
x \downarrow & \downarrow x \\
G \xrightarrow{L_a} G
\end{array}$$
If \( X \) is a left-invariant vector field, setting \( b = 1 \), we see that
\[
d(L_a)_1(X(1)) = X(L_a(1)) = X(a),
\]
which shows that \( X \) is determined by its value \( X(1) \in \mathfrak{g} \) at the identity (and similarly for right-invariant vector fields).

Conversely, given any \( v \in \mathfrak{g} \), since \((dL_a)_1 : \mathfrak{g} \to T_aG\) is a linear isomorphism between \( \mathfrak{g} \) and \( T_aG \) for every \( a \in G \), we can define the vector field \( v^L \) by
\[
v^L(a) = d(L_a)_1(v), \quad \text{for all } a \in G.
\]

We claim that \( v^L \) is left-invariant. This follows by an easy application of the chain rule:
\[
v^L(ab) = d(L_{ab})_1(v) = d(L_a \circ L_b)_1(v) = d(L_a)_b(d(L_b)_1(v)) = d(L_a)_b(v^L(b)).
\]
Furthermore, \( v^L(1) = v \).

Therefore, we showed that the map \( X \mapsto X(1) \) establishes an isomorphism between the space of left-invariant vector fields on \( G \) and \( \mathfrak{g} \). In fact, the map \( G \times \mathfrak{g} \to TG \) given by \((a, v) \mapsto v^L(a)\) is an isomorphism between \( G \times \mathfrak{g} \) and the tangent bundle, \( TG \).

We denote the vector space of left-invariant vector fields on \( G \) by \( \mathfrak{g}^L \). Because Proposition 15.9 implies that the derivative of any Lie group homomorphism is a Lie algebra homomorphism, \((dL_a)_b\) is a Lie algebra homomorphism, so if \( X \) and \( Y \) are left-invariant vector fields, then the vector field \([X, Y]\) is also left-invariant. In particular
\[
(dL_a)_b[X(b), Y(b)] = [(dL_a)_bX(b), (dL_a)_bY(b)], \quad (dL_a)_b \text{ is a Lie algebra homomorphism}
= [X(L_a(b)), Y(L_a(b))], \quad X \text{ and } Y \text{ are left-invariant vector fields}
= [X(ab), Y(ab)].
\]
It follows that \( \mathfrak{g}^L \) is a Lie algebra.

**Remark:** Given any \( v \in \mathfrak{g} \), since \((dR_a)_1 : \mathfrak{g} \to T_aG\) is a linear isomorphism between \( \mathfrak{g} \) and \( T_aG \) for every \( a \in G \), we can also define the vector field \( v^R \) by
\[
v^R(a) = d(R_a)_1(v), \quad \text{for all } a \in G.
\]
It is easily shown that \( v^R \) is right-invariant and we also have an isomorphism \( G \times \mathfrak{g} \to TG \) given by \((a, v) \mapsto v^R(a)\).

We denote the vector space of right-invariant vector fields on \( G \) by \( \mathfrak{g}^R \). Since \((dR_a)_b\) is a Lie algebra homomorphism, if \( X \) and \( Y \) are right-invariant vector fields, then the vector
15.2. LEFT AND RIGHT INVARIANT VECTOR FIELDS, EXPONENTIAL MAP

A similar argument applies to right-invariant vector fields.

**Proposition 15.3.** Given a Lie group \( G \), for every \( v \in \mathfrak{g} \), there is a unique smooth homomorphism \( h_v: (\mathbb{R}, +) \to G \) such that \( h_v(0) = v \). Furthermore, \( h_v(t) = \gamma_g(t) \) is the maximal integral curve of both \( v^L \) and \( v^R \) with initial condition 1, and the flows of \( v^L \) and \( v^R \) are defined for all \( t \in \mathbb{R} \).

**Proof.** Let \( \Phi_v^g(g) = \gamma_g(t) \) denote the flow of \( v^L \). As far as defined, we know that

\[
\Phi_{s+t,1}^v = \Phi^v_s(s, \Phi^v_t(1)) = \Phi^v_{s+t}(1),
\]

by Proposition 8.8,

\[
\Phi^v_s(1) = \Phi^v_s(1),
\]

by Proposition 15.2.

Now, if \( \Phi^v_t(1) = \gamma_1(t) \) is defined on \((-\epsilon, \epsilon)\), setting \( s = t \), we see that \( \Phi^v_{s+t}(1) \) is actually defined on \((-2^n\epsilon, 2^n\epsilon)\), for all \( n \geq 0 \), and so \( \Phi^v_s(1) \) is defined on \( \mathbb{R} \), and the map \( t \mapsto \Phi^v(t) \) is a homomorphism \( h_v: (\mathbb{R}, +) \to G \), with \( h_v(0) = v \). Since \( \Phi^v_t(g) = g\Phi^v_t(1) \), the flow \( \Phi^v_t(g) \) is defined for all \((t, g) \in \mathbb{R} \times G \). A similar proof applies to \( v^R \). To show that \( h_v \) is smooth, consider the map

\[
\mathbb{R} \times G \times \mathfrak{g} \longrightarrow G \times \mathfrak{g}, \quad (t, g, v) \mapsto (g\Phi^v_t(1), v).
\]
It can be shown that the above is the flow of the vector field
\[(g, v) \mapsto (v(g), 0),\]
and thus it is smooth. Consequently, the restriction of this smooth map to \(\mathbb{R} \times \{1\} \times \{v\}\), which is just \(t \mapsto \Phi^v_t(1) = h_v(t)\), is also smooth.

Assume \(h: (\mathbb{R}, +) \to G\) is a smooth homomorphism with \(\dot{h}(0) = v\). From
\[h(t + s) = h(t)h(s) = h(s)h(t),\]
we have
\[h(t + s) = L_{h(t)}h(s),\quad h(t + s) = R_{h(t)}h(s).\]
If we differentiate these equations with respect to \(s\) at \(s = 0\), we get via the chain rule
\[
\frac{dh}{ds}(t) = d(L_{h(t)})_1(v) = v^L(h(t))
\]
and
\[
\frac{dh}{ds}(t) = d(R_{h(t)})_1(v) = v^R(h(t)).
\]
Therefore, \(h(t)\) is an integral curve for \(v^L\) and \(v^R\) with initial condition \(h(0) = 1\), and \(h(t) = \Phi^v_t(1) = \gamma_1(t)\).

Since \(h_v: (\mathbb{R}, +) \to G\) is a homomorphism, the integral curve \(h_v\) if often referred to as a one-parameter group. Proposition 15.3 yields the definition of the exponential map in terms of maximal integral curves.

**Definition 15.6.** Given a Lie group \(G\), the exponential map \(\exp: \mathfrak{g} \to G\) is given by
\[\exp(v) = h_v(1) = \Phi^v_1(1) = \gamma_1(1), \quad \text{for all } v \in \mathfrak{g}.\]

We can see that \(\exp\) is smooth as follows. As in the proof of Proposition 15.3, we have the smooth map
\[\mathbb{R} \times G \times \mathfrak{g} \longrightarrow G \times \mathfrak{g}, \quad \text{where} \quad (t, g, v) \mapsto (g\Phi^v_t(1), v),\]
which is the flow of the vector field
\[(g, v) \mapsto (v(g), 0).\]
Consequently, the restriction of this smooth map to \(\{1\} \times \{1\} \times \mathfrak{g}\), which is just \(v \mapsto \Phi^v_1(1) = \exp(v)\), is also smooth.

Observe that for any fixed \(t \in \mathbb{R}\), the map
\[s \mapsto h_v(st) = \gamma_1(st)\]
is a smooth homomorphism \( h \) such that \( \dot{h}(0) = tv \). By uniqueness of the maximal integral curves, we have
\[
\Phi_{st}^v(1) = h_v(st) = h_{tv}(s) = \Phi_s^{tv}(1).
\]
Setting \( s = 1 \), we find that
\[
\gamma_1(t) = h_v(t) = \exp(tv), \quad \text{for all } v \in \mathfrak{g} \text{ and all } t \in \mathbb{R}.
\]
If \( G \) is a linear Lie group, the preceding equation is equivalent to Proposition 2.25.

Differentiating this equation with respect to \( t \) at \( t = 0 \), we get
\[
v = d\exp_0(v),
\]
i.e., \( d\exp_0 = \text{id}_\mathfrak{g} \). By the inverse function theorem, \( \exp \) is a local diffeomorphism at \( 0 \). This means that there is some open subset \( U \subseteq \mathfrak{g} \) containing \( 0 \), such that the restriction of \( \exp \) to \( U \) is a diffeomorphism onto \( \exp(U) \subseteq G \), with \( 1 \in \exp(U) \). In fact, by left-translation, the map \( v \mapsto g \exp(v) \) is a local diffeomorphism between some open subset \( U \subseteq \mathfrak{g} \) containing \( 0 \) and the open subset \( \exp(U) \) containing \( g \).

Given any Lie group \( G \), we have a notion of exponential map \( \exp: \mathfrak{g} \to G \) given by the maximal integral curves of left-invariant vector fields on \( G \) (see Proposition 15.3 and Definition 15.6). This exponential does not require any connection or any metric in order to be defined; let us call it the \textit{group exponential}. If \( G \) is endowed with a connection or a Riemannian metric (the Levi-Civita connection if \( G \) has a Riemannian metric), then we also have the notion of exponential induced by geodesics (see Definition 12.4); let us call this exponential the \textit{geodesic exponential}. To avoid ambiguities when both kinds of exponentials arise, we propose to denote the group exponential by \( \exp_{\text{gr}} \) and the geodesic exponential by \( \exp \), as before. Even if the geodesic exponential is defined on the whole of \( \mathfrak{g} \) (which may not be the case), these two notions of exponential differ in general.

The group exponential map is natural in the following sense:

**Proposition 15.4.** Given any two Lie groups \( G \) and \( H \), for every Lie group homomorphism \( f: G \to H \), the following diagram commutes:

\[
\begin{array}{ccc}
G & \xrightarrow{f} & H \\
\exp \downarrow & & \downarrow \exp \\
\mathfrak{g} & \xrightarrow{df_1} & \mathfrak{h}
\end{array}
\]

**Proof.** Observe that for every \( v \in \mathfrak{g} \), the map \( h: t \mapsto f(\exp(tv)) \) is a homomorphism from \((\mathbb{R}, +)\) to \( G \) such that \( \dot{h}(0) = df_1(v) \). On the other hand, Proposition 15.3 shows that the map \( t \mapsto \exp(tdf_1(v)) \) is the unique maximal integral curve whose tangent at \( 0 \) is \( df_1(v) \), so \( f(\exp(v)) = \exp(df_1(v)) \). \( \square \)
Proposition 15.4 is the generalization of Proposition 4.13.

A useful corollary of Proposition 15.4 is:

**Proposition 15.5.** Let $G$ be a connected Lie group and $H$ be any Lie group. For any two homomorphisms $\phi_1: G \to H$ and $\phi_2: G \to H$, if $d(\phi_1)_1 = d(\phi_2)_1$, then $\phi_1 = \phi_2$.

*Proof.* We know that the exponential map is a diffeomorphism on some small open subset $U$ containing 0. Now, by Proposition 15.4, for all $a \in \exp G(U)$, we have

$$
\phi_i(a) = \exp_H(d(\phi_i)_1(\exp_G^{-1}(a))), \quad i = 1, 2.
$$

Since $d(\phi_1)_1 = d(\phi_2)_1$, we conclude that $\phi_1 = \phi_2$ on $\exp G(U)$. However, as $G$ is connected, Proposition 5.9 implies that $G$ is generated by $\exp G(U)$ (we can easily find a symmetric neighborhood of 1 in $\exp G(U)$). Therefore, $\phi_1 = \phi_2$ on $G$. \qed

The above proposition shows that if $G$ is connected, then a homomorphism of Lie groups $\phi: G \to H$ is uniquely determined by the Lie algebra homomorphism $d\phi_1: \mathfrak{g} \to \mathfrak{h}$.

We obtain another useful corollary of Proposition 15.4 when we apply it to the adjoint representation of $G$

$$
\text{Ad}: G \to \text{GL}(\mathfrak{g}),
$$

and to the conjugation map

$$
\text{Ad}_a: G \to G,
$$

where $\text{Ad}_a(b) = aba^{-1}$. In the first case, $d\text{Ad}_1 = \text{ad}$, with $\text{ad}: \mathfrak{g} \to \text{gl}(\mathfrak{g})$, and in the second case, $d(\text{Ad}_a)_1 = \text{Ad}_a$.

**Proposition 15.6.** Given any Lie group $G$, the following properties hold:

1. $$
\text{Ad}(\exp(u)) = e^{\text{ad}(u)}, \quad \text{for all } u \in \mathfrak{g},
$$

where $\exp: \mathfrak{g} \to G$ is the exponential of the Lie group $G$, and $f \mapsto e^f$ is the exponential map given by

$$
e^f = \sum_{k=0}^{\infty} \frac{f^k}{k!},
$$

for any linear map (matrix) $f \in \text{gl}(\mathfrak{g})$. Equivalently, the following diagram commutes:

\[
\begin{array}{ccc}
G & \xrightarrow{\text{Ad}} & \text{GL}(\mathfrak{g}) \\
\exp & & f \mapsto e^f \\
\mathfrak{g} & \xrightarrow{\text{ad}} & \text{gl}(\mathfrak{g}).
\end{array}
\]
15.2. LEFT AND RIGHT INVARIANT VECTOR FIELDS, EXPONENTIAL MAP

\[(2)\]

\[
\exp(t \text{Ad}_g(u)) = g \exp(tu) g^{-1},
\]

for all \(u \in \mathfrak{g}\), all \(g \in G\) and all \(t \in \mathbb{R}\). Equivalently, the following diagram commutes:

\[
\begin{array}{ccc}
G & \xrightarrow{\text{Ad}_g} & G \\
\exp & \downarrow & \exp \\
\mathfrak{g} & \xrightarrow{\text{Ad}_g} & \mathfrak{g}
\end{array}
\]

Since the Lie algebra \(\mathfrak{g} = T_1G\) is isomorphic to the vector space of left-invariant vector fields on \(G\) and since the Lie bracket of vector fields makes sense (see Definition 8.3), it is natural to ask if there is any relationship between \([u, v]\), where \([u, v] = \text{ad}(u)(v)\), and the Lie bracket \([u^L, v^L]\) of the left-invariant vector fields associated with \(u, v \in \mathfrak{g}\). The answer is: Yes, they coincide (via the correspondence \(u \mapsto u^L\)). This fact is recorded in the proposition below whose proof involves some rather acrobatic uses of the chain rule found in Warner [175] (Chapter 3), Bröcker and tom Dieck [31] (Chapter 1, Section 2), or Marsden and Ratiu [121] (Chapter 9).

**Proposition 15.7.** Given a Lie group \(G\), we have

\[
[u^L, v^L](1) = \text{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g},
\]

where \([u^L, v^L](1)\) is the element of the vector field \([u^L, v^L]\) at the identity.

Proposition 15.7 shows that the Lie algebras \(\mathfrak{g}\) and \(\mathfrak{g}^L\) are isomorphic (where \(\mathfrak{g}^L\) is the Lie algebra of left-invariant vector fields on \(G\)). In view of this isomorphism, if \(X\) and \(Y\) are any two left-invariant vector fields on \(G\), we define \(\text{ad}(X)(Y)\) by

\[
\text{ad}(X)(Y) = [X, Y],
\]

where the Lie bracket on the right-hand side is the Lie bracket on vector fields.

It is shown in Marsden and Ratiu [121] (Chapter 9) that if \(\iota: G \to G\) is the inversion map \(\iota(g) = g^{-1}\), then for any \(u \in \mathfrak{g}\), the vector fields \(u^L\) and \(u^R\) are related by the equation

\[
\iota_*(u^L) = -u^R,
\]

where \(\iota_*(u^L)\) is the push-forward of \(u^L\) (that is,

\[
\iota_*(u^L) = d\iota_{g^{-1}}(u^L(g^{-1}))
\]

for all \(g \in G\).) This implies that

\[
[u, v]^R = -[u^R, v^R],
\]
and so

\[ [u^R, v^R](1) = -\text{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g}. \]

It follows that the Lie algebras \( \mathfrak{g} \) and \( \mathfrak{g}^R \) are anti-isomorphic (where \( \mathfrak{g}^R \) is the Lie algebra of right-invariant vector fields on \( G \)).

We can apply Proposition 5.10 and use the exponential map to prove a useful result about Lie groups. If \( G \) is a Lie group, let \( G_0 \) be the connected component of the identity. We know \( G_0 \) is a topological normal subgroup of \( G \) and it is a submanifold in an obvious way, so it is a Lie group.

**Proposition 15.8.** If \( G \) is a Lie group and \( G_0 \) is the connected component of 1, then \( G_0 \) is generated by \( \exp(\mathfrak{g}) \). Moreover, \( G_0 \) is countable at infinity.

**Proof.** We can find a symmetric open \( U \) in \( \mathfrak{g} \) containing 0, on which \( \exp \) is a diffeomorphism. Then, apply Proposition 5.10 to \( V = \exp(U) \). That \( G_0 \) is countable at infinity follows from Proposition 5.11. \qed

### 15.3 Homomorphisms of Lie Groups and Lie Algebras, Lie Subgroups

If \( G \) and \( H \) are two Lie groups and \( \phi: G \to H \) is a homomorphism of Lie groups, then \( d\phi_1: \mathfrak{g} \to \mathfrak{h} \) is a linear map between the Lie algebras \( \mathfrak{g} \) and \( \mathfrak{h} \) of \( G \) and \( H \). In fact, it is a Lie algebra homomorphism, as shown below. This proposition is the generalization of Proposition 4.14.

**Proposition 15.9.** If \( G \) and \( H \) are two Lie groups and \( \phi: G \to H \) is a homomorphism of Lie groups, then

\[ d\phi_1 \circ \text{Ad}_g = \text{Ad}_{\phi(g)} \circ d\phi_1, \quad \text{for all } g \in G; \]

that is, the following diagram commutes

\[
\begin{array}{ccc}
\mathfrak{g} & \xrightarrow{d\phi_1} & \mathfrak{h} \\
\text{Ad}_g \downarrow & & \downarrow \text{Ad}_{\phi(g)} \\
\mathfrak{g} & \xrightarrow{d\phi_1} & \mathfrak{h}
\end{array}
\]

and \( d\phi_1: \mathfrak{g} \to \mathfrak{h} \) is a Lie algebra homomorphism.

**Proof.** Recall that

\[ \text{Ad}_a(b) = R_{a^{-1}}L_a(b) = aba^{-1}, \quad \text{for all } a, b \in G \]

and that the derivative

\[ d(\text{Ad}_a)_1: \mathfrak{g} \to \mathfrak{g} \]
of $\text{Ad}_a$ at $1$ is an isomorphism of Lie algebras, denoted by $\text{Ad}_a: g \to g$. The map $a \mapsto \text{Ad}_a$ is a map of Lie groups

$$\text{Ad}: G \to \text{GL}(g),$$

(where $\text{GL}(g)$ denotes the Lie group of all bijective linear maps on $g$) and the derivative

$$d\text{Ad}_1: g \to \text{gl}(g)$$

of $\text{Ad}$ at $1$ is map of Lie algebras, denoted by

$$\text{ad}: g \to \text{gl}(g),$$

called the adjoint representation of $g$ (where $\text{gl}(g)$ denotes the Lie algebra of all linear maps on $g$). Then the Lie bracket is defined by

$$[u, v] = \text{ad}(u)(v), \quad \text{for all } u, v \in g.$$

Now, as $\phi$ is a homomorphism, we have

$$\phi(\text{Ad}_a(b)) = \phi(aba^{-1}) = \phi(a)\phi(b)\phi(a)^{-1} = R_{\phi(a)^{-1}}L_{\phi(a)}(\phi(b)) = \text{Ad}_{\phi(a)}(\phi(b)),$$

and by differentiating w.r.t. $b$ at $b = 1$ in the direction, $v \in g$, we get

$$d\phi_1(\text{Ad}_a(v)) = \text{Ad}_{\phi(a)}(d\phi_1(v)),$$

proving the first part of the proposition. Differentiating again with respect to $a$ at $a = 1$ in the direction, $u \in g$, (and using the chain rule), we get

$$d\phi_1(\text{ad}(u)(v)) = \text{ad}(d\phi_1(u))(d\phi_1(v)),$$

i.e.,

$$d\phi_1[u, v] = [d\phi_1(u), d\phi_1(v)],$$

which proves that $d\phi_1$ is indeed a Lie algebra homomorphism.

\[ \square \]

**Remark:** If we identify the Lie algebra $g$ of $G$ with the space of left-invariant vector fields on $G$, then the map $d\phi_1: g \to h$ is viewed as the map such that, for every left-invariant vector field $X$ on $G$, the vector field $d\phi_1(X)$ is the unique left-invariant vector field on $H$ such that

$$d\phi_1(X)(1) = d\phi_1(X(1)),$$

i.e., $d\phi_1(X) = d\phi_1(X(1))^L$. Then we can give another proof of the fact that $d\phi_1$ is a Lie algebra homomorphism using the notion of $\phi$-related vector fields.

**Proposition 15.10.** If $G$ and $H$ are two Lie groups and $\phi: G \to H$ is a homomorphism of Lie groups, if we identify $g$ (resp. $h$) with the space of left-invariant vector fields on $G$ (resp. left-invariant vector fields on $H$), then,
(a) $X$ and $d\phi_1(X)$ are $\phi$-related, for every left-invariant vector field $X$ on $G$;

(b) $d\phi_1 : \mathfrak{g} \to \mathfrak{h}$ is a Lie algebra homomorphism.

Proof. The proof uses Proposition 8.5. For details see Warner [175].

We now consider Lie subgroups. As a preliminary result, note that if $\phi : G \to H$ is an injective Lie group homomorphism, we prove that $d\phi_g : T_gG \to T_{\phi(g)}H$ is injective for all $g \in G$. As $\mathfrak{g} = T_1G$ and $T_gG$ are isomorphic for all $g \in G$ (and similarly for $\mathfrak{h} = T_1H$ and $T_hH$ for all $h \in H$), it is sufficient to check that $d\phi_1 : \mathfrak{g} \to \mathfrak{h}$ is injective. However, by Proposition 15.4, the diagram

$$
\begin{array}{ccc}
G & \xrightarrow{\phi} & H \\
\exp & \downarrow & \exp \\
\mathfrak{g} & \xrightarrow{d\phi_1} & \mathfrak{h}
\end{array}
$$

commutes, and since the exponential map is a local diffeomorphism at 0, as $\phi$ is injective, then $d\phi_1$ is injective, too. Therefore, if $\phi : G \to H$ is injective, it is automatically an immersion.

**Definition 15.7.** Let $G$ be a Lie group. A set $H$ is an immersed (Lie) subgroup of $G$ iff

(a) $H$ is a Lie group;

(b) There is an injective Lie group homomorphism $\phi : H \to G$ (and thus, $\phi$ is an immersion, as noted above).

We say that $H$ is a Lie subgroup (or closed Lie subgroup) of $G$ iff $H$ is a Lie group which is a subgroup of $G$, and also a submanifold of $G$.

Observe that an immersed Lie subgroup $H$ is an immersed submanifold, since $\phi$ is an injective immersion. However, $\phi(H)$ may not have the subspace topology inherited from $G$ and $\phi(H)$ may not be closed.

An example of this situation is provided by the 2-torus $T^2 \cong \text{SO}(2) \times \text{SO}(2)$, which can be identified with the group of $2 \times 2$ complex diagonal matrices of the form

$$
\begin{pmatrix}
e^{i\theta_1} & 0 \\
0 & e^{i\theta_2}
\end{pmatrix}
$$

where $\theta_1, \theta_2 \in \mathbb{R}$. For any $c \in \mathbb{R}$, let $S_c$ be the subgroup of $T^2$ consisting of all matrices of the form

$$
\begin{pmatrix}
e^{it} & 0 \\
0 & e^{ict}
\end{pmatrix}, \quad t \in \mathbb{R}.
$$

It is easily checked that $S_c$ is an immersed Lie subgroup of $T^2$ iff $c$ is irrational. However, when $c$ is irrational, one can show that $S_c$ is dense in $T^2$ but not closed.
As we will see below, a Lie subgroup is always closed. We borrowed the terminology “immersed subgroup” from Fulton and Harris [70] (Chapter 7), but we warn the reader that most books call such subgroups “Lie subgroups” and refer to the second kind of subgroups (that are submanifolds) as “closed subgroups.”

**Theorem 15.11.** Let $G$ be a Lie group and let $(H, \phi)$ be an immersed Lie subgroup of $G$. Then, $\phi$ is an embedding iff $\phi(H)$ is closed in $G$. As as consequence, any Lie subgroup of $G$ is closed.

**Proof.** The proof can be found in Warner [175] (Chapter 1, Theorem 3.21) and uses a little more machinery than we have introduced. However, we prove that a Lie subgroup $H$ of $G$ is closed. The key to the argument is this: Since $H$ is a submanifold of $G$, there is chart $(U, \varphi)$ of $G$, with $1 \in U$, so that

$$\varphi(U \cap H) = \varphi(U) \cap (R^m \times \{0_{n-m}\}).$$

By Proposition 5.4, we can find some open subset $V \subseteq U$ with $1 \in V$, so that $V = V^{-1}$ and $V \subseteq U$. Observe that

$$\varphi(V \cap H) = \varphi(V) \cap (R^m \times \{0_{n-m}\})$$

and since $V$ is closed and $\varphi$ is a homeomorphism, it follows that $V \cap H$ is closed. Thus, $V \cap H = V \cap H$ (as $V \cap H = V \cap H$). Now, pick any $y \in H$. As $1 \in V^{-1}$, the open set $yV^{-1}$ contains $y$ and since $y \in H$, we must have $yV^{-1} \cap H \neq \emptyset$. Let $x \in yV^{-1} \cap H$, then $x \in H$ and $y \in xV$. Then, $y \in xV \cap H$, which implies $x^{-1}y \in V \cap H \subseteq \overline{V} \cap H = V \cap H$. Therefore, $x^{-1}y \in H$ and since $x \in H$, we get $y \in H$ and $H$ is closed.

We also have the following important and useful theorem: If $G$ is a Lie group, say that a subset $H \subseteq G$ is an abstract subgroup iff it is just a subgroup of the underlying group of $G$ (i.e., we forget the topology and the manifold structure).

**Theorem 15.12.** Let $G$ be a Lie group. An abstract subgroup $H$ of $G$ is a submanifold (i.e., a Lie subgroup) of $G$ iff $H$ is closed (i.e, $H$ with the induced topology is closed in $G$).

**Proof.** We proved the easy direction of this theorem above. Conversely, we need to prove that if the subgroup $H$ with the induced topology is closed in $G$, then it is a manifold. This can be done using the exponential map, but it is harder. For details, see Bröcker and tom Dieck [31] (Chapter 1, Section 3) or Warner [175], Chapter 3.

### 15.4 The Correspondence Lie Groups–Lie Algebras

Historically, Lie was the first to understand that a lot of the structure of a Lie group is captured by its Lie algebra, a simpler object (since it is a vector space). In this short section, we state without proof some of the “Lie theorems,” although not in their original form.
**Definition 15.8.** If \( \mathfrak{g} \) is a Lie algebra, a subalgebra \( \mathfrak{h} \) of \( \mathfrak{g} \) is a (linear) subspace of \( \mathfrak{g} \) such that \([u,v] \in \mathfrak{h}\), for all \( u,v \in \mathfrak{h} \). If \( \mathfrak{h} \) is a (linear) subspace of \( \mathfrak{g} \) such that \([u,v] \in \mathfrak{h}\) for all \( u \in \mathfrak{h} \) and all \( v \in \mathfrak{g} \), we say that \( \mathfrak{h} \) is an ideal in \( \mathfrak{g} \).

For a proof of the theorem below see Warner [175] (Chapter 3) or Duistermaat and Kolk [64] (Chapter 1, Section 10).

**Theorem 15.13.** Let \( G \) be a Lie group with Lie algebra \( \mathfrak{g} \), and let \((H, \phi)\) be an immersed Lie subgroup of \( G \) with Lie algebra \( \mathfrak{h} \); then \( d\phi_1 \mathfrak{h} \) is a Lie subalgebra of \( \mathfrak{g} \). Conversely, for each subalgebra \( \tilde{\mathfrak{h}} \) of \( \mathfrak{g} \), there is a unique connected immersed subgroup \((H, \phi)\) of \( G \) so that \( d\phi_1 \mathfrak{h} = \tilde{\mathfrak{h}} \). In fact, as a group, \( \phi(H) \) is the subgroup of \( G \) generated by \( \exp(\tilde{\mathfrak{h}}) \). Furthermore, normal subgroups correspond to ideals.

Theorem 15.13 shows that there is a one-to-one correspondence between connected immersed subgroups of a Lie group and subalgebras of its Lie algebra.

**Theorem 15.14.** Let \( G \) and \( H \) be Lie groups with \( G \) connected and simply connected and let \( \mathfrak{g} \) and \( \mathfrak{h} \) be their Lie algebras. For every homomorphism \( \psi : \mathfrak{g} \to \mathfrak{h} \), there is a unique Lie group homomorphism \( \phi : G \to H \) so that \( d\phi_1 = \psi \).

Again a proof of the theorem above is given in Warner [175] (Chapter 3) or Duistermaat and Kolk [64] (Chapter 1, Section 10).

**Corollary 15.15.** If \( G \) and \( H \) are connected and simply connected Lie groups, then \( G \) and \( H \) are isomorphic iff \( \mathfrak{g} \) and \( \mathfrak{h} \) are isomorphic.

It can also be shown that for every finite-dimensional Lie algebra \( \mathfrak{g} \), there is a connected and simply connected Lie group \( G \) such that \( \mathfrak{g} \) is the Lie algebra of \( G \). This is a consequence of deep theorem (whose proof is quite hard) known as Ado’s theorem. For more on this, see Knapp [106], Fulton and Harris [70], or Bourbaki [28].

In summary, following Fulton and Harris, we have the following two principles of the Lie group/Lie algebra correspondence:

**First Principle:** (restatement of Proposition 15.5:) If \( G \) and \( H \) are Lie groups, with \( G \) connected, then a homomorphism of Lie groups \( \phi : G \to H \) is uniquely determined by the Lie algebra homomorphism \( d\phi_1 : \mathfrak{g} \to \mathfrak{h} \).

**Second Principle:** (restatement of Theorem 15.14:) Let \( G \) and \( H \) be Lie groups with \( G \) connected and simply connected and let \( \mathfrak{g} \) and \( \mathfrak{h} \) be their Lie algebras. A linear map \( \psi : \mathfrak{g} \to \mathfrak{h} \) is a Lie algebra map iff there is a unique Lie group homomorphism \( \phi : G \to H \) so that \( d\phi_1 = \psi \).
15.5 Semidirect Products of Lie Algebras and Lie Groups

The purpose of this section is to construct an entire class of Lie algebras and Lie groups by combining two "smaller" pieces in a manner which preserves the algebraic structure. We begin with two Lie algebras and form a new vector space via the direct sum. If \( a \) and \( b \) are two Lie algebras, recall that the direct sum \( a \oplus b \) of \( a \) and \( b \) is \( a \times b \) with the product vector space structure where

\[
(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)
\]

for all \( a_1, a_2 \in a \) and all \( b_1, b_2 \in b \), and

\[
\lambda(a, b) = (\lambda a, \lambda b)
\]

for all \( \lambda \in \mathbb{R}, a \in a, \) and all \( b \in b \). The map \( a \mapsto (a, 0) \) is an isomorphism of \( a \) with the subspace \( \{(a, 0) \mid a \in a\} \) of \( a \oplus b \) and the map \( b \mapsto (0, b) \) is an isomorphism of \( b \) with the subspace \( \{(0, b) \mid b \in b\} \) of \( a \oplus b \). These isomorphisms allow us to identify \( a \) with the subspace \( \{(a, 0) \mid a \in a\} \) and \( b \) with the subspace \( \{(0, b) \mid b \in b\} \). The simplest way to make the direct sum \( a \oplus b \) into a Lie algebra is by defining the Lie bracket \([-,-]\) such that \([a_1, a_2]\) agrees with the Lie bracket on \( a \) for all \( a_1, a_2 \in a \), \([b_1, b_2]\) agrees with the Lie bracket on \( b \) for all \( b_1, b_2 \in b \), and \([a, b] = [b, a] = 0\) for all \( a \in a \) and all \( b \in b \). This Lie algebra is called the Lie algebra direct sum of \( a \) and \( b \). Observe that with this Lie algebra structure, \( a \) and \( b \) are ideals. For example, let \( a = \mathbb{R}^n \) with the zero bracket, and let \( b = \mathfrak{so}(n) \) be the Lie algebra of \( n \times n \) skew symmetric matrices with the commutator bracket. Then \( g = \mathbb{R}^n \oplus \mathfrak{so}(n) \) is a Lie algebra with \([-,-]\) defined as \([u, v] = 0\) for all \( u, v \in \mathbb{R}^n \), \([A, B] = AB - BA\) for all \( A, B \in \mathfrak{so}(n) \), and \([u, A] = 0\) for all \( u \in \mathbb{R}^n \), \( A \in \mathfrak{so}(n) \).

The above construction is sometimes called an "external direct sum" because it does not assume that the constituent Lie algebras \( a \) and \( b \) are subalgebras of some given Lie algebra \( g \). If \( a \) and \( b \) are subalgebras of a given Lie algebra \( g \) such that \( g = a \oplus b \) is a direct sum as a vector space and if both \( a \) and \( b \) are ideals, then for all \( a \in a \) and all \( b \in b \), we have \([a, b] \in a \cap b = (0)\), so \( a \oplus b \) is the Lie algebra direct sum of \( a \) and \( b \). Sometimes, it is called an "internal direct sum."

We now would like to generalize this construction to the situation where the Lie bracket \([a, b]\) of some \( a \in a \) and some \( b \in b \) is given in terms of a map from \( b \) to \( \text{Hom}(a, a) \). For this to work, we need to consider derivations.

**Definition 15.9.** Given a Lie algebra \( g \), a **derivation** is a linear map \( D: g \to g \) satisfying the following condition:

\[
D([X, Y]) = [D(X), Y] + [X, D(Y)], \quad \text{for all } X, Y \in g.
\]

The vector space of all derivations on \( g \) is denoted by \( \text{Der}(g) \).
Given a Lie algebra with \([-,-]\), we may use this bracket structure to define \(\text{ad}: \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})\) as \(\text{ad}(u)(v) = [u,v]\). Then the Jacobi identity can be expressed as
\[
[Z, [X, Y]] = [[Z, X], Y] + [X, [Z, Y]],
\]
which holds iff
\[
(\text{ad} Z)[X, Y] = ([\text{ad} Z]X, Y) + [X, (\text{ad} Z)Y],
\]
and the above equation means that \(\text{ad}(Z)\) is a derivation. In fact, it is easy to check that the Jacobi identity holds iff \(\text{ad} Z\) is a derivation for every \(Z \in \mathfrak{g}\). It turns out that the vector space of derivations \(\text{Der}(\mathfrak{g})\) is a Lie algebra under the commutator bracket.

**Proposition 15.16.** For any Lie algebra \(\mathfrak{g}\), the vector space \(\text{Der}(\mathfrak{g})\) is a Lie algebra under the commutator bracket. Furthermore, the map \(\text{ad}: \mathfrak{g} \to \text{Der}(\mathfrak{g})\) is a Lie algebra homomorphism.

**Proof.** For any \(D, E \in \text{Der}(\mathfrak{g})\) and any \(X, Y \in \mathfrak{g}\), we have
\[
[D, E][X, Y] = (DE - ED)[X, Y] = DE[X, Y] - ED[X, Y]
= [DEX, Y] + [EX, DY] + [DX, EY] + [X, DEY]
- [EDX, Y] - [DX, EY] - [EX, DY] - [X, EDY]
= [DEX, Y] - [EDX, Y] + [X, DEY] - [X, EDY]
= [[D, E]X, Y] + [X, [D, E]Y],
\]
which proves that \([D, E]\) is a derivation. Thus, \(\text{Der}(\mathfrak{g})\) is a Lie algebra. We already know that \(\text{ad} X\) is a derivation for all \(X \in \mathfrak{g}\), so \(\text{ad} \mathfrak{g} \subseteq \text{Der}(\mathfrak{g})\). For all \(X, Y \in \mathfrak{g}\), we need to show that
\[
\text{ad} [X, Y] = (\text{ad} X) \circ (\text{ad} Y) - (\text{ad} Y) \circ (\text{ad} X).
\]
If we apply both sides to any \(Z \in \mathfrak{g}\), we get
\[
(\text{ad} [X, Y])(Z) = (\text{ad} X)((\text{ad} Y)(Z)) - (\text{ad} Y)((\text{ad} X)(Z)),
\]
that is,
\[
[[X, Y], Z] = [X, [Y, Z]] - [Y, [X, Z]],
\]
which is equivalent to
\[
[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0,
\]
which is the Jacobi identity. Therefore, \(\text{ad}\) is a Lie algebra homomorphism. \(\square\)

If \(D \in \text{Der}(\mathfrak{g})\) and \(X \in \mathfrak{g}\), it is easy to show that
\[
[D, \text{ad} X] = \text{ad} (DX),
\]
since for all $Z \in \mathfrak{g}$ we have
\[
[D, \text{ad} X] = D(\text{ad} X(Z)) - \text{ad} X(D(Z))
\]
\[
= D[X, Z] - [X, DZ] = [DX, Z] + [X, DZ] - [X, DZ]
\]
\[
= [DX, Z] = \text{ad} (DX)(Z).
\]

We would like to describe another way of defining a bracket structure on $\mathfrak{a} \oplus \mathfrak{b}$ using $\text{Der}(\mathfrak{a})$. To best understand this construction, let us go back to our previous example where $\mathfrak{a} = \mathbb{R}^n$ with $[\cdot, \cdot]_\mathfrak{a} = 0$, and $\mathfrak{b} = \mathfrak{so}(n)$ with $[A, B]_\mathfrak{b} = AB - BA$ for all $A, B \in \mathfrak{so}(n)$. The underlying vector space is $\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b} = \mathbb{R}^n \oplus \mathfrak{so}(n)$, but this time the bracket on $\mathfrak{g}$ is defined as
\[
[(u, A), (v, B)] = (Av - Bu, [A, B]_\mathfrak{b}), \quad u, v \in \mathbb{R}^n, \quad A, B \in \mathfrak{so}(n).
\]

By using the isomorphism between $\mathfrak{a}$ and $\{(a, 0) \mid a \in \mathfrak{a}\}$ and the isomorphism between $\mathfrak{b}$ and $\{(0, b) \mid b \in \mathfrak{b}\}$, we have
\[
[u, v]_\mathfrak{a} = [(u, 0), (v, 0)] = (0, 0),
\]
and
\[
[A, B]_\mathfrak{b} = [(0, A), (0, B)] = (0, [A, B]_\mathfrak{b}).
\]

Furthermore
\[
[(u, 0), (0, B)] = (-Bu, [0, B]_\mathfrak{b}) = (-Bu, 0) \in \mathfrak{a}.
\]

Hence, $\mathfrak{a}$ is an ideal in $\mathfrak{g}$. With this bracket structure $\mathfrak{g} = \mathfrak{se}(n)$, the Lie algebra of $\text{SE}(n)$. (See Section 1.6).

How does this bracket structure on $\mathfrak{g} = \mathfrak{se}(n)$ relate to $\text{Der}(\mathfrak{a})$? Since $\mathfrak{a} = \mathbb{R}^n$ is an abelian Lie algebra, $\text{Der}(\mathfrak{a}) = \mathfrak{gl}(n, \mathbb{R})$. Define $\tau : \mathfrak{b} \to \mathfrak{gl}(n, \mathbb{R})$ to be the inclusion map, i.e. $\tau(B) = B$ for $B \in \mathfrak{so}(n)$. Then
\[
[(u, A), (v, B)] = ((u, v)_\mathfrak{a} + \tau(A)v - \tau(B)v, [A, B]_\mathfrak{b}) = (Av - Bu, [A, B]_\mathfrak{b}).
\]

In other words
\[
[(0, A), (v, 0)] = (\tau(A)v, 0),
\]
and $[a, b]$ for $a \in \mathfrak{a} = \mathbb{R}^n$ and $b \in \mathfrak{b} = \mathfrak{so}(n)$ is determined by the map $\tau$.

The construction illustrated by this example is summarized in the following theorem.

**Proposition 15.17.** Let $\mathfrak{a}$ and $\mathfrak{b}$ be two Lie algebras, and suppose $\tau$ is a Lie algebra homomorphism $\tau : \mathfrak{b} \to \text{Der}(\mathfrak{a})$. Then there is a unique Lie algebra structure on the vector space
\[ g = a \oplus b \text{ whose Lie bracket agrees with the Lie bracket on } a \text{ and the Lie bracket on } b, \text{ and such that} \]
\[ [(0, B), (A, 0)]_g = \tau(B)(A) \quad \text{for all } A \in a \text{ and all } B \in b. \]

The Lie bracket on \( g = a \oplus b \) is given by
\[ [(A, B), (A', B')]_g = ([A, A']_a + \tau(B)(A') - \tau(B')(A), [B, B']_b), \]
for all \( A, A' \in a \) and all \( B, B' \in b \). In particular,
\[ [(0, B), (A', 0)]_g = \tau(B)(A') \in a. \]

With this Lie algebra structure, \( a \) is an ideal and \( b \) is a subalgebra.

Proof. Uniqueness of the Lie algebra structure is forced by the fact that the Lie bracket is bilinear and skew symmetric. The problem is to check the Jacobi identity. Pick \( X, Y, Z \in g \). If all three are in \( a \) or in \( b \), we are done. By skew symmetry, we are reduced to two cases:

1. \( X \) is in \( a \) and \( Y, Z \) are in \( b \), to simplify notation, write \( X \) for \((X, 0)\) and \( Y, Z \) for \((0, Y)\) and \((0, Z)\). Since \( \tau \) is a Lie algebra isomorphism,
\[ \tau([Y, Z]) = \tau(Y)\tau(Z) - \tau(Z)\tau(Y). \]
If we apply both sides to \( X \), we get
\[ \tau([Y, Z])(X) = (\tau(Y)\tau(Z))(X) - (\tau(Z)\tau(Y))(X), \]
that is,
\[ [[Y, Z], X] = [Y, [Z, X]] - [[Z, Y], X], \]
or equivalently
\[ [[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0, \]
which is the Jacobi identity.

2. \( X, Y \) are in \( a \) and \( Z \) is in \( b \), again to simplify notation, write \( X, Y \) for \((X, 0)\) and \((Y, 0)\) and \( Z \) for \((0, Z)\). Since \( \tau(Z) \) is a derivation, we have
\[ \tau(Z)([X, Y]) = [\tau(Z)(X), Y] + [X, \tau(Z)(Y)], \]
which is equivalent to
\[ [Z, [X, Y]] = [[Z, X], Y] + [X, [Z, Y]], \]
a version of the Jacobi identity.

Since both \( a \) and \( b \) bracket into \( a \), we conclude that \( a \) is an ideal. \( \square \)
The Lie algebra obtained in Proposition 15.17 is denoted by
\[ a \oplus_{\tau} b \quad \text{or} \quad a \rtimes_{\tau} b \]
and is called the \textit{semidirect product of} \( b \) \textit{by} \( a \) \textit{with respect to} \( \tau : b \to \text{Der}(a) \). When \( \tau \) is the zero map, we get back the Lie algebra direct sum.

\textbf{Remark}: A sequence of Lie algebra maps
\[ a \xrightarrow{\varphi} g \xrightarrow{\psi} b \]
with \( \varphi \) injective, \( \psi \) surjective, and with \( \text{Im} \varphi = \text{Ker} \psi = n \), is called an \textit{extension of} \( b \) \textit{by} \( a \) \textit{with kernel} \( n \). If there is a subalgebra \( p \) of \( g \) such that \( g = n \oplus p \), then we say that this extension is \textit{inessential}. Given a semidirect product \( g = a \rtimes_{\tau} b \) of \( b \) by \( a \), if \( \varphi : a \to g \) is the map given \( \varphi(a) = (a,0) \) and \( \psi : g \to b \) given by \( \psi(a,b) = b \), then \( g \) is an inessential extension of \( b \) by \( a \). Conversely, it is easy to see that every inessential extension of \( b \) by \( a \) is a semidirect product of \( b \) by \( a \).

Proposition 15.17 is an external construction. The notion of semidirect product has a corresponding internal construction. If \( g \) is a Lie algebra and if \( a \) and \( b \) are subspaces of \( g \) such that \( g = a \oplus b \), \( a \) is an ideal in \( g \) and \( b \) is a subalgebra of \( g \), then for every \( B \in b \), because \( a \) is an ideal, the restriction of \( \text{ad} B \) to \( a \) leaves \( a \) invariant, so by Proposition 15.16, the map \( B \mapsto \text{ad} B \upharpoonright a \) is a Lie algebra homomorphism \( \tau : b \to \text{Der}(a) \). Observe that \( [B, A] = \tau(B)(A) \), for all \( A \in a \) and all \( B \in b \), so the Lie bracket on \( g \) is completely determined by the Lie brackets on \( a \) and \( b \) and the homomorphism \( \tau \). We say that \( g \) is the \textit{semidirect product} of \( b \) and \( a \) and we write
\[ g = a \rtimes_{\tau} b. \]

Semidirect products of Lie algebras are discussed in Varadarajan [171] (Section 3.14), Bourbaki [28], (Chapter 1, Section 8), and Knapp [106] (Chapter 1, Section 4). However, beware that Knapp switches the roles of \( a \) and \( b \), and \( \tau \) is a Lie algebra map \( \tau : a \to \text{Der}(b) \).

Before turning our attention to semidirect products of Lie groups, let us consider the group \( \text{Aut}(g) \) of Lie algebra isomorphisms of a Lie algebra \( g \). The group \( \text{Aut}(g) \) is a subgroup of the groups \( \text{GL}(g) \) of linear automorphisms of \( g \), and it is easy to see that it is closed, so it is a Lie group. It turns out that its Lie algebra is \( \text{Der}(g) \).

\textbf{Proposition 15.18}. For any (real) Lie algebra \( g \), the Lie algebra \( L(\text{Aut}(g)) \) of the group \( \text{Aut}(g) \) is \( \text{Der}(g) \), the Lie algebra of derivations of \( g \).

\textit{Proof}. For any \( f \in L(\text{Aut}(g)) \), let \( \gamma(t) \) be a smooth curve in \( \text{Aut}(g) \) such that \( \gamma(0) = I \) and \( \gamma'(0) = f \). Since \( \gamma(t) \) is a Lie algebra automorphism
\[ \gamma(t)([X,Y]) = [\gamma(t)(X), \gamma(t)(Y)] \]
for all $X, Y \in \mathfrak{g}$, and using the product rule and taking the derivative for $t = 0$, we get
\[
\gamma'(0)([X, Y]) = f([X, Y]) = [\gamma'(0)(X), \gamma(0)(Y)] + [\gamma(0)(X), \gamma'(0)(Y)] \\
= [f(X), Y] + [Y, f(X)],
\]
which shows that $f$ is a derivation.

Conversely, pick any $f \in \text{Der}(\mathfrak{g})$. We prove that $e^{tf} \in \text{Aut}(\mathfrak{g})$ for all $t \in \mathbb{R}$, which shows that $\text{Der}(\mathfrak{g}) \subseteq \text{L}(\text{Aut}(\mathfrak{g}))$. For any $X, Y \in \mathfrak{g}$, consider the two curves in $\mathfrak{g}$ given by
\[
\gamma_1(t) = e^{tf}[X, Y] \quad \text{and} \quad \gamma_2(t) = [e^{tf}X, e^{tf}Y].
\]
For $t = 0$, we have $\gamma_1(0) = \gamma_2(0) = [X, Y]$. We find that
\[
\gamma_1'(t) = fe^{tf}[X, Y] = f\gamma_1(t),
\]
and since $f$ is a derivation
\[
\gamma_2'(t) = [fe^{tf}X, e^{tf}Y] + [e^{tf}X, fe^{tf}Y] \\
= f[e^{tf}X, e^{tf}Y] \\
= f\gamma_2(t).
\]
Since $\gamma_1$ and $\gamma_2$ are maximal integral curves for the linear vector field defined by $f$, and with the same initial condition, by uniqueness, we have
\[
e^{tf}[X, Y] = [e^{tf}X, e^{tf}Y] \quad \text{for all } t \in \mathbb{R},
\]
which shows that $e^{tf}$ is a Lie algebra automorphism. Therefore, $f \in \text{L}(\text{Aut}(\mathfrak{g}))$.

Remark: It can be shown that if $\mathfrak{g}$ is semisimple (see Section 17.4 for the definition of a semisimple Lie algebra), then $\text{ad}(\mathfrak{g}) = \text{Der}(\mathfrak{g})$.

We now define semidirect products of Lie groups and show how their algebras are semidirect products of Lie algebras. We begin with the definition of the semidirect product of two groups.
Proposition 15.19. Let $H$ and $K$ be two groups and let $\tau : K \to \text{Aut}(H)$ be a homomorphism of $K$ into the automorphism group of $H$, i.e. the set isomorphisms of $H$ with a group structure given by the composition of operators. Let $G = H \times K$ with multiplication defined as follows:

$$(h_1, k_1)(h_2, k_2) = (h_1\tau(k_1)(h_2), k_1k_2),$$

for all $h_1, h_2 \in H$ and all $k_1, k_2 \in K$. Then, the following properties hold:

1. This multiplication makes $G$ into a group with identity $(1, 1)$ and with inverse given by

$$(h, k)^{-1} = (\tau(k^{-1})(h^{-1}), k^{-1}).$$

2. The maps $h \mapsto (h, 1)$ for $h \in H$ and $k \mapsto (1, k)$ for $k \in K$ are isomorphisms from $H$ to the subgroup $\{(h, 1) \mid h \in H\}$ of $G$ and from $K$ to the subgroup $\{(1, k) \mid k \in K\}$ of $G$.

3. Using the isomorphisms from (2), the group $H$ is a normal subgroup of $G$.

4. Using the isomorphisms from (2), $H \cap K = (1)$.

5. For all $h \in H$ and all $k \in K$, we have

$$(1, k)(h, 1)(1, k)^{-1} = (\tau(k)(h), 1).$$

Proof. We leave the proof of these properties as an exercise, except for (5). Checking associativity takes a little bit of work.

Using the definition of multiplication, since $\tau(k_1)$ is an automorphism of $H$ for all $k_1 \in K$, we have $\tau(k_1)(1) = 1$, which means that

$$(1, k)^{-1} = (1, k^{-1}),$$

so we have

$$(1, k)(h, 1)(1, k)^{-1} = ((1, k)(h, 1))(1, k^{-1})$$

$$= (\tau(k)(h), k)(1, k^{-1})$$

$$= (\tau(k)(h)\tau(k)(1), kk^{-1})$$

$$= (\tau(k)(h), 1),$$

as claimed. 

In view of Proposition 15.19, we make the following definition.

Definition 15.10. Let $H$ and $K$ be two groups and let $\tau : K \to \text{Aut}(H)$ be a homomorphism of $K$ into the automorphism group of $H$. The group defined in Proposition 15.19 is called the *semidirect product of $K$ by $H$ with respect to $\tau$*, and it is denoted $H \rtimes_{\tau} K$ (or even $H \rtimes K$).
Note that $\tau: K \to \text{Aut}(H)$ can be viewed as a left action $\cdot: K \times H \to H$ of $K$ on $H$ “acting by automorphisms,” which means that for every $k \in K$, the map $h \mapsto \tau(k,h)$ is an automorphism of $H$.

Note that when $\tau$ is the trivial homomorphism (that is, $\tau(k) = \text{id}$ for all $k \in K$), the semidirect product is just the direct product $H \times K$ of the groups $H$ and $K$, and $K$ is also a normal subgroup of $G$.

Semidirect products are used to construct affine groups. For example, let $H = \mathbb{R}^n$ under addition, let $K = \text{SO}(n)$, and let $\tau$ be the inclusion map of $\text{SO}(n)$ into $\text{Aut}(\mathbb{R}^n)$. In other words, $\tau$ is the action of $\text{SO}(n)$ on $\mathbb{R}^n$ given by $R \cdot u = Ru$. Then, the semidirect product $\mathbb{R}^n \rtimes \text{SO}(n)$ is isomorphic to the group $\text{SE}(n)$ of direct affine rigid motions of $\mathbb{R}^n$ (translations and rotations), since the multiplication is given by

$$(u, R)(v, S) = (Ru + v, RS), \quad u, v \in \mathbb{R}^2, \quad R, S \in \text{SO}(n).$$

We obtain other affine groups by letting $K$ be $\text{SL}(n)$, $\text{GL}(n)$, etc.

Semidirect products of Lie algebras are discussed in Varadarajan [171] (Section 3.15), Bourbaki [28], (Chapter 3, Section 1.4), and Knapp [106] (Chapter 1, Section 15). Note that some authors (such as Knapp) define the semidirect product of two groups $H$ and $K$ by letting $H$ act on $K$. In this case, in order to work, the multiplication must be defined as

$$(h_1, k_1)(h_2, k_2) = (h_1h_2, \tau(h_2^{-1})(k_1)k_2),$$

which involves the inverse $h_2^{-1}$ of $h_2$. This is because $h_2$ acts on the element $k_1$ on its left, which makes it a right action. To work properly, we must use $h_2^{-1}$. In fact, $\tau: K \times H \to K$ is a right action of $H$ on $K$, and in this case, the map from $H$ to $\text{Aut}(K)$ should send $h$ to the map $k \mapsto \tau(h^{-1}, k)$, in order to be a homomorphism.

On the other hand, the way we have defined multiplication as

$$(h_1, k_1)(h_2, k_2) = (h_1\tau(k_1)(h_2), k_1k_2),$$

the element $k_1$ acts on the element $h_2$ on its right, which makes it a left action and works fine with no inversion needed. The left action seems simpler.

As in the case of Lie algebras, a sequence of groups homomorphisms

$$H \xrightarrow{\varphi} G \xrightarrow{\psi} K$$

with $\varphi$ injective, $\psi$ surjective, and with $\text{Im} \varphi = \text{Ker} \psi = N$, is called an extension of $K$ by $H$ with kernel $N$.

If $H \rtimes \tau K$ is a semidirect product, we have the homomorphisms $\varphi: H \to G$ and $\psi: G \to K$ given by

$$\varphi(h) = (h, 1), \quad \psi(h, k) = k,$$
and it is clear that we have an extension of $K$ by $H$ with kernel $N = \{(h, 1) \mid h \in H\}$. Note that we have a homomorphism $\gamma: K \to G$ (a section of $\psi$) given by

$$\gamma(k) = (1, k),$$

and that

$$\psi \circ \gamma = \text{id}.$$ 

Conversely, it can be shown that if an extension of $K$ by $H$ has a section $\gamma: K \to G$, then $G$ is isomorphic to a semidirect product of $K$ by $H$ with respect to a certain homomorphism $\tau$; find it!

I claim that if $H$ and $K$ are two Lie groups and if the map from $H \times K$ to $H$ given by $(h, k) \mapsto \tau(k)(h)$ is smooth, then the semidirect product $H \rtimes K$ is a Lie group (see Varadarajan [171] (Section 3.15), Bourbaki [28], (Chapter 3, Section 1.4)). This is because

$$(h_1, k_1)(h_2, k_2)^{-1} = (h_1, k_1)(\tau(k_2^{-1})(h_2^{-1}), k_2^{-1})$$

$$= (h_1\tau(k_1)(\tau(k_2^{-1})(h_2^{-1})), k_1k_2^{-1})$$

$$= (h_1\tau(k_1k_2^{-1})(h_2^{-1}), k_1k_2^{-1}),$$

which shows that multiplication and inversion in $H \rtimes K$ are smooth.

It it not very surprising that the Lie algebra of $H \rtimes K$ is a semidirect product of the Lie algebras $\mathfrak{h}$ of $H$ and $\mathfrak{k}$ of $K$.

For every $k \in K$, the derivative of $d(\tau(k))_1$ of $\tau(k)$ at 1 is a Lie algebra isomorphism of $\mathfrak{h}$, and just like $\text{Ad}$, it can be shown that the map $\tilde{\tau}: K \to \text{Aut}(\mathfrak{h})$ given by

$$\tilde{\tau}(k) = d(\tau(k))_1 \quad k \in K$$

is a smooth homomorphism from $K$ into $\text{Aut}(\mathfrak{h})$. It follows by Proposition 15.18 that its derivative $d\tilde{\tau}_1: \mathfrak{k} \to \text{Der}(\mathfrak{h})$ at 1 is a homomorphism of $\mathfrak{k}$ into $\text{Der}(\mathfrak{h})$.

**Proposition 15.20.** Using the notations just introduced, the Lie algebra of the semidirect product $H \rtimes K$ of $K$ by $H$ with respect to $\tau$ is the semidirect product $\mathfrak{h} \rtimes d\tilde{\tau}_1 \mathfrak{k}$ of $\mathfrak{k}$ by $\mathfrak{h}$ with respect to $d\tilde{\tau}_1$.

**Proof.** We follow Varadarajan [171] (Section 3.15), and provide a few more details. The tangent space at the identity of $H \rtimes K$ is $\mathfrak{h} \oplus \mathfrak{k}$ as a vector space. The bracket structure on $\mathfrak{h} \times \{0\}$ is inherited by the bracket on $\mathfrak{h}$, and similarly the bracket structure on $\{0\} \times \mathfrak{k}$ is inherited by the bracket on $\mathfrak{k}$. We need to figure out the bracket between elements of $\{0\} \times \mathfrak{k}$ and elements of $\mathfrak{h} \times \{0\}$. For any $X \in \mathfrak{h}$ and any $Y \in \mathfrak{k}$, for all $t, s \in \mathbb{R}$, using Proposition 15.6(2), Property (5) of Proposition 15.19, and the fact that $\exp(X, Y) = (\exp(X), \exp(Y))$, we have

$$\exp(\text{Ad}(\exp(t, 0, Y)))(s(X, 0)) = (\exp(t, 0, Y))(\exp(s(X, 0)))(\exp(t, 0, Y))^{-1}$$

$$= (1, \exp(tY))(\exp(sX), 1)(1, \exp(-tY))$$

$$= (\tau(\exp(tY))(\exp(sX)), 1).$$
For fixed $t$, taking the derivative with respect to $s$ at $s = 0$, we deduce that

$$\text{Ad}(\exp(t(0, Y)))(X, 0) = (\tilde{\tau}(\exp(tY))(X), 0).$$

Taking the derivative with respect to $t$ at $t = 0$, we get

$$[(0, Y), (X, 0)] = (\text{ad} (0, Y))(X, 0) = (d\tilde{\tau}_1(Y)(X), 0),$$

which shows that the Lie bracket between elements of $\{0\} \times \mathfrak{t}$ and elements of $\mathfrak{h} \times \{0\}$ is given by $d\tilde{\tau}_1$.

Proposition 15.20 applied to the semidirect product $\mathbb{R}^n \rtimes \tau \text{SO}(n) \cong \text{SE}(n)$ where $\tau$ is the inclusion map of $\text{SO}(n)$ into $\text{Aut}(\mathbb{R}^n)$ confirms that $\mathbb{R}^n \rtimes d\tilde{\tau}_1 \text{so}(n)$ is the Lie algebra of $\text{SE}(n)$, where $d\tilde{\tau}_1$ is inclusion map of $\text{so}(n)$ into $\text{gl}(n, \mathbb{R})$ (and $\tilde{\tau}$ is the inclusion of $\text{SO}(n)$ into $\text{Aut}(\mathbb{R}^n)$).

As a special case of Proposition 15.20, when our semidirect product is just a direct product $H \times K$ ($\tau$ is the trivial homomorphism mapping every $k \in K$ to id), we see that the Lie algebra of $H \times K$ is the Lie algebra direct sum $\mathfrak{h} \oplus \mathfrak{k}$ (where the bracket between elements of $\mathfrak{h}$ and elements of $\mathfrak{k}$ is 0).

### 15.6 Universal Covering Groups ⋆

Every connected Lie group $G$ is a manifold, and as such, from results in Section 9.2, it has a universal cover $\tilde{\pi}: \tilde{G} \to G$, where $\tilde{G}$ is simply connected. It is possible to make $\tilde{G}$ into a group so that $\tilde{G}$ is a Lie group and $\pi$ is a Lie group homomorphism. We content ourselves with a sketch of the construction whose details can be found in Warner [175], Chapter 3.

Consider the map $\alpha: \tilde{G} \times \tilde{G} \to G$, given by

$$\alpha(\tilde{a}, \tilde{b}) = \pi(\tilde{a})\pi(\tilde{b})^{-1},$$

for all $\tilde{a}, \tilde{b} \in \tilde{G}$, and pick some $\tilde{e} \in \pi^{-1}(1)$. Since $\tilde{G} \times \tilde{G}$ is simply connected, it follows by Proposition 9.12 that there is a unique map $\tilde{\alpha}: \tilde{G} \times \tilde{G} \to \tilde{G}$ such that

$$\alpha = \pi \circ \tilde{\alpha} \quad \text{and} \quad \tilde{e} = \tilde{\alpha}(\tilde{e}, \tilde{e}).$$

For all $\tilde{a}, \tilde{b} \in \tilde{G}$, define

$$\tilde{b}^{-1} = \tilde{\alpha}(\tilde{e}, \tilde{b}), \quad \tilde{a}\tilde{b} = \tilde{\alpha}(\tilde{a}, \tilde{b}^{-1}). \quad \text{(*)}$$

Using Proposition 9.12, it can be shown that the above operations make $\tilde{G}$ into a group, and as $\tilde{\alpha}$ is smooth, into a Lie group. Moreover, $\pi$ becomes a Lie group homomorphism. We summarize these facts as

**Theorem 15.21.** Every connected Lie group has a simply connected covering map $\pi: \tilde{G} \to G$, where $\tilde{G}$ is a Lie group and $\pi$ is a Lie group homomorphism.
The group $\tilde{G}$ is called the universal covering group of $G$. Consider $D = \ker \pi$. Since the fibres of $\pi$ are countable, the group $D$ is a countable closed normal subgroup of $\tilde{G}$; that is, a discrete normal subgroup of $\tilde{G}$. It follows that $G \cong \tilde{G}/D$, where $\tilde{G}$ is a simply connected Lie group and $D$ is a discrete normal subgroup of $\tilde{G}$.

We conclude this section by stating the following useful proposition whose proof can be found in Warner [175] (Chapter 3, Proposition 3.26):

**Proposition 15.22.** Let $\phi: G \to H$ be a homomorphism of connected Lie groups. Then $\phi$ is a covering map iff $d\phi_1: g \to h$ is an isomorphism of Lie algebras.

For example, we know that $\mathfrak{su}(2) = \mathfrak{so}(3)$, so the homomorphism from $\text{SU}(2)$ to $\text{SO}(3)$ provided by the representation of 3D rotations by the quaternions is a covering map.

### 15.7 The Lie Algebra of Killing Fields

In Section 14.4, we defined Killing vector fields. Recall that a Killing vector field $X$ on a manifold $M$ satisfies the condition

$$L_X g(Y, Z) = X(\langle Y, Z \rangle) - \langle [X, Y], Z \rangle - \langle Y, [X, Z] \rangle = 0,$$

for all $X, Y, Z \in \mathfrak{X}(M)$. By Proposition 14.8, $X$ is a Killing vector field iff the diffeomorphisms $\Phi_t$ induced by the flow $\Phi$ of $X$ are isometries (on their domain).

The isometries of a Riemannian manifold $(M, g)$ form a group $\text{Isom}(M, g)$, called the isometry group of $(M, g)$. An important theorem of Myers and Steenrod asserts that the isometry group $\text{Isom}(M, g)$ is a Lie group. It turns out that the Lie algebra $\mathfrak{i}(M)$ of the group $\text{Isom}(M, g)$ is closely related to a certain Lie subalgebra of the Lie algebra of Killing fields. In this section, we briefly explore this relationship.

We begin by observing that the Killing fields form a Lie algebra. This is because the Lie derivative $L_X$ is $\mathbb{R}$-linear in $X$, and since

$$[L_X, L_Y] = L_{[X,Y]} ,$$

the Killing vector fields on $M$ form a Lie subalgebra $\mathcal{K}i(M)$ of the Lie algebra $\mathfrak{X}(M)$ of vector fields on $M$. However, unlike $\mathfrak{X}(M)$, the Lie algebra $\mathcal{K}i(M)$ is finite-dimensional. In fact, the Lie subalgebra $c\mathcal{K}i(M)$ of complete Killing vector fields is anti-isomorphic to the Lie algebra $\mathfrak{i}(M)$ of the Lie group $\text{Isom}(M)$ of isometries of $M$ (complete vector fields are defined in Definition 8.10). The following result is proved in O’Neill [138] (Chapter 9, Lemma 28) and Sakai [150] (Chapter III, Lemma 6.4 and Proposition 6.5).

**Proposition 15.23.** Let $(M, g)$ be a connected Riemannian manifold of dimension $n$ (equipped with the Levi–Civita connection on $M$ induced by $g$). The Lie algebra $\mathcal{K}i(M)$ of Killing vector fields on $M$ has dimension at most $n(n + 1)/2$. 

We also have the following result proved in O’Neill [138] (Chapter 9, Proposition 30) and Sakai [150] (Chapter III, Corollary 6.3).

**Proposition 15.24.** Let \((M, g)\) be a Riemannian manifold of dimension \(n\) (equipped with the Levi–Civita connection on \(M\) induced by \(g\)). If \(M\) is complete, then every Killing vector fields on \(M\) is complete.

The relationship between the Lie algebra \(\mathfrak{i}(M)\) and Killing vector fields is obtained as follows. For every element \(X\) in the Lie algebra \(\mathfrak{i}(M)\) of \(\text{Isom}(M)\) (viewed as a left-invariant vector field), define the vector field \(X^+\) on \(M\) by

\[
X^+(p) = \left. \frac{d}{dt} (\varphi_t(p)) \right|_{t=0}, \quad p \in M,
\]

where \(t \mapsto \varphi_t = \exp(tX)\) is the one-parameter group associated with \(X\). Because \(\varphi_t\) is an isometry of \(M\), the vector field \(X^+\) is a Killing vector field, and it is also easy to show that \((\varphi_t)\) is the one-parameter group of \(X^+\). Since \(\varphi_t\) is defined for all \(t\), the vector field \(X^+\) is complete. The following result is shown in O’Neill [138] (Chapter 9, Proposition 33).

**Theorem 15.25.** Let \((M, g)\) be a Riemannian manifold (equipped with the Levi–Civita connection on \(M\) induced by \(g\)). The following properties hold:

1. The set \(\mathfrak{cK}(M)\) of complete Killing vector fields on \(M\) is a Lie subalgebra of the Lie algebra \(\mathfrak{K}(M)\) of Killing vector fields.

2. The map \(X \mapsto X^+\) is a Lie anti-isomorphism between \(\mathfrak{i}(M)\) and \(\mathfrak{cK}(M)\), which means that

\[
[X^+, Y^+] = -[X, Y]^+, \quad X, Y \in \mathfrak{i}(M).
\]

For more on Killing vector fields, see Sakai [150] (Chapter III, Section 6). In particular, complete Riemannian manifolds for which \(\mathfrak{i}(M)\) has the maximum dimension \(n(n+1)/2\) are characterized.
Chapter 16

The Derivative of exp and Dynkin’s Formula ⊗

16.1 The Derivative of the Exponential Map

We know that if $[X, Y] = 0$, then $\exp(X + Y) = \exp(X) \exp(Y)$, but this generally false if $X$ and $Y$ do not commute. For $X$ and $Y$ in a small enough open subset $U$ containing 0, we know that $\exp$ is a diffeomorphism from $U$ to its image, so the function $\mu: U \times U \to U$ given by

$$\mu(X, Y) = \log(\exp(X) \exp(Y))$$

is well-defined and it turns out that for $U$ small enough, it is analytic. Thus, it is natural to seek a formula for the Taylor expansion of $\mu$ near the origin.

This problem was investigated by Campbell (1897/98), Baker (1905) and in a more rigorous fashion by Hausdorff (1906). These authors gave recursive identities expressing the Taylor expansion of $\mu$ at the origin and the corresponding result is often referred to as the Campbell-Baker-Hausdorff Formula. F. Schur (1891) and Poincaré (1899) also investigated the exponential map, in particular formulae for its derivative and the problem of expressing the function $\mu$. However, it was Dynkin who finally gave an explicit formula (see Section 16.3) in 1947.

The proof that $\mu$ is analytic in a suitable domain can be proved using a formula for the derivative of the exponential map, a formula that was obtained by F. Schur and Poincaré. Thus, we begin by presenting such a formula.

First we introduce a convenient notation. If $A$ is any real (or complex) $n \times n$ matrix, the following formula is clear:

$$\int_0^1 e^{tA} dt = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}.$$
If $A$ is invertible, then the right-hand side can be written explicitly as
\[ \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!} = A^{-1}(e^A - I), \]
and we also write the latter as
\[ \frac{e^A - I}{A} = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}. \]

Even if $A$ is not invertible, we use (*) as the definition of $\frac{e^A - I}{A}$.

We can use the following trick to figure out what $(d\exp_X)(Y)$ is:
\[ (d\exp_X)(Y) = \frac{d}{d\epsilon} \bigg|_{\epsilon=0} \exp(X + \epsilon Y) = \frac{d}{d\epsilon} \bigg|_{\epsilon=0} (dR_{\exp(X+\epsilon Y)})_1, \]
since by Proposition 15.2, the map $s \mapsto R_{\exp s(X+\epsilon Y)}$ is the flow of the left-invariant vector field $(X + \epsilon Y)^L$ on $G$. Now, $(X + \epsilon Y)^L$ is an $\epsilon$-dependent vector field which depends on $\epsilon$ in a $C^1$ fashion. From the theory of ODE’s, if $p \mapsto v_\epsilon(p)$ is a smooth vector field depending in a $C^1$ fashion on a real parameter $\epsilon$ and if $\Phi^\epsilon_t$ denotes its flow (after time $t$), then we have the variational formula
\[ \frac{\partial \Phi^\epsilon_t}{\partial \epsilon}(x) = \int_0^t d(\Phi^\epsilon_{t-s})_{\Phi^\epsilon_s(x)} \frac{\partial v_\epsilon}{\partial \epsilon}(\Phi^\epsilon_s(x)) ds. \]

See Duistermaat and Kolk [64], Appendix B, Formula (B.10). Using this, the following is proved in Duistermaat and Kolk [64] (Chapter 1, Section 1.5):

**Proposition 16.1.** Given any Lie group $G$, for any $X \in \mathfrak{g}$, the linear map $d\exp_X : \mathfrak{g} \to T_{\exp(X)}G$ is given by
\[ d\exp_X = (dR_{\exp(X)})_1 \circ \int_0^1 e^{sadX} ds = (dR_{\exp(X)})_1 \circ \frac{e^{adX} - I}{adX} \]
\[ = (dL_{\exp(X)})_1 \circ \int_0^1 e^{-sadX} ds = (dL_{\exp(X)})_1 \circ \frac{I - e^{-adX}}{adX}. \]

**Remark:** If $G$ is a matrix group of $n \times n$ matrices, we see immediately that the derivative of left multiplication $(X \mapsto L_A X = AX)$ is given by
\[ (dL_A)_X Y = AY, \]
for all $n \times n$ matrices $X, Y$. Consequently, for a matrix group, we get
\[ d\exp_X = e^X \left( \frac{I - e^{-adX}}{adX} \right). \]
16.2. THE PRODUCT IN LOGARITHMIC COORDINATES

An alternative proof sketch of this fact is provided in Section 2.4.

Now, if $A$ is a real matrix, it is clear that the (complex) eigenvalues of $\int_0^1 e^{sA} ds$ are of the form

$$\frac{e^\lambda - 1}{\lambda} \quad (= 1 \quad \text{if } \lambda = 0),$$

where $\lambda$ ranges over the (complex) eigenvalues of $A$. Consequently, we get

**Proposition 16.2.** The singular points of the exponential map $\exp : g \to G$, that is, the set of $X \in g$ such that $d\exp_X$ is singular (not invertible), are the $X \in g$ such that the linear map $\text{ad} X : g \to g$ has an eigenvalue of the form $k2\pi i$, with $k \in \mathbb{Z}$ and $k \neq 0$.

Another way to describe the singular locus $\Sigma$ of the exponential map is to say that it is the disjoint union

$$\Sigma = \bigcup_{k \in \mathbb{Z} - \{0\}} k\Sigma_1,$$

where $\Sigma_1$ is the algebraic variety in $g$ given by

$$\Sigma_1 = \{X \in g \mid \det(\text{ad} X - 2\pi i I) = 0\}.$$

For example, for $\text{SL}(2, \mathbb{R})$,

$$\Sigma_1 = \left\{ \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2) \mid a^2 + bc = -\pi^2 \right\},$$

a two-sheeted hyperboloid mapped to $-I$ by $\exp$.

Let $g_e = g - \Sigma$ be the set of $X \in g$ such that $\frac{e^{\text{ad} X} - I}{\text{ad} X}$ is invertible. This is an open subset of $g$ containing 0.

### 16.2 The Product in Logarithmic Coordinates

Since the map,

$$X \mapsto \frac{e^{\text{ad} X} - I}{\text{ad} X}$$

is invertible for all $X \in g_e = g - \Sigma$, in view of the chain rule, the reciprocal (multiplicative inverse) of the above map

$$X \mapsto \frac{\text{ad} X}{e^{\text{ad} X} - I},$$

is an analytic function from $g_e$ to $\mathfrak{gl}(g, g)$. Let $g_e^2$ be the subset of $g \times g_e$ consisting of all $(X, Y)$ such that the solution $t \mapsto Z(t)$ of the differential equation

$$\frac{dZ(t)}{dt} = \frac{\text{ad} Z(t)}{e^{\text{ad} Z(t)} - I}(X)$$

...
with initial condition $Z(0) = Y(\in \mathfrak{g}_e)$ is defined for all $t \in [0, 1]$. Set

$$
\mu(X, Y) = Z(1), \quad (X, Y) \in \mathfrak{g}_e^2.
$$

The following theorem is proved in Duistermaat and Kolk [64] (Chapter 1, Section 1.6, Theorem 1.6.1):

**Theorem 16.3.** Given any Lie group $G$ with Lie algebra $\mathfrak{g}$, the set $\mathfrak{g}_e^2$ is an open subset of $\mathfrak{g} \times \mathfrak{g}$ containing $(0, 0)$, and the map $\mu: \mathfrak{g}_e^2 \to \mathfrak{g}$ is real-analytic. Furthermore, we have

$$
\exp(X) \exp(Y) = \exp(\mu(X, Y)), \quad (X, Y) \in \mathfrak{g}_e^2,
$$

where $\exp: \mathfrak{g} \to G$. If $\mathfrak{g}$ is a complex Lie algebra, then $\mu$ is complex-analytic.

We may think of $\mu$ as the product in logarithmic coordinates. It is explained in Duistermaat and Kolk [64] (Chapter 1, Section 1.6) how Theorem 16.3 implies that a Lie group can be provided with the structure of a real-analytic Lie group. Rather than going into this, we will state a remarkable formula due to Dynkin expressing the Taylor expansion of $\mu$ at the origin.

### 16.3 Dynkin’s Formula

As we said in Section 16.1, the problem of finding the Taylor expansion of $\mu$ near the origin was investigated by Campbell (1897/98), Baker (1905) and Hausdorff (1906). However, it was Dynkin who finally gave an explicit formula in 1947. There are actually slightly different versions of Dynkin’s formula. One version is given (and proved convergent) in Duistermaat and Kolk [64] (Chapter 1, Section 1.7). Another slightly more explicit version (because it gives a formula for the homogeneous components of $\mu(X, Y)$) is given (and proved convergent) in Bourbaki [28] (Chapter II, §6, Section 4) and Serre [159] (Part I, Chapter IV, Section 8). We present the version in Bourbaki and Serre without proof. The proof uses formal power series and free Lie algebras.

Given $X, Y \in \mathfrak{g}_e^2$, we can write

$$
\mu(X, Y) = \sum_{n=1}^{\infty} z_n(X, Y),
$$

where $z_n(X, Y)$ is a homogeneous polynomial of degree $n$ in the non-commuting variables $X, Y$.

**Theorem 16.4.** (Dynkin’s Formula) If we write $\mu(X, Y) = \sum_{n=1}^{\infty} z_n(X, Y)$, then we have

$$
z_n(X, Y) = \frac{1}{n} \sum_{p+q=n} (z_{p,q}^'(X, Y) + z_{p,q}''(X, Y)),
$$

where $z_{p,q}^'(X, Y)$ and $z_{p,q}''(X, Y)$ are homogeneous polynomials of degree $p$ and $q$, respectively.
with

\[ z'_{p,q}(X,Y) = \sum_{p_1+\cdots+p_m=p, \, \sum_{i=1}^{m-1} q_i = q-1, \, p_i, q_i \geq 1, \, p_m, q_m \geq 1} \frac{(-1)^{m+1}}{m} \left( \prod_{i=1}^{m-1} \frac{(\text{ad } X)^{p_i}}{p_i!} \frac{(\text{ad } Y)^{q_i}}{q_i!} \right) \left( \frac{\text{ad } X}{p_m!} \right) (Y) \]

and

\[ z''_{p,q}(X,Y) = \sum_{p_1+\cdots+p_m-1=p-1, \, \sum_{i=1}^{m-1} q_i = q, \, p_i, q_i \geq 1, \, m \geq 1} \frac{(-1)^{m+1}}{m} \left( \prod_{i=1}^{m-1} \frac{(\text{ad } X)^{p_i}}{p_i!} \frac{(\text{ad } Y)^{q_i}}{q_i!} \right) (X). \]

As a concrete illustration of Dynkin's formula, after some labor, the following Taylor expansion up to order 4 is obtained:

\[ \mu(X,Y) = X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \frac{1}{12}[Y, [Y, X]] - \frac{1}{24}[X, [Y, [X, Y]]] + \text{higher order terms}. \]

Observe that due the lack of associativity of the Lie bracket quite different looking expressions can be obtained using the Jacobi identity. For example,

\[ -[X, [Y, [X, Y]]] = [Y, [X, [Y, X]]]. \]

There is also an integral version of the Campbell-Baker-Hausdorff formula; see Hall [84] (Chapter 3).
CHAPTER 16. THE DERIVATIVE OF EXP AND DYNKIN’S FORMULA
Chapter 17

Metrics, Connections, and Curvature on Lie Groups

Since a Lie group $G$ is a smooth manifold, we can endow $G$ with a Riemannian metric. Among all the Riemannian metrics on a Lie groups, those for which the left translations (or the right translations) are isometries are of particular interest because they take the group structure of $G$ into account. As a consequence, it is possible to find explicit formulae for the Levi-Civita connection and the various curvatures, especially in the case of metrics which are both left and right-invariant.

In Section 17.1 we define left-invariant and right-invariant metrics on a Lie group. We show that left-invariant metrics are obtained by picking some inner product on $\mathfrak{g}$ and moving it around the group to the other tangent spaces $T_gG$ using the maps $(dL_g^{-1})_g$ (with $g \in G$). Right-invariant metrics are obtained by using the maps $(dR_g^{-1})_g$.

In Section 17.2 we give four characterizations of bi-invariant metrics. The first one refines the criterion of the existence of a left-invariant metric and states that every bi-invariant metrics on a Lie group $G$ arises from some Ad-invariant inner products on the Lie algebra $\mathfrak{g}$.

In Section 17.3 we show that if $G$ is a Lie group equipped with a left-invariant metric, then it is possible to express the Levi-Civita connection and the sectional curvature in terms of quantities defined over the Lie algebra of $G$, at least for left-invariant vector fields. When the metric is bi-invariant, much nicer formulae are be obtained. In particular the geodesics coincide with the one-parameter groups induced by left-invariant vector fields.

Section 17.4 introduces simple and semisimple Lie algebras. They play a major role in the structure theory of Lie groups.

Section 17.5 is devoted to the Killing form. It is an important concept, and we establish some of its main properties. Remarkably, the Killing form yields a simple criterion due to Élie Cartan for testing whether a Lie algebra is semisimple. Indeed, a Lie algebra $\mathfrak{g}$ is semisimple iff its Killing form $B$ is non-degenerate. We also show that a connected Lie group is compact and semisimple iff its Killing form is negative definite.
We conclude this chapter with a section on Cartan connections (Section 17.6). Unfortunately, if a Lie group $G$ does not admit a bi-invariant metric, under the Levi-Civita connection, geodesics are generally not given by the exponential map $\exp : \mathfrak{g} \to G$. If we are willing to consider connections not induced by a metric, then it turns out that there is a fairly natural connection for which the geodesics coincide with integral curves of left-invariant vector fields. We are led to consider left-invariant connections. It turns out that there is a one-to-one correspondence between left-invariant connections and bilinear maps $\alpha : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. Connections for which the geodesics are given by the exponential map are those for which $\alpha$ is skew-symmetric. These connections are called Cartan connections.

This chapter makes extensive use of results from a beautiful paper of Milnor [128].

17.1 Left (resp. Right) Invariant Metrics

In a Lie group $G$, since the operations $dL_a$ and $dR_a$ are diffeomorphisms for all $a \in G$, it is natural to consider the metrics for which these maps are isometries.

**Definition 17.1.** A metric $\langle - , - \rangle$ on a Lie group $G$ is called left-invariant (resp. right-invariant) iff

$$\langle u,v \rangle_b = \langle (dL_a)_b u, (dL_a)_b v \rangle_a$$

(resp. $\langle u,v \rangle_b = \langle (dR_a)_b u, (dR_a)_b v \rangle_a$),

for all $a,b \in G$ and all $u,v \in T_b G$. A Riemannian metric that is both left and right-invariant is called a bi-invariant metric.

As shown in the next proposition, left-invariant (resp. right-invariant) metrics on $G$ are induced by inner products on the Lie algebra, $\mathfrak{g}$, of $G$. In what follows the identity element of the Lie group, $G$, will be denoted by $e$ or $1$.

**Proposition 17.1.** There is a bijective correspondence between left-invariant (resp. right invariant) metrics on a Lie group $G$, and inner products on the Lie algebra $\mathfrak{g}$ of $G$.

**Proof.** If the metric on $G$ is left-invariant, then for all $a \in G$ and all $u,v \in T_a G$, we have

$$\langle u,v \rangle_a = \langle (dL_{a^{-1}})_a u, (dL_{a^{-1}})_a v \rangle_e,$$

which shows that our metric is completely determined by its restriction to $\mathfrak{g} = T_e G$. Conversely, let $\langle - , - \rangle$ be an inner product on $\mathfrak{g}$ and set

$$\langle u,v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle,$$

for all $u,v \in T_g G$ and all $g \in G$. Obviously, the family of inner products, $\langle - , - \rangle_g$, yields a Riemannian metric on $G$. To prove that it is left-invariant, we use the chain rule and the
17.2. BI-IN VARIANT METRICS

fact that left translations are group isomorphisms. For all \( a, b \in G \) and all \( u, v \in T_bG \), we have

\[
\langle (dL_a)_b u, (dL_a)_b v \rangle_{ab} = \langle (dL_{(ab)^{-1}})_b ((dL_a)_b u), (dL_{(ab)^{-1}})_b ((dL_a)_b v) \rangle \\
= \langle d(L_{(ab)^{-1}} \circ L_a)_b u, d(L_{(ab)^{-1}} \circ L_a)_b v \rangle \\
= \langle d(L_{b^{-1}a^{-1}} \circ L_a)_b u, d(L_{b^{-1}a^{-1}} \circ L_a)_b v \rangle \\
= \langle (dL_{b^{-1}})_b u, (dL_{b^{-1}})_b v \rangle \\
= \langle u, v \rangle_b,
\]

as desired.

To get a right-invariant metric on \( G \), set

\[
\langle u, v \rangle_g = \langle (dR_{g^{-1}})_g u, (dR_{g^{-1}})_g v \rangle,
\]

for all \( u, v \in T_gG \) and all \( g \in G \). The verification that this metric is right-invariant is analogous.

If \( G \) has dimension \( n \), then since inner products on \( g \) are in one-to-one correspondence with \( n \times n \) positive definite matrices, we see that \( G \) possesses a family of left-invariant metrics of dimension \( \frac{1}{2} n(n + 1) \).

If \( G \) has a left-invariant (resp. right-invariant) metric, since left-invariant (resp. right-invariant) translations are isometries and act transitively on \( G \), the space \( G \) is called a homogeneous Riemannian manifold.

**Proposition 17.2.** Every Lie group \( G \) equipped with a left-invariant (resp. right-invariant) metric is complete.

**Proof.** As \( G \) is locally compact, we can pick some \( \epsilon > 0 \) small enough so that the closed \( \epsilon \)-ball about the identity is compact. By translation, every \( \epsilon \)-ball is compact, hence every Cauchy sequence eventually lies within a compact set, and thus, converges.

We now give four characterizations of bi-invariant metrics.

### 17.2 Bi-Invariant Metrics

Recall that the adjoint representation \( \text{Ad}: G \rightarrow \text{GL}(g) \) of the Lie group \( G \) is the map defined such that \( \text{Ad}_a: g \rightarrow g \) is the linear isomorphism given by

\[
\text{Ad}_a = d(\text{Ad}_a)_e = d(R_{a^{-1}} \circ L_a)_e, \quad \text{for every } a \in G.
\]

Clearly,

\[
\text{Ad}_a = (dR_{a^{-1}})_a \circ (dL_a)_e.
\]

Here is the first of four criteria for the existence of a bi-invariant metric on a Lie group.
Proposition 17.3. There is a bijective correspondence between bi-invariant metrics on a Lie group $G$ and $\text{Ad}$-invariant inner products on the Lie algebra $\mathfrak{g}$ of $G$, namely inner products $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}$ such that $\text{Ad}_a$ is an isometry of $\mathfrak{g}$ for all $a \in G$; more explicitly, $\text{Ad}$-invariant inner inner products satisfy the condition

$$\langle \text{Ad}_a u, \text{Ad}_a v \rangle = \langle u, v \rangle,$$

for all $a \in G$ and all $u, v \in \mathfrak{g}$.

Proof. If $\langle \cdot, \cdot \rangle$ is a bi-invariant metric on $G$, as

$$\text{Ad}_a = (dR_a^{-1})_a \circ (dL_a)_e,$$

we claim that

$$\langle \text{Ad}_a u, \text{Ad}_a v \rangle = \langle u, v \rangle,$$

which means that $\text{Ad}_a$ is an isometry on $\mathfrak{g}$. To prove this claim, first observe that the left-invariance of the metric gives

$$\langle (dL_a)_e u, (dL_a)_e v \rangle_a = \langle u, v \rangle.$$

Define $U = (dL_a)_e u \in T_a G$ and $V = (dL_a)_e v \in T_a G$. This time, the right-invariance of the metric implies

$$\langle (dR_a^{-1})_a U, (dR_a^{-1})_a V \rangle = \langle U, V \rangle_a = \langle (dL_a)_e u, (dL_a)_e v \rangle_a.$$

Since $\langle (dR_a^{-1})_a U, (dR_a^{-1})_a V \rangle = \langle \text{Ad}_a u, \text{Ad}_a v \rangle$, the previous equation verifies the claim.

Conversely, if $\langle \cdot, \cdot \rangle$ is any inner product on $\mathfrak{g}$ such that $\text{Ad}_a$ is an isometry of $\mathfrak{g}$ for all $a \in G$, we need to prove that the metric on $G$ given by

$$\langle u, v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle$$

is also right-invariant. We have

$$\langle (dR_a)_b u, (dR_a)_b v \rangle_{ba} = \langle (dL_{(ba)^{-1}})_b (dR_a)_b u, (dL_{(ba)^{-1}})_b (dR_a)_b v \rangle$$

$$= \langle d(L_{(a^{-1})} \circ L_{b^{-1}} \circ R_a)_b u, d(L_{(a^{-1})} \circ L_{b^{-1}} \circ R_a)_b v \rangle$$

$$= \langle d(R_a \circ L_{a^{-1}} \circ L_{b^{-1}})_b u, d(R_a \circ L_{a^{-1}} \circ L_{b^{-1}})_b v \rangle$$

$$= \langle d(R_a \circ L_{a^{-1}})_e \circ d(L_{b^{-1}})_b u, d(R_a \circ L_{a^{-1}})_e \circ d(L_{b^{-1}})_b v \rangle$$

$$= \langle \text{Ad}_{a^{-1}} \circ d(L_{b^{-1}})_b u, \text{Ad}_{a^{-1}} \circ d(L_{b^{-1}})_b v \rangle$$

$$= \langle u, v \rangle_b,$$

as $\langle \cdot, \cdot \rangle$ is left-invariant and $\text{Ad}_g$-invariant for all $g \in G$. 

$\square$
Proposition 17.3 shows that if a Lie group $G$ possesses a bi-invariant metric, then every linear map $\text{Ad}_a$ is an orthogonal transformation of $\mathfrak{g}$. It follows that $\text{Ad}(G)$ is a subgroup of the orthogonal group of $\mathfrak{g}$, and so its closure $\overline{\text{Ad}(G)}$ is compact. It turns out that this condition is also sufficient!

To prove the above fact, we make use of an “averaging trick” used in representation theory. A representation of a Lie group $G$ is a (smooth) homomorphism $\rho: G \to \text{GL}(V)$, where $V$ is some finite-dimensional vector space. For any $g \in G$ and any $u \in V$, we often write $g \cdot u$ for $\rho(g)(u)$. We say that an inner-product $\langle -, - \rangle$ on $V$ is $G$-invariant iff

$$\langle g \cdot u, g \cdot v \rangle = \langle u, v \rangle, \quad \text{for all } g \in G \text{ and all } u, v \in V.$$ 

If $G$ is compact, then the “averaging trick,” also called “Weyl’s unitarian trick,” yields the following important result:

**Theorem 17.4.** If $G$ is a compact Lie group, then for every representation $\rho: G \to \text{GL}(V)$, there is a $G$-invariant inner product on $V$.

**Proof.** This proof uses a fact shown in Section 24.7, namely that a notion of integral invariant with respect to left and right multiplication can be defined on any compact Lie group. Thus the reader may skip this proof until she/he reads Chapter 24.

In Section 24.7 it is shown that a Lie group is orientable, has a left-invariant volume form $\omega$, and for every continuous function $f$ with compact support, we can define the integral $\int_G f = \int_G f \omega$. Furthermore, when $G$ is compact, we may assume that our integral is normalized so that $\int_G \omega = 1$ and in this case, our integral is both left and right invariant. Given any inner product $\langle -, - \rangle$ on $V$, set

$$\langle\langle u, v \rangle \rangle = \int_G \langle g \cdot u, g \cdot v \rangle, \quad \text{for all } u, v \in V,$$

where $\langle g \cdot u, g \cdot v \rangle$ denotes the function $g \mapsto \langle g \cdot u, g \cdot v \rangle$. It is easily checked that $\langle\langle -, - \rangle \rangle$ is an inner product on $V$. Furthermore, using the right-invariance of our integral (that is, $\int_G f = \int_G (f \circ R_h)$, for all $h \in G$), we have

$$\langle\langle h \cdot u, h \cdot v \rangle \rangle = \int_G \langle g \cdot (h \cdot u), g \cdot (h \cdot v) \rangle$$

$$= \int_G \langle (gh) \cdot u, (gh) \cdot v \rangle$$

$$= \int_G \langle g \cdot u, g \cdot v \rangle$$

$$= \langle\langle u, v \rangle \rangle,$$

which shows that $\langle\langle -, - \rangle \rangle$ is $G$-invariant. \qed
CHAPTER 17. METRICS, CONNECTIONS, AND CURVATURE ON LIE GROUPS

Using Theorem 17.4, we can prove the following result giving a criterion for the existence of a $G$-invariant inner product for any representation of a Lie group $G$ (see Sternberg [166], Chapter 5, Theorem 5.2).

**Theorem 17.5.** Let $\rho: G \rightarrow \text{GL}(V)$ be a (finite-dimensional) representation of a Lie group $G$. There is a $G$-invariant inner product on $V$ iff $\overline{\rho(G)}$ is compact. In particular, if $G$ is compact, then there is a $G$-invariant inner product on $V$.

**Proof.** If $V$ has a $G$-invariant inner product on $V$, then each linear map, $\rho(g)$, is an isometry, so $\rho(G)$ is a subgroup of the orthogonal group $O(V)$, of $V$. As $O(V)$ is compact, $\rho(G)$ is also compact.

Conversely, assume that $\overline{\rho(G)}$ is compact. In this case, $H = \overline{\rho(G)}$ is a closed subgroup of the Lie group $\text{GL}(V)$, so by Theorem 15.12, $H$ is a compact Lie subgroup of $\text{GL}(V)$. Now, the inclusion homomorphism $H \hookrightarrow \text{GL}(V)$ is a representation of $H$ ($f \cdot u = f(u)$, for all $f \in H$ and all $u \in V$), so by Theorem 17.4, there is an inner product on $V$ which is $H$-invariant. However, for any $g \in G$, if we write $f = \rho(g) \in H$, then we have

$$\langle g \cdot u, g \cdot v \rangle = \langle f(u), f(v) \rangle = \langle u, v \rangle,$$

proving that $\langle - , - \rangle$ is $G$-invariant as well. \(\square\)

Applying Theorem 17.5 to the adjoint representation $\text{Ad}: G \rightarrow \text{GL}(\mathfrak{g})$, we get our second criterion for the existence of a bi-invariant metric on a Lie group.

**Proposition 17.6.** Given any Lie group $G$, an inner product $\langle - , - \rangle$ on $\mathfrak{g}$ induces a bi-invariant metric on $G$ iff $\overline{\text{Ad}(G)}$ is compact. In particular, every compact Lie group has a bi-invariant metric.

**Proof.** Proposition 17.3 is equivalent to the fact that $G$ possesses a bi-invariant metric iff there is some $\text{Ad}$-invariant inner product on $\mathfrak{g}$. By Theorem 17.5, there is some $\text{Ad}$-invariant inner product on $\mathfrak{g}$ iff $\overline{\text{Ad}(G)}$ is compact, which is the statement of our theorem. \(\square\)

Proposition 17.6 can be used to prove that certain Lie groups do not have a bi-invariant metric. For example, Arsigny, Pennec and Ayache use Proposition 17.6 to give a short and elegant proof of the fact that $\text{SE}(n)$ does not have any bi-invariant metric for all $n \geq 2$. As noted by these authors, other proofs found in the literature are a lot more complicated and only cover the case $n = 3$.

Recall the adjoint representation of the Lie algebra $\mathfrak{g}$,

$$\text{ad}: \mathfrak{g} \rightarrow \text{gl}(\mathfrak{g}),$$

given by $\text{ad} = d\text{Ad}_1$. Here is our third criterion for the existence of a bi-invariant metric on a connected Lie group.
Proposition 17.7. If $G$ is a connected Lie group, an inner product $\langle - , - \rangle$ on $\mathfrak{g}$ induces a bi-invariant metric on $G$ iff the linear map $\text{ad}(u): \mathfrak{g} \rightarrow \mathfrak{g}$ is skew-adjoint for all $u \in \mathfrak{g}$, which means that

$$\langle \text{ad}(u)(v), w \rangle = -\langle v, \text{ad}(u)(w) \rangle, \text{ for all } u, v, w \in \mathfrak{g},$$

or equivalently that

$$\langle [v, u], w \rangle = \langle v, [u, w] \rangle, \text{ for all } u, v, w \in \mathfrak{g}.$$

Proof. We follow Milnor [128], Lemma 7.2. By Proposition 17.3 an inner product on $\mathfrak{g}$ induces a bi-invariant metric on $G$ iff $\text{Ad}_g$ is an isometry for all $g \in G$. Recall the notion of adjoint of a linear map. Given a linear map $f: V \rightarrow V$ on a vector space $V$ equipped with an inner product $\langle - , - \rangle$, we define $f^* : V \rightarrow V$ to be the unique linear map such that

$$\langle f(u), v \rangle = \langle u, f^*(v) \rangle, \text{ for all } u, v \in V,$$

and call $f^*$ the adjoint of $f$. It is a standard fact of linear algebra that $f$ is an isometry iff $f^{-1} = f^*$. Thus $\text{Ad}(g)$ is an isometry iff $\text{Ad}(g)^{-1} = \text{Ad}(g)^*$. The paragraph before Proposition 15.4 shows that we can choose a small enough open subset $U$ of $\mathfrak{g}$ containing 0 so that $\exp: \mathfrak{g} \rightarrow G$ is a diffeomorphism from $U$ to $\exp(U)$. For any $g \in \exp(U)$, there is a unique $u \in \mathfrak{g}$ so that $g = \exp(u)$. By Proposition 15.6,

$$\text{Ad}(g) = \text{Ad}(\exp(u)) = e^{\text{ad}(u)}.$$

Since $\text{Ad}(g)^{-1} = \text{Ad}(g)^*$, the preceding equation implies that

$$\text{Ad}(g)^{-1} = e^{-\text{ad}(u)} \text{ and } \text{Ad}(g)^* = e^{\text{ad}(u)},$$

so we deduce that $\text{Ad}(g)^{-1} = \text{Ad}(g)^*$ iff

$$\text{ad}(u)^* = -\text{ad}(u),$$

which means that $\text{ad}(u)$ is skew-adjoint. Since a connected Lie group is generated by any open subset containing the identity and since products of isometries are isometries, our results holds for all $g \in G$.

The skew-adjointness of $\text{ad}(u)$ means that

$$\langle \text{ad}(u)(v), w \rangle = -\langle v, \text{ad}(u)(w) \rangle \text{ for all } u, v, w \in \mathfrak{g},$$

and since $\text{ad}(u)(v) = [u, v]$ and $[u, v] = -[v, u]$, we get

$$\langle [v, u], w \rangle = \langle v, [u, w] \rangle$$

which is the last claim of the proposition. \qed

It will be convenient to say that an inner product on $\mathfrak{g}$ is bi-invariant iff every $\text{ad}(u)$ is skew-adjoint.

The following variant of Proposition 17.7 will also be needed. This is a special case of Lemma 3 in O'Neill [138] (Chapter 11).
Proposition 17.8. If $G$ is a Lie group equipped with an inner product $\langle -,- \rangle$ on $\mathfrak{g}$ that induces a bi-invariant metric on $G$, then $\text{ad}(X): \mathfrak{g}^L \to \mathfrak{g}^L$ is skew-adjoint for all left-invariant vector fields $X \in \mathfrak{g}^L$, which means that
\[ \langle \text{ad}(X)(Y), Z \rangle = -\langle Y, \text{ad}(X)(Z) \rangle, \quad \text{for all } X,Y,Z \in \mathfrak{g}^L, \]
or equivalently that
\[ \langle [Y,X], Z \rangle = \langle Y, [X,Z] \rangle, \quad \text{for all } X,Y,Z \in \mathfrak{g}^L. \]

Proof. By the bi-invariance of the metric, Proposition 17.3 implies that the inner product $\langle -,- \rangle$ on $\mathfrak{g}$ is Ad-invariant. For any two left-invariant vector fields $X,Y \in \mathfrak{g}^L$, we have
\[ \langle \text{Ad}_a X, \text{Ad}_a Y \rangle_e := \langle \text{Ad}_a X(e), \text{Ad}_a Y(e) \rangle = \langle X(e), Y(e) \rangle, \]
which shows that the function $a \mapsto \langle \text{Ad}_a X, \text{Ad}_a Y \rangle_e$ is constant. For any left-invariant vector field $Z$, by taking the derivative of this function with $a = \exp(tZ(e))$ at $t = 0$, we get
\[ \langle [Z(e),X(e)], Y(e) \rangle + \langle X(e), [Z(e),Y(e)] \rangle = 0. \]
Since $dL_g$ is a diffeomorphism for every $g \in G$, the metric on $G$ is assumed to be bi-invariant, and $X(g) = (dL_g)_e(X(e))$ for any left-invariant vector field $X$, we have
\[ \langle [Z(g),X(g)], Y(g) \rangle_g + \langle X(g), [Z(g),Y(g)] \rangle_g = \langle [[dL_g]_e(Z(e)), (dL_g)_e(X(e))], (dL_g)_e(Y(e)) \rangle_g + \langle (dL_g)_e(X(e)), [[dL_g]_e(Z(e)), (dL_g)_e(Y(e))] \rangle_g = \langle [Z(e),X(e)], Y(e) \rangle + \langle X(e), [Z(e),Y(e)] \rangle = 0. \]
Therefore,
\[ \langle [Z,X], Y \rangle + \langle X, [Z,Y] \rangle = 0, \]
which is equivalent to
\[ \langle [X,Z], Y \rangle = \langle X, [Z,Y] \rangle, \]
and to
\[ \langle \text{ad}(Z)(X), Y \rangle = -\langle X, \text{ad}(Z)(Y) \rangle. \]
If we apply the permutation $(X,Y,Z) \mapsto Y,Z,X$, we obtain our proposition. \qed

We now turn to our fourth criterion. If $G$ is a connected Lie group, then the existence of a bi-invariant metric on $G$ places a heavy restriction on its group structure, as shown by the following result from Milnor’s paper [128] (Lemma 7.5):

Theorem 17.9. A connected Lie group $G$ admits a bi-invariant metric iff it is isomorphic to the cartesian product of a compact group and a vector space $(\mathbb{R}^m, \text{for some } m \geq 0)$. 
A proof of Theorem 17.9 can be found in Milnor [128] (Lemma 7.4 and Lemma 7.5). The proof uses the universal covering group and it is a bit involved. We will outline the structure of the proof, because it is really quite beautiful.

First, recall from Definition 15.8 that a subset $h$ of a Lie algebra $g$ is a *Lie subalgebra* iff it is a subspace of $g$ (as a vector space) and if it is closed under the bracket operation on $g$. A subalgebra $h$ of $g$ is *abelian* iff $[x, y] = 0$ for all $x, y \in h$. An *ideal* in $g$ is a Lie subalgebra $h$ such that $[h, g] \subseteq h$, for all $h \in h$ and all $g \in g$.

**Definition 17.2.** A Lie algebra $g$ is *simple* iff it is non-abelian and if it has no ideal other than $(0)$ and $g$. A Lie group is *simple* iff its Lie algebra is simple.

In a first step for the proof of Theorem 17.9, it is shown that if $G$ has a bi-invariant metric, then its Lie algebra $g$ can be written as an orthogonal direct sum

$$ g = g_1 \oplus \cdots \oplus g_k, $$

where each $g_i$ is either a simple ideal or a one-dimensional abelian ideal; that is, $g_i \cong \mathbb{R}$.

The next step is to lift the ideals $g_i$ to simply connected normal subgroups $G_i$ of the universal covering group $\widetilde{G}$ of $G$. For every simple ideal $g_i$ in the decomposition, it is proved that there is some constant $c_i > 0$, so that all Ricci curvatures are strictly positive and bounded from below by $c_i$. Therefore, by Myers’ Theorem (Theorem 13.22), $G_i$ is compact. It follows that $\widetilde{G}$ is isomorphic to a product of compact simple Lie groups and some vector space, $\mathbb{R}^m$. Finally, we know that $G$ is isomorphic to the quotient of $\widetilde{G}$ by a discrete normal subgroup of $\widetilde{G}$, which yields our theorem.

Because it is a fun proof, we prove the statement about the structure of a Lie algebra for which each $\text{ad}(u)$ is skew-adjoint.

**Proposition 17.10.** Let $g$ be a Lie algebra with an inner product such that the linear map $\text{ad}(u)$ is skew-adjoint for every $u \in g$. Then the orthogonal complement $a^\perp$ of any ideal $a$ is itself an ideal. Consequently, $g$ can be expressed as an orthogonal direct sum

$$ g = g_1 \oplus \cdots \oplus g_k, $$

where each $g_i$ is either a simple ideal or a one-dimensional abelian ideal ($g_i \cong \mathbb{R}$).

**Proof.** Assume $u \in g$ is orthogonal to $a$. We need to prove that $[u, v]$ is orthogonal to $a$ for all $v \in g$. But, as $\text{ad}(u)$ is skew-adjoint, $\text{ad}(u)(v) = [u, v]$, and $a$ is an ideal, we have

$$ \langle [u, v], a \rangle = -\langle u, [v, a] \rangle = 0, \quad \text{for all } a \in a, $$

which shows that $a^\perp$ is an ideal.
For the second statement we use induction on the dimension of $g$, but for this proof, we redefine a simple Lie algebra to be an algebra with no nontrivial proper ideals. The case where $\dim g = 1$ is clear.

For the induction step, if $g$ is simple, we are done. Else, $g$ has some nontrivial proper ideal $h$, and if we pick $h$ of minimal dimension $p$, with $1 \leq p < n = \dim g$, then $h$ is simple. Now, $h^\perp$ is also an ideal and $\dim h^\perp < n$, so the induction hypothesis applies. Therefore, we have an orthogonal direct sum

$$g = g_1 \oplus \cdots \oplus g_k,$$

where each $g_i$ is simple in our relaxed sense. However, if $g_i$ is not abelian, then it is simple in the usual sense, and if $g_i$ is abelian, having no proper nontrivial ideal, it must be one-dimensional and we get our decomposition.

We now investigate connections and curvature on Lie groups with a left-invariant metric.

### 17.3 Connections and Curvature of Left-Invariant Metrics on Lie Groups

If $G$ is a Lie group equipped with a left-invariant metric, then it is possible to express the Levi-Civita connection and the sectional curvature in terms of quantities defined over the Lie algebra of $G$, at least for left-invariant vector fields. When the metric is bi-invariant, much nicer formulae are be obtained. In this section, we always assume that our Lie groups are equipped with the Levi-Civita connection.

If $\langle - , - \rangle$ is a left-invariant metric on $G$, then for any two left-invariant vector fields $X, Y$, we have

$$\langle X, Y \rangle_g = \langle X(g), Y(g) \rangle_g = \langle (dL_g)eX(e), (dL_g)eY(e) \rangle_g = \langle X_e, Y_e \rangle_e = \langle X, Y \rangle_e,$$

which shows that the function $g \mapsto \langle X, Y \rangle_g$ is constant. Therefore, for any vector field $Z$,

$$Z(\langle X, Y \rangle) = 0.$$

If we go back to the Koszul formula (Proposition 11.8)

$$2\langle \nabla_X Y, Z \rangle = X(\langle Y, Z \rangle) + Y(\langle X, Z \rangle) - Z(\langle X, Y \rangle) - \langle Y, [X, Z] \rangle - \langle X, [Y, Z] \rangle - \langle Z, [Y, X] \rangle,$$

we deduce that for all left-invariant vector fields $X, Y, Z \in \mathfrak{g}^L$, we have

$$2\langle \nabla_X Y, Z \rangle = -\langle Y, [X, Z] \rangle - \langle X, [Y, Z] \rangle - \langle Z, [Y, X] \rangle,$$

which can be rewritten as

$$2\langle \nabla_X Y, Z \rangle = \langle [X, Y], Z \rangle - \langle [Y, Z], X \rangle + \langle [Z, X], Y \rangle.$$

(†)
The above yields the formula

$$\nabla_X Y = \frac{1}{2} ([X, Y] - \text{ad}(X)^* Y - \text{ad}(Y)^* X), \quad X, Y \in g^L,$$

where $\text{ad}(X)^*$ denotes the adjoint of $\text{ad}(X)$, where $\text{ad}(X)$ is defined just after Proposition 15.7.

Given any two vector $u, v \in g$, it is common practice (even though this is quite confusing) to denote by $\nabla_u v$ the result of evaluating the vector field $\nabla_{u^L} v^L$ at $e$ (so, $\nabla_u v = (\nabla_{u^L} v^L)(e)$).

Following Milnor, if we pick an orthonormal basis $(e_1, \ldots, e_n)$ w.r.t. our inner product on $g$, and if we define the structure constants $\alpha_{ijk}$ by

$$\alpha_{ijk} = \langle [e_i, e_j], e_k \rangle,$$

we see that

$$\nabla_{e_i} e_j = \frac{1}{2} \sum_k (\alpha_{ijk} - \alpha_{jki} + \alpha_{kij}) e_k. \quad (*)$$

For example, let $G = \text{SO}(3)$, the group of $3 \times 3$ rotation matrices. Then $g = \text{so}(3)$ is the vector space of skew symmetric $3 \times 3$ matrices with orthonormal basis

$$e_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad e_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad e_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

since the left invariant, indeed bi-invariant, metric on $g$ is

$$\langle B_1, B_2 \rangle = \text{tr}(B_1^\top B_2) = -\text{tr}(B_1 B_2).$$

Matrix multiplication shows

$$[e_1, e_2] = e_1 e_2 - e_2 e_1 = \frac{1}{\sqrt{2}} e_3$$
$$[e_2, e_3] = e_2 e_3 - e_3 e_2 = \frac{1}{\sqrt{2}} e_1$$
$$[e_3, e_1] = e_3 e_1 - e_1 e_3 = \frac{1}{\sqrt{2}} e_2$$
$$[e_1, e_1] = [e_2, e_2] = [e_3, e_3] = 0.$$

Hence

\[-\alpha_{213} = \alpha_{123} = \langle [e_1, e_2], e_3 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_3 \rangle = \frac{1}{\sqrt{2}}\]

\[-\alpha_{211} = \alpha_{121} = \langle [e_1, e_2], e_1 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_1 \rangle = 0\]

\[-\alpha_{212} = \alpha_{122} = \langle [e_1, e_2], e_2 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_2 \rangle = 0\]

\[\alpha_{112} = \langle [e_1, e_1], e_2 \rangle = 0, \quad \alpha_{221} = \langle [e_2, e_2], e_1 \rangle = 0\]

\[-\alpha_{321} = \alpha_{231} = \langle [e_2, e_3], e_1 \rangle = \frac{1}{\sqrt{2}} \langle e_1, e_1 \rangle = \frac{1}{\sqrt{2}}\]

\[-\alpha_{131} = \alpha_{311} = \langle [e_3, e_1], e_1 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_3 \rangle = 0\]

\[-\alpha_{132} = \alpha_{312} = \langle [e_3, e_1], e_2 \rangle = \frac{1}{\sqrt{2}} \langle e_2, e_2 \rangle = \frac{1}{\sqrt{2}}\]

\[-\alpha_{322} = \alpha_{232} = \langle [e_2, e_3], e_2 \rangle = \frac{1}{\sqrt{2}} \langle e_1, e_2 \rangle = 0\]

\[-\alpha_{323} = \alpha_{233} = \langle [e_2, e_3], e_3 \rangle = \frac{1}{\sqrt{2}} \langle e_1, e_3 \rangle = 0\]

\[\alpha_{223} = \langle [e_2, e_2], e_3 \rangle = 0, \quad \alpha_{332} = \langle [e_3, e_3], e_2 \rangle = 0,\]

and

\[\nabla_{e_1} e_2 = -\nabla_{e_2} e_1 = \frac{1}{2} \sum_{k=1}^{3} (\alpha_{12k} - \alpha_{2k1} + \alpha_{k12}) e_k\]

\[= \frac{1}{2} \left( \alpha_{123} - \alpha_{231} + \alpha_{312} \right) = \frac{1}{2\sqrt{2}} e_3 = \frac{1}{2} \langle e_1, e_2 \rangle\]

\[\nabla_{e_1} e_3 = -\nabla_{e_3} e_1 = \frac{1}{2} \sum_{k=1}^{3} (\alpha_{13k} - \alpha_{3k1} + \alpha_{k13}) e_k\]

\[= \frac{1}{2} \left( \alpha_{132} - \alpha_{321} + \alpha_{213} \right) = -\frac{1}{2\sqrt{2}} e_2 = \frac{1}{2} \langle e_1, e_3 \rangle\]

\[\nabla_{e_2} e_3 = -\nabla_{e_3} e_2 = \frac{1}{2} \sum_{k=1}^{3} (\alpha_{23k} - \alpha_{3k2} + \alpha_{k23}) e_k\]

\[= \frac{1}{2} \left( \alpha_{213} - \alpha_{312} + \alpha_{123} \right) = -\frac{1}{2\sqrt{2}} e_1 = \frac{1}{2} \langle e_2, e_2 \rangle\]
17.3. CONNECTIONS AND CURVATURE OF LEFT-INARIANT METRICS

\[ \nabla_{e_1} e_1 = \frac{1}{2} \sum_{k=1}^{3} (\alpha_{11k} - \alpha_{1k1} + \alpha_{k11})e_k = 0 \]

\[ \nabla_{e_2} e_2 = \frac{1}{2} \sum_{k=1}^{3} (\alpha_{22k} - \alpha_{2k2} + \alpha_{k22})e_k = 0 \]

\[ \nabla_{e_3} e_3 = \frac{1}{2} \sum_{k=1}^{3} (\alpha_{33k} - \alpha_{3k3} + \alpha_{k33})e_k = 0. \]

Now for orthonormal vectors \( u, v \), the sectional curvature is given by

\[ K(u, v) = \langle R(u, v)u, v \rangle, \]

with

\[ R(u, v) = \nabla_{[u,v]} - \nabla_u \nabla_v + \nabla_v \nabla_u. \]

If we plug the expressions from equation (*) into the definitions we obtain the following proposition from Milnor [128] (Lemma 1.1):

**Proposition 17.11.** Given a Lie group \( G \) equipped with a left-invariant metric, for any orthonormal basis \( e_1, \ldots, e_n \) of \( g \), and with the structure constants \( \alpha_{ijk} = \langle [e_i, e_j], e_k \rangle \), the sectional curvature \( K(e_1, e_2) \) is given by

\[
K(e_i, e_j) = \sum_k \left( \frac{1}{2} \alpha_{ijk} (-\alpha_{ijk} + \alpha_{jki} + \alpha_{kij}) \right. \\
\left. - \frac{1}{4} (\alpha_{ijk} - \alpha_{jki} + \alpha_{kij})(\alpha_{ijk} + \alpha_{jki} - \alpha_{kij}) - \alpha_{kii}\alpha_{kj} \right).
\]

For \( \text{SO}(3) \), the formula of Proposition 17.11, when evaluated (via Maple) with the previously computed structure constants gives

\[
K(e_1, e_2) = \frac{1}{8} = \frac{1}{4} \langle e_3, e_3 \rangle = \frac{1}{4} \langle [e_1, e_2], [e_1, e_2] \rangle \\
K(e_1, e_3) = \frac{1}{8} = \frac{1}{4} \langle e_2, e_2 \rangle = \frac{1}{4} \langle [e_1, e_3], [e_1, e_3] \rangle \\
K(e_2, e_3) = \frac{1}{8} = \frac{1}{4} \langle e_1, e_1 \rangle = \frac{1}{4} \langle [e_2, e_3], [e_3, e_3] \rangle \\
K(e_1, e_1) = K(e_2, e_2) = K(e_3, e_3) = 0.
\]

Although the above formula is not too useful in general, in some cases of interest, a great deal of cancellation takes place so that a more useful formula can be obtained. An example of this situation is provided by the next proposition (Milnor [128], Lemma 1.2).
Proposition 17.12. Given a Lie group $G$ equipped with a left-invariant metric, for any $u \in \mathfrak{g}$, if the linear map $\text{ad}(u)$ is skew-adjoint, then

$$K(u, v) \geq 0 \quad \text{for all } v \in \mathfrak{g},$$

where equality holds iff $u$ is orthogonal to $[v, \mathfrak{g}] = \{[v, x] \mid x \in \mathfrak{g}\}$.

Proof. We may assume that $u$ and $v$ are orthonormal. If we pick an orthonormal basis such that $e_1 = u$ and $e_2 = v$, the fact that $\text{ad}(e_1)$ is skew-adjoint means that the array $(\alpha_{1jk})$ is skew-symmetric (in the indices $j$ and $k$). It follows that the formula of Proposition 17.11 reduces to

$$K(e_1, e_2) = \frac{1}{4} \sum_k \alpha_{2k1}^2,$$

so $K(e_1, e_2) \geq 0$, as claimed. Furthermore, $K(e_1, e_2) = 0$ iff $\alpha_{2k1} = 0$ for $k = 1, \ldots, n$; that is, $\langle [e_2, e_k], e_1 \rangle = 0$ for $k = 1, \ldots, n$, which means that $e_1$ is orthogonal to $[e_2, \mathfrak{g}]$. \square

For the next proposition we need the following definition:

Definition 17.3. The center $Z(\mathfrak{g})$ of a Lie algebra $\mathfrak{g}$ is the set of all elements $u \in \mathfrak{g}$ such that $[u, v] = 0$ for all $v \in \mathfrak{g}$, or equivalently, such that $\text{ad}(u) = 0$.

Proposition 17.13. Given a Lie group $G$ equipped with a left-invariant metric, for any $u$ in the center $Z(\mathfrak{g})$ of $\mathfrak{g}$,

$$K(u, v) \geq 0 \quad \text{for all } v \in \mathfrak{g}.$$

Proof. For any element $u$ in the center of $\mathfrak{g}$, we have $\text{ad}(u) = 0$, and the zero map is obviously skew-adjoint. \square

Recall that the Ricci curvature $\text{Ric}(u, v)$ is the trace of the linear map $y \mapsto R(u, y)v$. With respect to any orthonormal basis $(e_1, \ldots, e_n)$ of $\mathfrak{g}$, we have

$$\text{Ric}(u, v) = \sum_{j=1}^n \langle R(u, e_j)v, e_j \rangle = \sum_{j=1}^n R(u, e_j, v, e_j).$$

The Ricci curvature is a symmetric form, so it is completely determined by the quadratic form

$$r(u) = \text{Ric}(u, u) = \sum_{j=1}^n R(u, e_j, u, e_j).$$

When $u$ is a unit vector, $r(u)$ is called the Ricci curvature in the direction $u$. If we pick an orthonormal basis such that $e_1 = u$, then

$$r(e_1) = \sum_{i=2}^n K(e_1, e_i).$$
For computational purposes it may be more convenient to introduce the Ricci transformation $\hat{r}$, defined by

$$\hat{r}(x) = \sum_{i=1}^{n} R(e_i, x)e_i.$$ 

Observe that

$$\langle \hat{r}(x), y \rangle = \langle \sum_{i=1}^{n} R(e_i, x)e_i, y \rangle = \sum_{i=1}^{n} \langle R(e_i, x), y \rangle = \sum_{i=1}^{n} R(e_i, e_i, x),$$

by Prop. 13.2 (2)

$$= \sum_{i=1}^{n} \langle R(e_i, y)e_i, x \rangle = \langle \sum_{i=1}^{n} (R(e_i, y)e_i, x) = \langle x, \hat{r}(y) \rangle.$$ 

Hence, the Ricci transformation is self-adjoint, and it is also the unique map so that

$$r(x) = \langle \hat{r}(x), x \rangle, \quad \text{for all } x \in \mathfrak{g}.$$ 

The eigenvalues of $\hat{r}$ are called the principal Ricci curvatures.

**Proposition 17.14.** Given a Lie group $G$ equipped with a left-invariant metric, if the linear map $\text{ad}(u)$ is skew-adjoint, then $r(u) \geq 0$, where equality holds iff $u$ is orthogonal to the commutator ideal $[\mathfrak{g}, \mathfrak{g}]$.

**Proof.** This follows from Proposition 17.12. \qed

In particular, if $u$ is in the center of $\mathfrak{g}$, then $r(u) \geq 0$.

As a corollary of Proposition 17.14, we have the following result which is used in the proof of Theorem 17.9:

**Proposition 17.15.** If $G$ is a connected Lie group equipped with a bi-invariant metric and if the Lie algebra of $G$ is simple, then there is a constant $c > 0$ so that $r(u) \geq c$ for all unit vector $u \in T_gG$ and for all $g \in G$.

**Proof.** First of all, the linear maps $\text{ad}(u)$ are skew-adjoint for all $u \in \mathfrak{g}$, which implies that $r(u) \geq 0$. As $\mathfrak{g}$ is simple, the commutator ideal $[\mathfrak{g}, \mathfrak{g}]$ is either (0) or $\mathfrak{g}$. But, if $[\mathfrak{g}, \mathfrak{g}] = (0)$, then $\mathfrak{g}$ is abelian, which is impossible since $\mathfrak{g}$ is simple. Therefore $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$, which implies $r(u) > 0$ for all $u \neq 0$ (otherwise, $u$ would be orthogonal to $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$, which is impossible). As the set of unit vectors in $\mathfrak{g}$ is compact, the function $u \mapsto r(u)$ achieves it minimum $c$, and $c > 0$ as $r(u) > 0$ for all $u \neq 0$. But, $dL_g: \mathfrak{g} \to T_gG$ is an isometry for all $g \in G$, so $r(u) \geq c$ for all unit vectors $u \in T_gG$, and for all $g \in G$. \qed
By Myers’ Theorem (Theorem 13.22), the Lie group $G$ is compact and has a finite fundamental group.

The following interesting theorem is proved in Milnor (Milnor [128], Theorem 2.2):

**Theorem 17.16.** A connected Lie group $G$ admits a left-invariant metric with $r(u) > 0$ for all unit vectors $u \in \mathfrak{g}$ (all Ricci curvatures are strictly positive) iff $G$ is compact and has a finite fundamental group.

The following criterion for obtaining a direction of negative curvature is also proved in Milnor (Milnor [128], Lemma 2.3):

**Proposition 17.17.** Given a Lie group $G$ equipped with a left-invariant metric, if $u$ is orthogonal to the commutator ideal $[\mathfrak{g}, \mathfrak{g}]$, then $r(u) \leq 0$, where equality holds iff $\text{ad}(u)$ is self-adjoint.

When $G$ possesses a bi-invariant metric and $G$ is equipped with the Levi-Civita connection, the group exponential coincides with the exponential defined in terms of geodesics. Much nicer formulae are also obtained for the Levi-Civita connection and the curvatures.

First of all, since by Proposition 17.8,$$
\langle [Y, Z], X \rangle = \langle Y, [Z, X] \rangle,$$
the last two terms in equation (†), namely
$$2\langle \nabla_X Y, Z \rangle = \langle [X, Y], Z \rangle - \langle [Y, Z], X \rangle + \langle [Z, X], Y \rangle,$$
cancel out, and we get
$$\nabla_X Y = \frac{1}{2} [X, Y], \quad \text{for all } X, Y \in \mathfrak{g}^L.$$
Then since
$$R(u, v) = \nabla_{[u, v]} - \nabla_u \nabla_v + \nabla_v \nabla_u,$$
we get
$$R(u, v) = \frac{1}{2} \text{ad}([u, v]) - \frac{1}{4} \text{ad}(u)\text{ad}(v) + \frac{1}{4} \text{ad}(v)\text{ad}(u).$$
Using the Jacobi identity,
$$\text{ad}([u, v]) = \text{ad}(u)\text{ad}(v) - \text{ad}(v)\text{ad}(u),$$
we get
$$R(u, v) = \frac{1}{4} \text{ad}[u, v],$$
so
$$R(u, v)w = \frac{1}{4} [[u, v], w].$$
Hence, for unit orthogonal vectors \( u, v \), the sectional curvature \( K(u, v) = \langle R(u, v)u, v \rangle \) is given by
\[
K(u, v) = \frac{1}{4} \langle [[u, v], u], v \rangle,
\]
which (as \( \langle [x, y], z \rangle = \langle x, [y, z] \rangle \)) is rewritten as
\[
K(u, v) = \frac{1}{4} \langle [u, v], [u, v] \rangle.
\]

To compute the Ricci curvature \( \text{Ric}(u, v) \), we observe that \( \text{Ric}(u, v) \) is the trace of the linear map
\[
y \mapsto R(u, y)v = \frac{1}{4} [[u, y], v] = -\frac{1}{4} [v, [u, y]] = -\frac{1}{4} \text{ad}(v) \circ \text{ad}(u)(y).
\]
However, the bilinear form \( B \) on \( g \) given by
\[
B(u, v) = \text{tr} \left( \text{ad}(u) \circ \text{ad}(v) \right)
\]
is a famous object known as the Killing form of the Lie algebra \( g \). We will take a closer look at the Killing form shortly. For the time being, we observe that as \( \text{tr}(\text{ad}(u) \circ \text{ad}(v)) = \text{tr}(\text{ad}(v) \circ \text{ad}(u)) \), we get
\[
\text{Ric}(u, v) = -\frac{1}{4} B(u, v), \quad \text{for all } u, v \in g.
\]

We summarize all this in

**Proposition 17.18.** For any Lie group \( G \) equipped with a bi-invariant metric, the following properties hold:

(a) The Levi-Civita connection \( \nabla_X Y \) is given by
\[
\nabla_X Y = \frac{1}{2} [X, Y], \quad \text{for all } X, Y \in g^L.
\]

(b) The curvature tensor \( R(u, v) \) is given by
\[
R(u, v) = \frac{1}{4} \text{ad}[u, v], \quad \text{for all } u, v \in g,
\]
or equivalently,
\[
R(u, v)w = \frac{1}{4} [[u, v], w], \quad \text{for all } u, v, w \in g.
\]

(c) The sectional curvature \( K(u, v) \) is given by
\[
K(u, v) = \frac{1}{4} \langle [u, v], [u, v] \rangle,
\]
for all pairs of orthonormal vectors \( u, v \in g \).
(d) The Ricci curvature $\text{Ric}(u,v)$ is given by

$$\text{Ric}(u,v) = -\frac{1}{4} B(u,v),$$ for all $u,v \in \mathfrak{g},$

where $B$ is the Killing form, with

$$B(u,v) = \text{tr}(\text{ad}(u) \circ \text{ad}(v)), \text{ for all } u,v \in \mathfrak{g}.$$ 

Consequently, $K(u,v) \geq 0$, with equality iff $[u,v] = 0$ and $r(u) \geq 0$, with equality iff $u$ belongs to the center of $\mathfrak{g}$.

**Remark:** Proposition 17.18 shows that if a Lie group admits a bi-invariant metric, then its Killing form is negative semi-definite.

What are the geodesics in a Lie group equipped with a bi-invariant metric and the Levi-Civita connection? The answer is simple: they are the integral curves of left-invariant vector fields.

**Proposition 17.19.** For any Lie group $G$ equipped with a bi-invariant metric, we have:

1. The inversion map $\iota: g \mapsto g^{-1}$ is an isometry.
2. For every $a \in G$, if $I_a$ denotes the map given by
   $$I_a(b) = ab^{-1}a, \text{ for all } a,b \in G,$$

then $I_a$ is an isometry fixing $a$ which reverses geodesics; that is, for every geodesic $\gamma$ through $a$, we have
   $$I_a(\gamma)(t) = \gamma(-t).$$
3. The geodesics through $e$ are the integral curves $t \mapsto \exp_{gr}(tu)$, where $u \in \mathfrak{g}$; that is, the one-parameter groups. Consequently, the Lie group exponential map $\exp_{gr}: \mathfrak{g} \to G$ coincides with the Riemannian exponential map (at $e$) from $T_eG$ to $G$, where $G$ is viewed as a Riemannian manifold.

**Proof.** (1) Since

$$\iota(g) = g^{-1} = g^{-1}h^{-1}h = (hg)^{-1}h = (R_h \circ \iota \circ L_h)(g),$$

we have

$$\iota = R_h \circ \iota \circ L_h, \text{ for all } h \in G.$$ 

In particular, for $h = g^{-1}$, we get

$$d\iota_g = (dR_{g^{-1}})_e \circ dt_e \circ (dL_{g^{-1}})_g.$$
As $(dR_g^{-1})_e$ and $d(L_g^{-1})_e$ are isometries (since $G$ has a bi-invariant metric), $d\iota_g$ is an isometry iff $d\iota_e$ is. Thus, it remains to show that $d\iota_e$ is an isometry. However we can prove that $d\iota_e = -id$, so $d\iota_g$ is an isometry for all $g \in G$.

It remains to prove that $d\iota_e = -id$. This can be done in several ways. If we denote the multiplication of the group by $\mu: G \times G \to G$, then $T_e(G \times G) = T_e G \oplus T_e G = g \oplus g$, and it is easy to see that

$$d\mu(e,e)(u,v) = u + v, \quad \text{for all } u, v \in g.$$  

This is because $d\mu(e,e)$ is a homomorphism, and because $g \mapsto \mu(e,g)$ and $g \mapsto \mu(g,e)$ are the identity map. As the map $g \mapsto \mu(e,\iota(g))$ is the constant map with value $e$, by differentiating and using the chain rule, we get

$$d\iota_e(u) = -u,$$

as desired. (Another proof makes use of the fact that for every $u \in g$, the integral curve $\gamma$ through $e$ with $\gamma'(0) = u$ is a group homomorphism. Therefore,

$$\iota(\gamma(t)) = \gamma(t)^{-1} = \gamma(-t),$$

and by differentiating, we get $d\iota_e(u) = -u$.)

(2) We follow Milnor [125] (Lemma 21.1). From (1), the map $\iota$ is an isometry, so by Proposition 14.2 (3), it preserves geodesics through $e$. Since $d\iota_e$ reverses $T_e G = g$, it reverses geodesics through $e$. Observe that

$$I_a = R_a \circ \iota \circ R_{a^{-1}},$$

so by (1), $I_a$ is an isometry, and obviously $I_a(a) = a$. Again, by Proposition 14.2 (3), the isometry $I_a$ preserve geodesics, and since $R_a$ and $R_{a^{-1}}$ translate geodesics but $\iota$ reverses geodesics, it follows that $I_a$ reverses geodesics.

(3) We follow Milnor [125] (Lemma 21.2). Assume $\gamma$ is the unique geodesic through $e$ such that $\gamma'(0) = u$, and let $X = u^t$ be the left invariant vector field such that $X(e) = u$. The first step is to prove that $\gamma$ has domain $\mathbb{R}$ and that it is a group homomorphism; that is,

$$\gamma(s + t) = \gamma(s) \gamma(t).$$

Details of this argument are given in Milnor [125] (Lemma 20.1 and Lemma 21.2) and in Gallot, Hulin and Lafontaine [73] (Appendix B, Solution of Exercise 2.90). We present Milnor’s proof.

**Claim.** The isometries $I_a$ have the following property: For every geodesic $\omega$ through $a$, if we let $p = \omega(0)$ and $q = \omega(r)$, then

$$I_q \circ I_p(\omega(t)) = \omega(t + 2r),$$

whenever $\omega(t)$ and $\omega(t + 2r)$ are defined.
Let \( \alpha(t) = \omega(t + r) \). Then, \( \alpha \) is a geodesic with \( \alpha(0) = q \). As \( I_p \) reverses geodesics through \( p \) (and similarly for \( I_q \)), we get

\[
I_q \circ I_p(\omega(t)) = I_q(\omega(-t)) = I_q(\alpha(-t - r)) = \alpha(t + r) = \omega(t + 2r).
\]

It follows from the claim that \( \omega \) can be indefinitely extended; that is, the domain of \( \omega \) is \( \mathbb{R} \).

Next we prove that \( \gamma \) is a homomorphism. By the claim, \( I_{\gamma(t)} \circ I_e \) takes \( \gamma(u) \) into \( \gamma(u + 2t) \). Now by definition of \( I_a \) and \( I_e \),

\[
I_{\gamma(t)} \circ I_e(a) = \gamma(t)a_\gamma(t),
\]

so, with \( a = \gamma(u) \), we get

\[
\gamma(t)\gamma(u)\gamma(t) = \gamma(u + 2t).
\]

By induction, it follows that

\[
\gamma(nt) = \gamma(t)^n, \quad \text{for all } n \in \mathbb{Z}.
\]

We now use the (usual) trick of approximating every real by a rational number. For all \( r, s \in \mathbb{R} \) with \( s \neq 0 \), if \( r/s \) is rational, say \( r/s = m/n \) where \( m, n \) are integers, then \( r = mt \) and \( s = nt \) with \( t = r/m = s/n \) and we get

\[
\gamma(r + s) = \gamma(t)^{m+n} = \gamma(t)^m\gamma(t)^n = \gamma(r)\gamma(s).
\]

Given any \( t_1, t_2 \in \mathbb{R} \) with \( t_2 \neq 0 \), since \( t_1 \) and \( t_2 \) can be approximated by rationals \( r \) and \( s \), as \( r/s \) is rational, \( \gamma(r + s) = \gamma(r)\gamma(s) \), and by continuity, we get

\[
\gamma(t_1 + t_2) = \gamma(t_1)\gamma(t_2),
\]

as desired (the case \( t_2 = 0 \) is trivial as \( \gamma(0) = e \)).

As \( \gamma \) is a homomorphism, by differentiating the equation \( \gamma(s + t) = \gamma(s)\gamma(t) = L_{\gamma(s)}\gamma(t) \), we get

\[
\frac{d}{dt}(\gamma(s + t))|_{t=0} = (dL_{\gamma(s)})_e \left( \frac{d}{dt}(\gamma(t))|_{t=0} \right),
\]

that is

\[
\gamma'(s) = (dL_{\gamma(s)})_e(\gamma'(0)) = X(\gamma(s)),
\]

which means that \( \gamma \) is the integral curve of the left-invariant vector field \( X \), a one-parameter group.

Conversely, let \( c \) be the one-parameter group determined by a left-invariant vector field \( X = u^L \), with \( X(e) = u \) and let \( \gamma \) be the unique geodesic through \( e \) such that \( \gamma'(0) = u \). Since we have just shown that \( \gamma \) is a homomorphism with \( \gamma'(0) = u \), by uniqueness of one-parameter groups, \( c = \gamma \); that is, \( c \) is a geodesic. \( \Box \)
17.4 SIMPLE AND SEMISIMPLE LIE ALGEBRAS AND LIE GROUPS

Remarks:

(1) As \( R_g = \iota \circ L_g^{-1} \circ \iota \), we deduce that if \( G \) has a left-invariant metric, then this metric is also right-invariant if \( \iota \) is an isometry.

(2) Property (2) of Proposition 17.19 says that a Lie group with a bi-invariant metric is a symmetric space, an important class of Riemannian spaces invented and studied extensively by Élie Cartan. Symmetric spaces are briefly discussed in Section 19.8.

(3) The proof of 17.19 (3) given in O'Neill [138] (Chapter 11, equivalence of (5) and (6) in Proposition 9) appears to be missing the “hard direction,” namely, that a geodesic is a one-parameter group. Also, since left and right translations are isometries and since isometries map geodesics to geodesics, the geodesics through any point \( a \in G \) are the left (or right) translates of the geodesics through \( e \), and thus are expressed in terms of the group exponential. Therefore, the geodesics through \( a \in G \) are of the form

\[ \gamma(t) = L_a(\exp_{\mathfrak{g}}(tu)), \]

where \( u \in \mathfrak{g} \). Observe that \( \gamma'(0) = (dL_a)_e(u) \).

(4) Some of the other facts stated in Proposition 17.18 and Proposition 17.19 are equivalent to the fact that a left-invariant metric is also bi-invariant; see O'Neill [138] (Chapter 11, Proposition 9).

Many more interesting results about left-invariant metrics on Lie groups can be found in Milnor’s paper [128]. For example, flat left-invariant metrics on Lie a group are characterized (Theorem 1.5). We conclude this section by stating the following proposition (Milnor [128], Lemma 7.6):

**Proposition 17.20.** If \( G \) is any compact, simple, Lie group, then the bi-invariant metric is unique up to a constant. Such a metric necessarily has constant Ricci curvature.

17.4 Simple and Semisimple Lie Algebras and Lie Groups

In this section we introduce semisimple Lie algebras. They play a major role in the structure theory of Lie groups, but we only scratch the surface.

**Definition 17.4.** A Lie algebra \( \mathfrak{g} \) is simple iff it is non-abelian and if it has no ideal other than \((0)\) and \( \mathfrak{g} \). A Lie algebra \( \mathfrak{g} \) is semisimple iff it has no abelian ideal other than \((0)\). A Lie group is simple (resp. semisimple) iff its Lie algebra is simple (resp. semisimple).

Clearly, the trivial subalgebras \((0)\) and \( \mathfrak{g} \) itself are ideals, and the center of a Lie algebra is an abelian ideal. It follows that the center \( Z(\mathfrak{g}) \) of a semisimple Lie algebra must be the trivial ideal \((0)\).
Definition 17.5. Given two subsets \( a \) and \( b \) of a Lie algebra \( g \), we let \([a, b]\) be the subspace of \( g \) consisting of all linear combinations \([a, b]\) with \( a \in a \) and \( b \in b \).

If \( a \) and \( b \) are ideals in \( g \), then \( a + b \), \( a \cap b \), and \([a, b]\), are also ideals (for \([a, b]\), use the Jacobi identity). In particular, \([g, g]\) is an ideal in \( g \) called the \textit{commutator ideal} of \( g \). The commutator ideal \([g, g]\) is also denoted by \( D^1g \) (or \( Dg \)).

If \( g \) is a simple Lie algebra, then \([g, g] = g \) (because \([g, g]\) is an ideal, so the simplicity of \( g \) implies that either \([g, g] = (0) \) or \([g, g] = g \). However, if \([g, g] = (0) \), then \( g \) is abelian, a contradiction).

The \textit{derived series} (or \textit{commutator series}) \( (D^k g) \) of a Lie algebra (or ideal) \( g \) is defined as follows:

\[
D^0 g = g
\]
\[
D^{k+1} g = [D^k g, D^k g], \quad k \geq 0.
\]

The first three \( D^k g \) are

\[
D^0 g = g
\]
\[
D^1 g = [g, g]
\]
\[
D^2 g = [D^1 g, D^1 g].
\]

We have a decreasing sequence

\[
g = D^0 g \supseteq D^1 g \supseteq D^2 g \supseteq \cdots.
\]

Since \( g \) is an ideal, by induction we see that each \( D^k g \) is an ideal.

We say that \( g \) is \textit{solvable} iff \( D^k g = (0) \) for some \( k \geq 0 \). If \( g \) is abelian, then \([g, g] = 0 \), so \( g \) is solvable. Observe that a nonzero solvable Lie algebra has a nonzero abelian ideal, namely, the last nonzero \( D^j g \). As a consequence, a Lie algebra is semisimple iff it has no nonzero solvable ideal.

It can be shown that every Lie algebra \( g \) has a largest solvable ideal \( r \), called the \textit{radical} of \( g \) (see Knapp [106], Chapter I, Proposition 1.12). The radical of \( g \) is also denoted \( \text{rad} g \). Then a Lie algebra is semisimple iff \( \text{rad} g = (0) \).

The \textit{lower central series} \( (C^k g) \) of a Lie algebra (or ideal) \( g \) is defined as follows:

\[
C^0 g = g
\]
\[
C^{k+1} g = [g, C^k g], \quad k \geq 0.
\]

Since \( g \) is an ideal, by induction, each \( C^k g \) is an ideal. We have a decreasing sequence

\[
g = C^0 g \supseteq C^1 g \supseteq C^2 g \supseteq \cdots.
\]
We say that $\mathfrak{g}$ is nilpotent iff $C^k\mathfrak{g} = (0)$ for some $k \geq 0$. By induction, it is easy to show that
\[ D^k\mathfrak{g} \subseteq C^k\mathfrak{g} \quad k \geq 0. \]
Consequently, every nilpotent Lie algebra is solvable.

Note that, by definition, simple and semisimple Lie algebras are non-abelian, and a simple algebra is a semisimple algebra. It turns out that a Lie algebra $\mathfrak{g}$ is semisimple iff it can be expressed as a direct sum of ideals $\mathfrak{g}_i$, with each $\mathfrak{g}_i$ a simple algebra (see Knapp [106], Chapter I, Theorem 1.54). As a consequence if $\mathfrak{g}$ is semisimple, then we also have $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$. This is because if
\[ \mathfrak{g} = \bigoplus_{i=1}^{m} \mathfrak{g}_i \]
where each $\mathfrak{g}_i$ is a simple ideal, then
\[ [\mathfrak{g}, \mathfrak{g}] = \left[ \bigoplus_{i=1}^{m} \mathfrak{a}_i, \bigoplus_{j=1}^{m} \mathfrak{g}_j \right] = \bigoplus_{i,j=1}^{m} [\mathfrak{g}_i, \mathfrak{g}_j] = \bigoplus_{i=1}^{m} \mathfrak{g}_i = \mathfrak{g}, \]
since the $\mathfrak{g}_i$ being simple and forming a direct sum, $[\mathfrak{g}_i, \mathfrak{g}_j] = (0)$ whenever $i \neq j$ and $[\mathfrak{g}_i, \mathfrak{g}_i] = \mathfrak{g}_i$.

If we drop the requirement that a simple Lie algebra be non-abelian, thereby allowing one dimensional Lie algebras to be simple, we run into the trouble that a simple Lie algebra is no longer semisimple, and the above theorem fails for this stupid reason. Thus, it seems technically advantageous to require that simple Lie algebras be non-abelian.

Nevertheless, in certain situations, it is desirable to drop the requirement that a simple Lie algebra be non-abelian and this is what Milnor does in his paper because it is more convenient for one of his proofs. This is a minor point but it could be confusing for uninitiated readers.

### 17.5 The Killing Form

The Killing form showed the tip of its nose in Proposition 17.18. It is an important concept, and in this section we establish some of its main properties. First we recall its definition.

**Definition 17.6.** For any Lie algebra $\mathfrak{g}$ over the field $K$ (where $K = \mathbb{R}$ or $K = \mathbb{C}$), the \textit{Killing form} $B$ of $\mathfrak{g}$ is the symmetric $K$-bilinear form $B: \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}$ given by
\[ B(u, v) = \text{tr}(\text{ad}(u) \circ \text{ad}(v)), \quad \text{for all } u, v \in \mathfrak{g}. \]
If $\mathfrak{g}$ is the Lie algebra of a Lie group $G$, we also refer to $B$ as the \textit{Killing form of $G$}.

**Remark:** According to the experts (see Knapp [106], page 754) the \textit{Killing form} as above, was not defined by Killing, and is closer to a variant due to Élie Cartan. On the other hand, the notion of “Cartan matrix” is due to Wilhelm Killing!
For example, consider the group $\text{SU}(2)$. Its Lie algebra $\mathfrak{su}(2)$ is the three-dimensional Lie algebra consisting of all skew-Hermitian $2 \times 2$ matrices with zero trace; that is, matrices of the form

$$X = \begin{pmatrix} a & b + i c \\ -b + i c & -a \\ \end{pmatrix}, \quad a, b, c \in \mathbb{R}.$$ 

Let

$$Y = \begin{pmatrix} d & e + i f \\ -e + i f & -d \end{pmatrix}, \quad d, e, f \in \mathbb{R}.$$ 

By picking a suitable basis of $\mathfrak{su}(2)$, namely

$$e_1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad e_3 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$

it can be shown that

$$\text{ad}_X(e_1) = L_X(e_1) - R_X(e_1)$$

$$= \begin{pmatrix} a & b + i c \\ -b + i c & -a \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a & b + i c \\ -b + i c & -a \end{pmatrix}$$

$$= \begin{pmatrix} -2ic & 2ia \\ 2ia & 2ic \end{pmatrix} = -2ce_3 + 2ae_2$$

$$\text{ad}_X(e_2) = L_X(e_2) - R_X(e_2) = \begin{pmatrix} 2ib & -2a \\ 2a & -2ib \end{pmatrix} = -2ae_1 + 2be_3$$

$$\text{ad}_X(e_3) = L_X(e_3) - R_X(e_3) = \begin{pmatrix} 0 & 2c - 2ib \\ -2c - 2ib & 0 \end{pmatrix} = 2ce_1 - 2be_2,$$

which in turn implies that

$$\text{ad}_X = \begin{pmatrix} 0 & -2a & 2c \\ 2a & 0 & -2b \\ -2c & 2b & 0 \end{pmatrix}.$$ 

Similarly

$$\text{ad}_Y = \begin{pmatrix} 0 & -2d & 2f \\ 2d & 0 & -2e \\ -2f & 2e & 0 \end{pmatrix}.$$ 

Thus

$$B(X, Y) = \text{tr}(\text{ad}_X \circ \text{ad}_Y) = \text{tr} \begin{pmatrix} -4ad - 4cf & 4ce & 4ae \\ 4bf & -4ad - 4be & 4af \\ 4bd & 4cd & -4be - 4cf \end{pmatrix}$$

$$= -8ad - 8be - 8cf.$$
17.5. THE KILLING FORM

However

\[
\text{tr}(XY) = \text{tr}\left( \begin{pmatrix} -ad - cf - be + i(bf - ce) & -af + cd + i(ae - bd) \\ af - cd + i(ae - bd) & -ad - cf - be + i(-bf + ce) \end{pmatrix} \right) \\
= -2ad - 2be - 2cf.
\]

Hence

\[B(X, Y) = 4\text{tr}(XY).\]

Now, if we consider the group \(U(2)\), its Lie algebra \(u(2)\) is the four-dimensional Lie algebra consisting of all skew-Hermitian \(2 \times 2\) matrices; that is, matrices of the form

\[
\begin{pmatrix} ai & b + ic \\ -b + ic & id \end{pmatrix}, \quad a, b, c, d \in \mathbb{R},
\]

By using the basis

\[
e_1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad e_3 = \begin{pmatrix} i & 0 \\ 0 & 0 \end{pmatrix}, \quad e_4 = \begin{pmatrix} 0 & 0 \\ 0 & i \end{pmatrix},
\]

it can be shown that

\[B(X, Y) = 4\text{tr}(XY) - 2\text{tr}(X)\text{tr}(Y).\]

For \(SO(3)\), we know that \(so(3) \cong su(2)\), and we get

\[B(X, Y) = \text{tr}(XY).\]

Actually, it can be shown that

\[
\begin{align*}
\text{GL}(n, \mathbb{R}), U(n): & \quad B(X, Y) = 2n\text{tr}(XY) - 2\text{tr}(X)\text{tr}(Y) \\
\text{SL}(n, \mathbb{R}), SU(n): & \quad B(X, Y) = 2n\text{tr}(XY) \\
SO(n): & \quad B(X, Y) = (n - 2)\text{tr}(XY).
\end{align*}
\]

It suffices to compute the quadratic form \(B(X, X)\), because \(B(X, Y)\) is symmetric bilinear so it can be recovered using the polarization identity

\[B(X, Y) = \frac{1}{2}(B(X + Y, X + Y) - B(X, X) - B(Y, Y)).\]

Furthermore, if \(g\) is the Lie algebra of a matrix group, since \(\text{ad}_X = L_X - R_X\) and \(L_X\) and \(R_X\) commute, for all \(X, Z \in g\), we have

\[(\text{ad}_X \circ \text{ad}_X)(Z) = (L_X^2 - 2L_X \circ R_X + R_X^2)(Z) = X^2Z - 2XZX + ZX^2.\]

Therefore, to compute \(B(X, X) = \text{tr}(\text{ad}_X \circ \text{ad}_X)\), we can pick a convenient basis of \(g\) and compute the diagonal entries of the matrix representing the linear map

\[
Z \mapsto X^2Z - 2XZX + ZX^2.
\]
Unfortunately, this is usually quite laborious. Some of the computations can be found in Jost [99] (Chapter 5, Section 5.5) and in Helgason [88] (Chapter III, §8).

Recall that a homomorphism of Lie algebras $\varphi: \mathfrak{g} \to \mathfrak{h}$ is a linear map that preserves brackets; that is,

$$
\varphi([u, v]) = [\varphi(u), \varphi(v)].
$$

**Proposition 17.21.** The Killing form $B$ of a Lie algebra $\mathfrak{g}$ has the following properties:

1. It is a symmetric bilinear form invariant under all automorphisms of $\mathfrak{g}$. In particular, if $\mathfrak{g}$ is the Lie algebra of a Lie group $G$, then $B$ is $\text{Ad}_g$-invariant, for all $g \in G$.

2. The linear map $\text{ad}(u)$ is skew-adjoint w.r.t $B$ for all $u \in \mathfrak{g}$; that is,

$$
B(\text{ad}(u)(v), w) = -B(v, \text{ad}(u)(w)), \quad \text{for all } u, v, w \in \mathfrak{g},
$$

or equivalently,

$$
B([u, v], w) = B(u, [v, w]), \quad \text{for all } u, v, w \in \mathfrak{g}.
$$

**Proof.** (1) The form $B$ is clearly bilinear, and as $\text{tr}(AB) = \text{tr}(BA)$, it is symmetric. If $\varphi$ is an automorphism of $\mathfrak{g}$, the preservation of the bracket implies that

$$
\text{ad}(\varphi(u)) \circ \varphi = \varphi \circ \text{ad}(u),
$$

so

$$
\text{ad}(\varphi(u)) = \varphi \circ \text{ad}(u) \circ \varphi^{-1}.
$$

From $\text{tr}(XY) = \text{tr}(YX)$, we get $\text{tr}(A) = \text{tr}(BAB^{-1})$, so we get

$$
B(\varphi(u), \varphi(v)) = \text{tr}(\text{ad}(\varphi(u)) \circ \text{ad}(\varphi(v)))
$$

$$
= \text{tr}(\varphi \circ \text{ad}(u) \circ \varphi^{-1} \circ \varphi \circ \text{ad}(v) \circ \varphi^{-1})
$$

$$
= \text{tr}(\text{ad}(u) \circ \text{ad}(v)) = B(u, v).
$$

Since $\text{Ad}_g$ is an automorphism of $\mathfrak{g}$ for all $g \in G$, $B$ is $\text{Ad}_g$-invariant.

(2) We have

$$
B(\text{ad}(u)(v), w) = B([u, v], w) = \text{tr}(\text{ad}([u, v]) \circ \text{ad}(w))
$$

and

$$
B(v, \text{ad}(u)(w)) = B(v, [u, w]) = \text{tr}(\text{ad}(v) \circ \text{ad}([u, w])).
$$

However, the Jacobi identity is equivalent to

$$
\text{ad}([u, v]) = \text{ad}(u) \circ \text{ad}(v) - \text{ad}(v) \circ \text{ad}(u).
$$
Consequently,
\[
\text{tr}(\text{ad}([u, v]) \circ \text{ad}(w)) = \text{tr}((\text{ad}(u) \circ \text{ad}(v) - \text{ad}(v) \circ \text{ad}(u)) \circ \text{ad}(w)) = \text{tr}(\text{ad}(u) \circ \text{ad}(v) \circ \text{ad}(w)) - \text{tr}(\text{ad}(v) \circ \text{ad}(u) \circ \text{ad}(w))
\]
and
\[
\text{tr}(\text{ad}(v) \circ \text{ad}([u, w])) = \text{tr}(\text{ad}(v) \circ (\text{ad}(u) \circ \text{ad}(w) - \text{ad}(w) \circ \text{ad}(u))) = \text{tr}(\text{ad}(v) \circ \text{ad}(u) \circ \text{ad}(w)) - \text{tr}(\text{ad}(v) \circ \text{ad}(w) \circ \text{ad}(u)).
\]
As
\[
\text{tr}(\text{ad}(u) \circ \text{ad}(v) \circ \text{ad}(w)) = \text{tr}(\text{ad}(v) \circ \text{ad}(w) \circ \text{ad}(u)),
\]
we deduce that
\[
B(\text{ad}(u)(v), w) = \text{tr}(\text{ad}([u, v]) \circ \text{ad}(w)) = -\text{tr}(\text{ad}(v) \circ \text{ad}([u, w])) = -B(v, \text{ad}(u)(w)),
\]
as claimed. \(\Box\)

Remarkably, the Killing form yields a simple criterion due to Élie Cartan for testing whether a Lie algebra is semisimple. Recall that a bilinear form \(f : \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}\) is non-degenerate if and only if \(f(u, u) = 0\) implies that \(u = 0\).

**Theorem 17.22.** (Cartan’s Criterion for Semisimplicity) A lie algebra \(\mathfrak{g}\) is semisimple iff its Killing form \(B\) is non-degenerate.

As far as we know, all the known proofs of Cartan’s criterion are quite involved. A fairly easy going proof can be found in Knapp [106] (Chapter 1, Theorem 1.45). A more concise proof is given in Serre [159] (Chapter VI, Theorem 2.1). As a corollary of Theorem 17.22, we get:

**Proposition 17.23.** If \(G\) is a semisimple Lie group, then the center of its Lie algebra is trivial; that is, \(Z(\mathfrak{g}) = (0)\).

**Proof.** Since \(u \in Z(\mathfrak{g})\) iff \(\text{ad}(u) = 0\), we have
\[
B(u, u) = \text{tr}(\text{ad}(u) \circ \text{ad}(u)) = 0.
\]
As \(B\) is nondegenerate, we must have \(u = 0\). \(\Box\)

Since a Lie group with trivial Lie algebra is discrete, this implies that the center of a simple Lie group is discrete (because the Lie algebra of the center of a Lie group is the center of its Lie algebra. Prove it!).

We can also characterize which Lie groups have a Killing form which is negative definite.
Theorem 17.24. A connected Lie group is compact and semisimple iff its Killing form is negative definite.

Proof. First, assume that $G$ is compact and semisimple. Then, by Proposition 17.6, there is an inner product on $\mathfrak{g}$ inducing a bi-invariant metric on $G$, and by Proposition 17.7, every linear map $\text{ad}(u)$ is skew-adjoint. Therefore, if we pick an orthonormal basis of $\mathfrak{g}$, the matrix $X$ representing $\text{ad}(u)$ is skew-symmetric, and

$$B(u, u) = \text{tr}(\text{ad}(u) \circ \text{ad}(u)) = \text{tr}(XX) = \sum_{i,j=1}^{n} a_{ij}a_{ji} = -\sum_{i,j=1}^{n} a_{ij}^2 \leq 0.$$ 

Since $G$ is semisimple, $B$ is nondegenerate, and so, it is negative definite.

Now, assume that $B$ is negative definite. If so, $-B$ is an inner product on $\mathfrak{g}$, and by Proposition 17.21, it is $\text{Ad}$-invariant. By Proposition 17.3, the inner product $-B$ induces a bi-invariant metric on $G$, and by Proposition 17.18 (d), the Ricci curvature is given by

$$\text{Ric}(u, v) = -\frac{1}{4} B(u, v),$$

which shows that $r(u) > 0$ for all units vectors $u \in \mathfrak{g}$. As in the proof of Proposition 17.15, there is some constant $c > 0$, which is a lower bound on all Ricci curvatures $r(u)$, and by Myers' Theorem (Theorem 13.22), $G$ is compact (with finite fundamental group). By Cartan's criterion, as $B$ is non-degenerate, $G$ is also semisimple. \hfill $\square$

Remark: A compact semisimple Lie group equipped with $-B$ as a metric is an Einstein manifold, since Ric is proportional to the metric (see Definition 13.5).

By using Theorems 17.22 and 17.24, since the Killing forms for $\text{U}(n)$, $\text{SU}(n)$ and $\text{SO}(n)$ are given by

$$\text{GL}(n, \mathbb{R}), \text{U}(n): \quad B(X, Y) = 2n\text{tr}(XY) - 2\text{tr}(X)\text{tr}(Y)$$
$$\text{SL}(n, \mathbb{R}), \text{SU}(n): \quad B(X, Y) = 2n\text{tr}(XY)$$
$$\text{SO}(n): \quad B(X, Y) = (n-2)\text{tr}(XY),$$

we see that $\text{SU}(n)$ is compact and semisimple for $n \geq 2$, $\text{SO}(n)$ is compact and semisimple for $n \geq 3$, and $\text{SL}(n, \mathbb{R})$ is noncompact and semisimple for $n \geq 2$. However, $\text{U}(n)$, even though it is compact, is not semisimple.

Another way to determine whether a Lie algebra is semisimple is to consider reductive Lie algebras. We give a quick exposition without proofs. Details can be found in Knapp [106] (Chapter I, Sections, 7, 8).

Definition 17.7. A Lie algebra $\mathfrak{g}$ is reductive iff for every ideal $\mathfrak{a}$ in $\mathfrak{g}$, there is some ideal $\mathfrak{b}$ in $\mathfrak{g}$ such that $\mathfrak{g}$ is the direct sum

$$\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b}.$$
If $\mathfrak{g}$ is semisimple, we can pick $\mathfrak{b} = \mathfrak{a}^\perp$, the orthogonal complement of $\mathfrak{a}$ with respect to the Killing form of $\mathfrak{g}$. Therefore, every semisimple Lie algebra is reductive. More generally, if $\mathfrak{g}$ is the direct sum of a semisimple Lie algebra and an abelian Lie algebra, then $\mathfrak{g}$ is reductive. In fact, there are no other reductive Lie algebra. The following result is proved in Knapp [106] (Chapter I, Corollary 1.56).

**Proposition 17.25.** If $\mathfrak{g}$ is a reductive Lie algebra, then

$$\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}] \oplus Z(\mathfrak{g}),$$

with $[\mathfrak{g}, \mathfrak{g}]$ semisimple and $Z(\mathfrak{g})$ abelian.

Consequently, if $\mathfrak{g}$ is reductive, then it is semisimple iff its center $Z(\mathfrak{g})$ is trivial. For Lie algebras of matrices, a simple condition implies that a Lie algebra is reductive. The following result is proved in Knapp [106] (Chapter I, Proposition 1.59).

**Proposition 17.26.** If $\mathfrak{g}$ is a real Lie algebra of matrices over $\mathbb{R}$ or $\mathbb{C}$, and if $\mathfrak{g}$ is closed under conjugate transpose (that is, if $A \in \mathfrak{g}$, then $A^* \in \mathfrak{g}$), then $\mathfrak{g}$ is reductive.

The familiar Lie algebras $\mathfrak{gl}(n, \mathbb{R})$, $\mathfrak{sl}(n, \mathbb{R})$, $\mathfrak{gl}(n, \mathbb{C})$, $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{so}(n)$, $\mathfrak{so}(n, \mathbb{C})$, $\mathfrak{u}(n)$, $\mathfrak{su}(n)$, $\mathfrak{so}(p, q)$, $\mathfrak{u}(p, q)$, $\mathfrak{su}(p, q)$ are all closed under conjugate transpose. Among those, by computing their center, we find that $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{sl}(n, \mathbb{C})$ are semisimple for $n \geq 2$, $\mathfrak{so}(n)$, $\mathfrak{so}(n, \mathbb{C})$ are semisimple for $n \geq 3$, $\mathfrak{su}(n)$ is semisimple for $n \geq 2$, $\mathfrak{so}(p, q)$ is semisimple for $p + q \geq 3$, and $\mathfrak{su}(p, q)$ is semisimple for $p + q \geq 2$.

Semisimple Lie algebras and semisimple Lie groups have been investigated extensively, starting with the complete classification of the complex semisimple Lie algebras by Killing (1888) and corrected by Élie Cartan in his thesis (1894). One should read the Notes, especially on Chapter II, at the end of Knapp’s book [106] for a fascinating account of the history of the theory of semisimple Lie algebras.

The theories and the body of results that emerged from these classification investigations play a very important role, not only in mathematics, but also in physics, and constitute one of the most beautiful chapters of mathematics. A quick introduction to these theories can be found in Arvanitoyeorgos [11] and in Carter, Segal, Macdonald [38]. A more comprehensive but yet still introductory presentation is given in Hall [84]. The most comprehensive treatment is probably Knapp [106]. An older is classic is Helgason [88], which also discusses differential geometric aspects of Lie groups. Other “advanced” presentations can be found in Bröcker and tom Dieck [31], Serre [160, 159], Samelson [151], Humphreys [96] and Kirillov [102]. A fascinating account of the history of Lie groups and Lie algebras is found in Armand Borel [23].
17.6 Left-Invariant Connections and Cartan Connections

Unfortunately, if a Lie group $G$ does not admit a bi-invariant metric, under the Levi-Civita connection, geodesics are generally not given by the Lie group exponential map $\exp_g: g \to G$. If we are willing to consider connections not induced by a metric, then it turns out that there is a fairly natural connection for which the geodesics coincide with integral curves of left-invariant vector fields. These connections are called Cartan connections. Such connections are torsion-free (symmetric), but the price that we pay is that in general they are not compatible with the chosen metric. As a consequence, even though geodesics exist for all $t \in \mathbb{R}$, Hopf–Rinow’s Theorem fails; worse, it is generally false that any two points can be connected by a geodesic. This has to do with the failure of the exponential to be surjective. This section is heavily inspired by Postnikov [144] (Chapter 6, Sections 3–6); see also Kobayashi and Nomizu [107] (Chapter X, Section 2).

Recall that a vector field $X$ on a Lie group $G$ is left-invariant if the following diagram commutes for all $a \in G$:

$$
\begin{array}{c}
TG \\
\downarrow d(L_a) \\
G
\end{array}
\quad
\begin{array}{c}
TG \\
\downarrow X \\
G
\end{array}
$$

In this section, we use freely the fact that there is an isomorphism between the Lie algebra $g$ and the Lie algebra $gl$ of left-invariant vector fields on $G$. For every $X \in g$, we denote by $X^L \in gl$ the unique left-invariant vector field such that $X^L_1 = X$.

**Definition 17.8.** A connection $\nabla$ on a Lie group $G$ is left-invariant if for any two left-invariant vector fields $X^L, Y^L$ with $X, Y \in g$, the vector field $\nabla_{X^L} Y^L$ is also left-invariant.

By analogy with left-invariant metrics, there is a version of Proposition 17.1 stating that there is a one-to-one correspondence between left-invariant connections and bilinear maps $\alpha: g \times g \to g$. This is shown as follows.

Given a left-invariant connection $\nabla$ on $G$, we get the map $\alpha: g \times g \to g$ given by

$$
\alpha(X, Y) = (\nabla_{X^L} Y^L)_1, \quad X, Y \in g.
$$

To define a map in the opposite direction, pick any basis $X_1, \ldots, X_n$ of $g$. Then every vector field $X$ on $G$ can be written as

$$
X = f_1 X^L_1 + \cdots + f_n X^L_n,
$$

for some smooth functions $f_1, \ldots, f_n$ on $G$. If $\nabla$ is a left-invariant connection on $G$, for any
left-invariant vector fields $X = \sum_{i=1}^{n} f_i X_i^L$ and $Y = \sum_{j=1}^{n} g_j X_j^L$, we have
\[
\nabla_X Y = \nabla_{\sum_{i=1}^{n} f_i X_i^L} Y = \sum_{i=1}^{n} f_i \nabla_{X_i^L} Y \\
= \sum_{i=1}^{n} f_i \nabla_{X_i^L} \sum_{i=1}^{n} g_j X_j^L \\
= \sum_{i,j=1}^{n} f_i ((X_i^L g_j) X_j^L + g_j \nabla_{X_i^L} X_j^L).
\]
This shows that $\nabla$ is completely determined by the matrix with entries
\[
\alpha_{ij} = \alpha(X_i, X_j) = (\nabla_{X_i^L} X_j^L)_1.
\]
Conversely, any bilinear map $\alpha$ on $\mathfrak{g}$ is determined by the matrix $(\alpha_{ij})$ with $\alpha_{ij} = \alpha(X_i, X_j) \in \mathfrak{g}$, and it is immediately checked that formula ($\dagger$) shown below
\[
\nabla_X Y = \sum_{i,j=1}^{n} f_i ((X_i^L g_j) X_j^L + g_j \alpha_{ij}),
\]
defines a left-invariant connection such that $(\nabla_{X_i^L} X_j^L)_1 = \alpha_{ij}$ for $i, j = 1, \ldots, n$. In summary, we proved the following result.

**Proposition 17.27.** There is a one-to-one correspondence between left-invariant connections on $G$ and bilinear maps $\alpha : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$.

Let us now investigate the conditions under which the geodesic curves coincide with the integral curves of left-invariant vector fields. Let $X^L$ be any left-invariant vector field, and let $\gamma$ be the integral curve such that $\gamma(0) = 1$ and $\gamma'(0) = X$ (in other words, $\gamma(t) = \exp_{\mathfrak{g}_t}(tX) = e^{tX}$). Since the vector field $t \mapsto \gamma'(t)$ along $\gamma$ is the restriction of the vector field $X^L$, we have
\[
D/dt(\gamma'(t)) = (\nabla_{X^L} X^L)_{\gamma(t)} = \alpha(X, X^L)_{\gamma(t)}, \quad \text{for all } t \in \mathbb{R}.
\]
Since a left-invariant vector field is determined by its value at 1, and $\gamma$ is a geodesic iff $D\gamma'/dt = 0$, we have $(\nabla_{X^L} X^L)_{\gamma(t)} = 0$ for all $t \in \mathbb{R}$ iff
\[
\alpha(X, X) = 0.
\]
Every bilinear map $\alpha$ can be written as the sum of a symmetric bilinear map
\[
\alpha_H(X, Y) = \frac{\alpha(X, Y) + \alpha(Y, X)}{2}
\]
and a skew-symmetric bilinear map

$$\alpha_S(X, Y) = \frac{\alpha(X, Y) - \alpha(Y, X)}{2},$$

Clearly $\alpha_S(X, X) = 0$. Thus $\alpha(X, X) = 0$ implies that $\alpha_H(X, X) = 0$. Hence we conclude that for every $X \in \mathfrak{g}$, the curve $t \mapsto e^{tX}$ is a geodesic iff $\alpha$ is skew-symmetric.

**Proposition 17.28.** The left-invariant connection $\nabla$ induced by a bilinear map $\alpha$ on $\mathfrak{g}$ has the property that, for every $X \in \mathfrak{g}$, the curve $t \mapsto \exp_{\mathfrak{g}}(tX) = e^{tX}$ is a geodesic iff $\alpha$ is skew-symmetric.

A left-invariant connection satisfying the property that for every $X \in \mathfrak{g}$, the curve $t \mapsto e^{tX}$ is a geodesic, is called a Cartan connection.

Let us find out when the Cartan connection $\nabla$ associated with a bilinear map $\alpha$ on $\mathfrak{g}$ is torsion-free (symmetric). We must have

$$\nabla_X Y^L - \nabla_Y X^L = [X, Y]^L,$$

that is,

$$\alpha(X, Y) - \alpha(Y, X) = [X, Y],$$

so we deduce that the Cartan connection induced by $\alpha$ is torsion-free iff

$$\alpha_S(X, Y) = \frac{1}{2} [X, Y],$$

for all $X, Y \in \mathfrak{g}$.

In view of the fact that the connection induced by $\alpha$ is torsion-free iff

$$\alpha_S(X, Y) = \frac{1}{2} [X, Y],$$

for all $X, Y \in \mathfrak{g}$,

we have the following fact:

**Proposition 17.29.** Given any Lie group $G$, there is a unique torsion-free (symmetric) Cartan connection $\nabla$ given by

$$\nabla_X Y^L = \frac{1}{2} [X, Y]^L,$$

for all $X, Y \in \mathfrak{g}$.

Then the same calculation that we used in the case of a bi-invariant metric on a Lie group shows that the curvature tensor is given by

$$R(X, Y)Z = \frac{1}{4} [[X, Y], Z],$$

for all $X, Y, Z \in \mathfrak{g}$.

It is easy to check that for any $X \in \mathfrak{g}$ and any point $a \in G$, the unique geodesic $\gamma_{a,X}$ such that $\gamma_{a,X}(0) = a$ and $\gamma_{a,X}'(0) = X$, is given by

$$\gamma_{a,X}(t) = e^{td(R(a^{-1}))aX}a;$$
that is,
\[ \gamma_{a,X} = R_a \circ \gamma_{d(R_a^{-1})aX}, \]
where \( \gamma_{d(R_a^{-1})aX}(t) = e^{td(R_a^{-1})aX}. \)

**Remark:** Observe that the bilinear maps given by
\[ \alpha(X, Y) = \lambda [X, Y] \quad \text{for some } \lambda \in \mathbb{R} \]
are skew-symmetric, and thus induce Cartan connections. Easy computations show that the torsion is given by
\[ T(X, Y) = (2\lambda - 1)[X, Y], \]
and the curvature by
\[ R(X, Y)Z = \lambda(\lambda - 1)[[X, Y], Z]. \]
It follows that for \( \lambda = 0 \) and \( \lambda = 1 \), we get connections where the curvature vanishes. However, these connections have torsion. Again, we see that \( \lambda = 1/2 \) is the only value for which the Cartan connection is symmetric.

In the case of a bi-invariant metric, the Levi-Civita connection coincides with the Cartan connection.
CHAPTER 17. METRICS, CONNECTIONS, AND CURVATURE ON LIE GROUPS
Chapter 18

The Log-Euclidean Framework Applied to SPD Matrices

18.1 Introduction

In this chapter, we present an application of Lie groups and Riemannian geometry. We describe an approach due to Arsigny, Fillard, Pennec and Ayache, to define a Lie group structure and a class of metrics on symmetric, positive-definite matrices (SPD matrices) which yield a new notion of mean on SPD matrices generalizing the standard notion of geometric mean.

SPD matrices are used in diffusion tensor magnetic resonance imaging (for short, DTI), and they are also a basic tool in numerical analysis, for example, in the generation of meshes to solve partial differential equations more efficiently.

As a consequence, there is a growing need to interpolate or to perform statistics on SPD matrices, such as computing the mean of a finite number of SPD matrices.

Recall that the set of $n \times n$ SPD matrices is not a vector space (because if $A \in \text{SPD}(n)$, then $\lambda A \notin \text{SPD}(n)$ if $\lambda < 0$), but it is a convex cone. Thus, the arithmetic mean of $n$ SPD matrices $S_1, \ldots, S_n$ can be defined as $(S_1 + \cdots + S_n)/n$, which is SPD. However, there are many situations, especially in DTI, where this mean is not adequate. There are essentially two problems:

1. The arithmetic mean is not invariant under inversion, which means that if $S = (S_1 + \cdots + S_n)/n$, then in general $S^{-1} \neq (S_1^{-1} + \cdots + S_n^{-1})/n$.

2. The swelling effect: the determinant $\det(S)$ of the mean $S$ may be strictly larger than the original determinants $\det(S_i)$. This effect is undesirable in DTI because it amounts to introducing more diffusion, which is physically unacceptable.

To circumvent these difficulties, various metrics on SPD matrices have been proposed. One class of metrics is the affine-invariant metrics (see Arsigny, Pennec and Ayache [9]).
The swelling effect disappears and the new mean is invariant under inversion, but computing this new mean has a high computational cost, and in general, there is no closed-form formula for this new kind of mean.

Arsigny, Fillard, Pennec and Ayache [8] have defined a new family of metrics on \( \text{SPD}(n) \) named Log-Euclidean metrics, and have also defined a novel structure of Lie group on \( \text{SPD}(n) \) which yields a notion of mean that has the same advantages as the affine mean but is a lot cheaper to compute. Furthermore, this new mean, called Log-Euclidean mean, is given by a simple closed-form formula. We will refer to this approach as the Log-Euclidean Framework.

The key point behind the Log-Euclidean Framework is the fact that the exponential map \( \exp: \mathbb{S}(n) \to \text{SPD}(n) \) is a bijection, where \( \mathbb{S}(n) \) is the space of \( n \times n \) symmetric matrices; see Proposition 1.8. Consequently, the exponential map has a well-defined inverse, the logarithm \( \log: \text{SPD}(n) \to \mathbb{S}(n) \).

But more is true. It turns out that \( \exp: \mathbb{S}(n) \to \text{SPD}(n) \) is a diffeomorphism, a fact stated as Theorem 2.8 in Arsigny, Fillard, Pennec and Ayache [8].

Since \( \exp \) is a bijection, the above result follows from the fact that \( \exp \) is a local diffeomorphism on \( \mathbb{S}(n) \), because \( d\exp_S \) is non-singular for all \( S \in \mathbb{S}(n) \). In Arsigny, Fillard, Pennec and Ayache [8], it is proved that the non-singularity of \( d\exp_I \) near 0, which is well-known, “propagates” to the whole of \( \mathbb{S}(n) \).

Actually, the non-singularity of \( d\exp \) on \( \mathbb{S}(n) \) is a consequence of a more general result stated in Theorem 2.27.

With this preparation, we are ready to present the natural Lie group structure on \( \text{SPD}(n) \) introduced by Arsigny, Fillard, Pennec and Ayache [8] (see also Arsigny’s thesis [6]).

### 18.2 A Lie-Group Structure on \( \text{SPD}(n) \)

Using the diffeomorphism \( \exp: \mathbb{S}(n) \to \text{SPD}(n) \) and its inverse \( \log: \text{SPD}(n) \to \mathbb{S}(n) \), an abelian group structure can be defined on \( \text{SPD}(n) \) as follows:

**Definition 18.1.** For any two matrices \( S_1, S_2 \in \text{SPD}(n) \), define the *logarithmic product* \( S_1 \odot S_2 \) by
\[
S_1 \odot S_2 = \exp(\log(S_1) + \log(S_2)).
\]

Obviously, the multiplication operation \( \odot \) is commutative. The following proposition is shown in Arsigny, Fillard, Pennec and Ayache [8] (Proposition 3.2):

**Proposition 18.1.** The set \( \text{SPD}(n) \) with the binary operation \( \odot \) is an abelian group with identity \( I \), and with inverse operation the usual inverse of matrices. Whenever \( S_1 \) and \( S_2 \) commute, then \( S_1 \odot S_2 = S_1 S_2 \) (the usual multiplication of matrices).
18.3. LOG-EUCLIDEAN METRICS ON SPD(N)

For the last statement, we need to show that if \(S_1, S_2 \in \text{SPD}(n)\) commute, then \(S_1S_2\) is also in \(\text{SPD}(n)\), and that \(\log(S_1)\) and \(\log(S_2)\) commute, which follows from the fact that if two diagonalizable matrices commute, then they can be diagonalized over the same basis of eigenvectors.

Actually, \((\text{SPD}(n), \odot, I)\) is an abelian Lie group isomorphic to the vector space (also an abelian Lie group!) \(\text{S}(n)\), as shown in Arsigny, Fillard, Pennec and Ayache [8] (Theorem 3.3 and Proposition 3.4):

**Theorem 18.2.** The abelian group \((\text{SPD}(n), \odot, I)\) is a Lie group isomorphic to its Lie algebra \(\text{spd}(n) = \text{S}(n)\). In particular, the Lie group exponential in \(\text{SPD}(n)\) is identical to the usual exponential on \(\text{S}(n)\).

We now investigate bi-invariant metrics on the Lie group, \(\text{SPD}(n)\).

### 18.3 Log-Euclidean Metrics on SPD(n)

In general, a Lie group does not admit a bi-invariant metric, but an abelian Lie group always does because \(\text{Ad}_g = \text{id} \in \text{GL}(g)\) for all \(g \in G\), and so the adjoint representation \(\text{Ad}: G \to \text{GL}(g)\) is trivial (that is, \(\text{Ad}(G) = \{\text{id}\}\)), and then the existence of bi-invariant metrics is a consequence of Proposition 17.3.

Then, given any inner product \(\langle -, - \rangle\) on \(g\), the induced bi-invariant metric on \(G\) is given by

\[
\langle u, v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle.
\]

Now, the geodesics on a Lie group equipped with a bi-invariant metric are the left (or right) translates of the geodesics through \(e\), and the geodesics through \(e\) are given by the group exponential, as stated in Proposition 17.19 (3).

Let us apply Proposition 17.19 to the abelian Lie group \(\text{SPD}(n)\) and its Lie algebra \(\text{spd}(n) = \text{S}(n)\). Let \(\langle -, - \rangle\) be any inner product on \(\text{S}(n)\) and let \(\langle -, - \rangle_{\text{S}}\) be the induced bi-invariant metric on \(\text{SPD}(n)\). We find that the geodesics through \(S \in \text{SPD}(n)\) are of the form

\[
\gamma(t) = S \odot e^{tV},
\]

where \(V \in \text{S}(n)\). But \(S = e^{\log S}\), so

\[
S \odot e^{tV} = e^{\log S} \odot e^{tV} = e^{\log S + tV},
\]

so every geodesic through \(S\) is of the form

\[
\gamma(t) = e^{\log S + tV} = \exp_{\text{gr}}(\log S + tV).
\]

To avoid confusion between the exponential and the logarithm as Lie group maps and as Riemannian manifold maps, we will denote the former by \(\exp\) (instead of \(\exp_{\text{gr}}\) and \(\log\) (instead of \(\log_{\text{gr}}\)), and their Riemannian counterparts by \(\text{Exp}\) and \(\text{Log}\).
We are going to show that \( \text{Exp}, \text{Log}, \) the bi-invariant metric on \( \text{SPD}(n) \), and the distance \( d(S,T) \) between two matrices \( S,T \in \text{SPD}(n) \) can be expressed in terms of \( \text{exp} and \text{log} \).

We begin with \( \text{Exp} \). Note that
\[
\gamma'(0) = \text{dexp}_{\text{log} S} (V),
\]
and since the exponential map of \( \text{SPD}(n) \), as a Riemannian manifold, is given by
\[
\text{Exp}_S(U) = \gamma_U(1) \quad U \in \text{S}(n),
\]
where \( \gamma_U \) is the unique geodesic such that \( \gamma_U(0) = S \) and \( \gamma_U'(0) = U \), we must have
\[
\text{dexp}_{\text{log} S} (V) = U,
\]
so
\[
V = (\text{dexp}_{\text{log} S})^{-1}(U) \quad \text{and} \quad \text{Exp}_S(U) = e^{\log S + V} = e^{\log S + (\text{dexp}_{\text{log} S})^{-1}(U)}.
\]
However, \( \text{exp} \circ \text{log} = \text{id} \), so by differentiation, we get
\[
(\text{dexp}_{\text{log} S})^{-1}(U) = \text{dlog}_S(U),
\]
which yields
\[
\text{Exp}_S(U) = e^{\log S + \text{dlog}_S(U)}.
\]
To get a formula for \( \text{Log}_S T \) with \( T \in \text{SPD}(n) \), we solve the equation \( T = \text{Exp}_S(U) \) with respect to \( U \), that is
\[
e^{\log S + (\text{dexp}_{\text{log} S})^{-1}(U)} = T,
\]
which yields
\[
\log S + (\text{dexp}_{\text{log} S})^{-1}(U) = \log T,
\]
so
\[
U = \text{dexp}_{\text{log} S}(\log T - \log S). \quad \text{Therefore,}
\]
\[
\text{Log}_S T = \text{dexp}_{\text{log} S}(\log T - \log S).
\]
Finally, we can find an explicit formula for the Riemannian metric,
\[
\langle U, V \rangle_S = \langle \text{d}(L_{S^{-1}})_S(U), \text{d}(L_{S^{-1}})_S(V) \rangle,
\]
because \( \text{d}(L_{S^{-1}})_S = \text{dlog}_S \), which can be shown as follows: Observe that
\[
(\log \circ L_{S^{-1}})(T) = \log(S^{-1} \circ T) = \log(\exp(\log(S^{-1}) + \log(T))) = \log S^{-1} + \log T,
\]
so \( \text{d}(\log \circ L_{S^{-1}})_T = \text{dlog}_T \) (because \( S \) is held fixed), that is
\[
\text{dlog}_{S^{-1} \circ T} \circ (L_{S^{-1}})_T = \text{dlog}_T,
\]
which, for \( T = S \), yields \( \text{d}(L_{S^{-1}})_S = \text{dlog}_S \) since \( \text{dlog}_I = I \). Therefore,
\[
\langle U, V \rangle_S = \langle \text{dlog}_S(U), \text{dlog}_S(V) \rangle.
Now, a Lie group with a bi-invariant metric is complete, so given any two matrices $S, T \in \text{SPD}(n)$, their distance is the length of the geodesic segment $\gamma_V$ such that $\gamma_V(0) = S$ and $\gamma_V(1) = T$, namely $\|V\|_S = \sqrt{\langle S, S \rangle_S}$, the norm given by the Riemannian metric. But $V = \text{Log}_S T$, so that $$d(S, T) = \|\text{Log}_S T\|_S.$$ Using the equation $$\text{Log}_S T = d \exp_{\log S} (\log T - \log S)$$ and the fact that $d \log \circ d \exp = \text{id}$, we get $$d(S, T) = \|\log T - \log S\|,$$ where $\| \|$ is the norm corresponding to the inner product on $\text{spd}(n) = \mathbf{S}(n)$. Since $\langle - , - \rangle$ is a bi-invariant metric on $\text{SPD}(n)$, and since $$\langle U, V \rangle_S = \langle d \log_S(U), d \log_S(V) \rangle,$$ we see that the map $\exp: \mathbf{S}(n) \to \text{SPD}(n)$ is an isometry (since $d \exp \circ d \log = \text{id}$).

In summary, we have proved Corollary 3.9 of Arsigny, Fillard, Pennec and Ayache [8]:

**Theorem 18.3.** For any inner product $\langle - , - \rangle$ on $\mathbf{S}(n)$, if we give the Lie group $\text{SPD}(n)$ the bi-invariant metric induced by $\langle - , - \rangle$, then the following properties hold:

1. For any $S \in \text{SPD}(n)$, the geodesics through $S$ are of the form $$\gamma(t) = e^{\log S + tV}, \quad V \in \mathbf{S}(n).$$

2. The exponential and logarithm associated with the bi-invariant metric on $\text{SPD}(n)$ are given by
   $$\text{Exp}_S(U) = e^{\log S + d \log_S(U)},$$
   $$\text{Log}_S(T) = d \exp_{\log S} (\log T - \log S),$$
   for all $S, T \in \text{SPD}(n)$ and all $U \in \mathbf{S}(n)$.

3. The bi-invariant metric on $\text{SPD}(n)$ is given by
   $$\langle U, V \rangle_S = \langle d \log_S(U), d \log_S(V) \rangle,$$
   for all $U, V \in \mathbf{S}(n)$ and all $S \in \text{SPD}(n)$, and the distance $d(S, T)$ between any two matrices $S, T \in \text{SPD}(n)$ is given by
   $$d(S, T) = \|\log T - \log S\|,$$
   where $\| \|$ is the norm corresponding to the inner product on $\text{spd}(n) = \mathbf{S}(n)$. 
(4) The map \( \exp: \mathbb{S}(n) \to \mathbb{SPD}(n) \) is an isometry.

In view of Theorem 18.3 part (3), bi-invariant metrics on the Lie group \( \mathbb{SPD}(n) \) are called *Log-Euclidean metrics*. Since \( \exp: \mathbb{S}(n) \to \mathbb{SPD}(n) \) is an isometry and \( \mathbb{S}(n) \) is a vector space, the Riemannian Lie group \( \mathbb{SPD}(n) \) is a complete, simply-connected, and flat manifold (the sectional curvature is zero at every point); that is, a flat *Hadamard manifold* (see Sakai [150], Chapter V, Section 4).

Although, in general, Log-Euclidean metrics are not invariant under the action of arbitrary invertible matrices, they are invariant under similarity transformations (an isometry composed with a scaling). Recall that \( \text{GL}(n) \) acts on \( \mathbb{SPD}(n) \) via

\[
A \cdot S = ASA^\top,
\]

for all \( A \in \text{GL}(n) \) and all \( S \in \mathbb{SPD}(n) \). We say that a Log-Euclidean metric is *invariant under* \( A \in \text{GL}(n) \) iff

\[
d(A \cdot S, A \cdot T) = d(S, T),
\]

for all \( S, T \in \mathbb{SPD}(n) \). The following result is proved in Arsigny, Fillard, Pennec and Ayache [8] (Proposition 3.11):

**Proposition 18.4.** There exist metrics on \( \mathbb{S}(n) \) that are invariant under all similarity transformations, for example the metric \( \langle S, T \rangle = \text{tr}(ST) \).

### 18.4 A Vector Space Structure on \( \mathbb{SPD}(n) \)

The vector space structure on \( \mathbb{S}(n) \) can also be transferred onto \( \mathbb{SPD}(n) \).

**Definition 18.2.** For any matrix \( S \in \mathbb{SPD}(n) \), for any scalar \( \lambda \in \mathbb{R} \), define the scalar multiplication \( \lambda \odot S \) by

\[
\lambda \odot S = \exp(\lambda \log(S)).
\]

It is easy to check that \( (\mathbb{SPD}(n), \odot, \otimes) \) is a vector space with addition \( \odot \) and scalar multiplication \( \otimes \). By construction, the map \( \exp: \mathbb{S}(n) \to \mathbb{SPD}(n) \) is a linear isomorphism. What happens is that the vector space structure on \( \mathbb{S}(n) \) is transferred onto \( \mathbb{SPD}(n) \) via the log and exp maps.

### 18.5 Log-Euclidean Means

One of the major advantages of Log-Euclidean metrics is that they yield a computationally inexpensive notion of mean with many desirable properties. If \( (x_1, \ldots, x_n) \) is a list of \( n \) data
18.5. LOG-EUCLIDEAN MEANS

points in $\mathbb{R}^m$, then it is a simple exercise to see that the mean $\bar{x} = (x_1 + \cdots + x_n)/n$ is the unique minimum of the map

$$x \mapsto \sum_{i=1}^n d(x, x_i)^2,$$

where $d_2$ is the Euclidean distance on $\mathbb{R}^m$. We can think of the quantity

$$\sum_{i=1}^n d(x, x_i)^2$$

as the dispersion of the data.

More generally, if $(X, d)$ is a metric space, for any $\alpha > 0$ and any positive weights $w_1, \ldots, w_n$, with $\sum_{i=1}^n w_i = 1$, we can consider the problem of minimizing the function

$$x \mapsto \sum_{i=1}^n w_i d(x, x_i)^\alpha.$$

The case $\alpha = 2$ corresponds to a generalization of the notion of mean in a vector space and was investigated by Fréchet. In this case, any minimizer of the above function is known as a Fréchet mean. Fréchet means are not unique, but if $X$ is a complete Riemannian manifold, certain sufficient conditions on the dispersion of the data are known that ensure the existence and uniqueness of the Fréchet mean (see Pennec [139]). The case $\alpha = 1$ corresponds to a generalization of the notion of median. When the weights are all equal, the points that minimize the map

$$x \mapsto \sum_{i=1}^n d(x, x_i)$$

are called Steiner points. On a Hadamard manifold, Steiner points can be characterized (see Sakai [150], Chapter V, Section 4, Proposition 4.9).

In the case where $X = \text{SPD}(n)$ and $d$ is a Log-Euclidean metric, it turns out that the Fréchet mean is unique and is given by a simple closed-form formula. We have the following theorem from Arsigny, Fillard, Pennec and Ayache [8] (Theorem 3.13):

**Theorem 18.5.** Given $N$ matrices $S_1, \ldots, S_N \in \text{SPD}(n)$, their Log-Euclidean Fréchet mean exists and is uniquely determined by the formula

$$E_{\text{LE}}(S_1, \ldots, S_N) = \exp \left( \frac{1}{N} \sum_{i=1}^N \log(S_i) \right).$$

Furthermore, the Log-Euclidean mean is similarity-invariant, invariant by group multiplication and inversion and exponential-invariant.
Similarity-invariance means that for any similarity $A$, 
\[ \mathbb{E}_{LE}(AS_1A^\top, \ldots, AS_NA^\top) = A\mathbb{E}_{LE}(S_1, \ldots, S_N)A^\top, \]
and similarly for the other types of invariance.

Observe that the Log-Euclidean mean is a generalization of the notion of geometric mean. Indeed, if $x_1, \ldots, x_n$ are $n$ positive numbers, then their geometric mean is given by 
\[ \mathbb{E}_{geom}(x_1, \ldots, x_n) = (x_1 \cdots x_n)^{\frac{1}{n}} = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \log(x_i)\right). \]

The Log-Euclidean mean also has a good behavior with respect to determinants. The following theorem is proved in Arsigny, Fillard, Pennec and Ayache [8] (Theorem 4.2):

**Theorem 18.6.** Given $N$ matrices $S_1, \ldots, S_N \in \text{SPD}(n)$, we have 
\[ \det(\mathbb{E}_{LE}(S_1, \ldots, S_N)) = \mathbb{E}_{geom}(\det(S_1), \ldots, \det(S_N)). \]

**Remark:** The last line of the proof in Arsigny, Fillard, Pennec and Ayache [8] seems incorrect.

Arsigny, Fillard, Pennec and Ayache [8] also compare the Log-Euclidean mean with the affine mean. We highly recommend the above paper as well as Arsigny’s thesis [6] for further details.
Chapter 19

Manifolds Arising from Group Actions

This chapter provides the culmination of the theory presented in the previous seventeen chapters, the concept of a homogenous naturally reductive space.

We saw in Chapter 5 that many topological spaces arise from a group action. The scenario is that we have a smooth action $\varphi: G \times M \to M$ of a Lie group $G$ acting on a manifold $M$. If $G$ acts transitively on $M$, then for any point $x \in M$, if $G_x$ is the stabilizer of $x$, then Theorem 5.14 ensures that $M$ is homeomorphic to $G/G_x$. For simplicity of notation, write $H = G_x$. What we would really like is that $G/H$ actually be a manifold. This is indeed the case, because the transitive action of $G$ on $G/H$ is equivalent to a right action of $H$ on $G$ which is no longer transitive, but which has some special properties (to be proper and free).

We are thus led to considering left (and right) actions $\varphi: G \times M \to M$ of a Lie group $G$ on a manifold $M$ that are not necessarily transitive. If the action is not transitive, then we consider the orbit space $M/G$ of orbits $G \cdot x$ ($x \in M$). However, in general, $M/G$ is not even Hausdorff. It is thus desirable to look for sufficient conditions that ensure that $M/G$ is Hausdorff. A sufficient condition can be given using the notion of a proper map. If our action is also free, then the orbit space $M/G$ is indeed a smooth manifold. These results are presented in Sections 19.1 and 19.2; see Theorem 19.8 and its Corollary Theorem 19.9.

Sharper results hold if we consider Riemannian manifolds. Given a Riemannian manifold $N$ and a Lie group $G$ acting on $N$, Theorem 19.11 gives us a method for obtaining a Riemannian manifold $N/G$ such that $\pi: N \to N/G$ is a Riemannian submersion (when $\cdot: G \times N \to N$ is a free and proper action and $G$ acts by isometries). Theorem 19.15 gives us a method for obtaining a Riemannian manifold $N/G$ such that $\pi: N \to N/G$ is a Riemannian covering (when $\cdot: G \times N \to N$ is a free and proper action of a discrete group $G$ acting by isometries).

In the rest of this chapter, we consider the situation where our Lie group $G$ acts transitively on a manifold $M$. In this case, we know that $M$ is diffeomorphic to $G/H$, where $H$ is the stabilizer of any given point in $M$. Our goal is to endow $G/H$ with Riemannian
metrics that arise from inner products on the Lie algebra $\mathfrak{g}$, in a way that is reminiscent of the way in which left-invariant metrics on a Lie group are in one-to-one correspondence with inner products on $\mathfrak{g}$ (see Proposition 17.1). Our goal is realized by the class of reductive homogeneous spaces, which is the object of much of the following sections.

The first step is to consider $G$-invariant metrics on $G/H$. For any $g \in G$, let $\tau_g : G/H \to G/H$ be the diffeomorphism given by

$$\tau_g(g_2H) = gg_2H.$$ 

The $\tau_g$ are left-multiplications on cosets. A metric on $G/H$ is said to be $G$-invariant iff the $\tau_g$ are isometries of $G/H$. The existence of $G$-invariant metrics on $G/H$ depends on properties of a certain representation of $H$ called the isotropy representation (see Proposition 19.17). We will also need to express the derivative $d\pi_1 : \mathfrak{g} \to T_o(G/H)$ of the natural projection $\pi : G \to G/H$ (where $o$ is the point of $G/H$ corresponding to the coset $H$). This can be done in terms of the Lie group exponential $\exp_{gr} : \mathfrak{g} \to G$ (see Definition 15.6). Then, it turns out that $\text{Ker}(d\pi_1) = \mathfrak{h}$, the Lie algebra of $\mathfrak{h}$, and $d\pi_1$ factors through $\mathfrak{g}/\mathfrak{h}$ and yields an isomorphism between $\mathfrak{g}/\mathfrak{h}$ and $T_o(G/H)$.

In general, it is difficult to deal with the quotient $\mathfrak{g}/\mathfrak{h}$, and this suggests considering the situation where $\mathfrak{g}$ splits as a direct sum

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}.$$ 

In this case, $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$, and $d\pi_1$ restricts to an isomorphism between $\mathfrak{m}$ and $T_o(G/H)$. This isomorphism can be used to transport an inner product on $\mathfrak{m}$ to an inner product on $T_o(G/H)$. It is remarkable that a simple condition on $\mathfrak{m}$, namely $\text{Ad}(H)$ invariance, yields a one-to-one correspondence between $G$-invariant metrics on $G/H$ and $\text{Ad}(H)$-invariant inner products on $\mathfrak{m}$ (see Proposition 19.18). This is a generalization of the situation of Proposition 17.3 characterizing the existence of bi-invariant metrics on Lie groups. All this is built into the definition of a reductive homogeneous space given by Definition 19.6.

It is possible to express the Levi-Civita connection on a reductive homogeneous space in terms of the Lie bracket on $\mathfrak{g}$, but in general this formula is not very useful. A simplification of this formula is obtained if a certain condition holds. The corresponding spaces are said to be naturally reductive; see Definition 19.7. A naturally reductive space has the ”nice” property that its geodesics at $o$ are given by applying the coset exponential map to $\mathfrak{m}$; see Proposition 19.20. As we will see from the explicit examples provided in Section 19.7, naturally reductive spaces ”behave” just as nicely as their Lie group counterpart $G$, and the coset exponential of $\mathfrak{m}$ will provide all the necessary geometric information.

### 19.1 Proper Maps

We saw in Chapter 5 that many manifolds arise from a group action. The scenario is that we have a smooth action $\varphi : G \times M \to M$ of a Lie group $G$ acting on a manifold $M$ (recall
that an action \( \varphi \) is smooth if it is a smooth map). If \( G \) acts transitively on \( M \), then for any point \( x \in M \), if \( G_x \) is the stabilizer of \( x \), then we know from Proposition 28.21 that \( G/G_x \) is diffeomorphic to \( M \) and that the projection \( \pi: G \to G/G_x \) is a submersion.

If the action is not transitive, then we consider the orbit space \( M/G \) of orbits \( G \cdot x \). However, in general, \( M/G \) is not even Hausdorff. It is thus desirable to look for sufficient conditions that ensure that \( M/G \) is Hausdorff. A sufficient condition can be given using the notion of a proper map.

Before we go any further, let us observe that the case where our action is transitive is subsumed by the more general situation of an orbit space. Indeed, if our action is transitive, for any \( x \in M \), we know that the stabilizer \( H = G_x \) of \( x \) is a closed subgroup of \( G \). Then we can consider the right action \( G \times H \to G \) of \( H \) on \( G \) given by

\[
g \cdot h = gh, \quad g \in G, h \in H.
\]

The orbits of this (right) action are precisely the left cosets \( gH \) of \( H \). Therefore, the set of left cosets \( G/H \) (the homogeneous space induced by the action \( \cdot: G \times M \to M \)) is the set of orbits of the right action \( G \times H \to G \).

Observe that we have a transitive left action of \( G \) on the space \( G/H \) of left cosets, given by

\[
g_1 \cdot g_2 H = g_1 g_2 H.
\]

The stabilizer of \( 1H \) is obviously \( H \) itself. Thus, we recover the original transitive left action of \( G \) on \( M = G/H \).

Now, it turns out that a right action of the form \( G \times H \to G \), where \( H \) is a closed subgroup of a Lie group \( G \), is a special case of a free and proper right action \( M \times G \to M \), in which case the orbit space \( M/G \) is a manifold, and the projection \( \pi: G \to M/G \) is a submersion.

Let us now define proper maps.

**Definition 19.1.** If \( X \) and \( Y \) are two Hausdorff topological spaces,\(^1\) a function a \( \varphi: X \to Y \) is proper iff it is continuous and for every topological space \( Z \), the map \( \varphi \times \text{id}: X \times Z \to Y \times Z \) is a closed map (recall that \( f \) is a closed map iff the image of any closed set by \( f \) is a closed set).

If we let \( Z \) be a one-point space, we see that a proper map is closed. The following proposition is easy to prove (see Bourbaki, General Topology [29], Chapter 1, Section 10).

**Proposition 19.1.** If \( \varphi: X \to Y \) is any proper map, then for any closed subset \( F \) of \( X \), the restriction of \( \varphi \) to \( F \) is proper.

\(^1\)It is not necessary to assume that \( X \) and \( Y \) are Hausdorff but, if \( X \) and/or \( Y \) are not Hausdorff, we have to replace “compact” by “quasi-compact.” We have no need for this extra generality.
The following result can be shown (see Bourbaki, General Topology [29], Chapter 1, Section 10):

**Proposition 19.2.** A continuous map \( \varphi: X \to Y \) is proper iff \( \varphi \) is closed and if \( \varphi^{-1}(y) \) is compact for every \( y \in Y \).

If \( \varphi \) is proper, it is easy to show that \( \varphi^{-1}(K) \) is compact in \( X \) whenever \( K \) is compact in \( Y \). Moreover, if \( Y \) is also locally compact, then we have the following result (see Bourbaki, General Topology [29], Chapter 1, Section 10).

**Proposition 19.3.** If \( Y \) is locally compact, a continuous map \( \varphi: X \to Y \) is a proper map iff \( \varphi^{-1}(K) \) is compact in \( X \) whenever \( K \) is compact in \( Y \).

In particular, this is true if \( Y \) is a manifold since manifolds are locally compact. This explains why Lee [117] (Chapter 9) takes the property stated in Proposition 19.3 as the definition of a proper map (because he only deals with manifolds).

Finally we can define proper actions.

**Remark:** It is remarkable that a great deal of material discussed in this chapter, especially in Sections 19.4–19.9, can be found in Volume IV of Dieudonné’s classical treatise on Analysis [56]. However, it is spread over 400 pages, which does not make it easy to read.

### 19.2 Proper and Free Actions

**Definition 19.2.** Given a Hausdorff topological group \( G \) and a topological space \( M \), a left action \( \cdot: G \times M \to M \) is proper if it is continuous and if the map

\[
\theta: G \times M \longrightarrow M \times M, \quad (g, x) \mapsto (g \cdot x, x)
\]

is proper.

The right actions associated with the transitive actions presented in Section 5.2 are examples of proper actions.

If \( H \) is a closed subgroup of \( G \) and if \( \cdot: G \times M \to M \) is a proper action, then the restriction of this action to \( H \) is also proper (by Proposition 19.1, because \( H \times M \) is closed in \( G \times M \)). If we let \( M = G \), then \( G \) acts on itself by left translation, and the map \( \theta: G \times G \to G \times G \) given by \( \theta(g, x) = (gx, x) \) is a homeomorphism, so it is proper. It follows that the action \( \cdot: H \times G \to G \) of a closed subgroup \( H \) of \( G \) on \( G \) (given by \( (h, g) \mapsto hg \)) is proper. The same is true for the right action of \( H \) on \( G \).

As desired, proper actions yield Hausdorff orbit spaces.

---

2However, Duistermaat and Kolk [64] seem to have overlooked the fact that a condition on \( Y \) (such as local compactness) is needed in their remark on lines 5-6, page 53, just before Lemma 1.11.3.
Proposition 19.4. If the action $\cdot : G \times M \to M$ is proper (where $G$ is Hausdorff), then the orbit space $M/G$ is Hausdorff. Furthermore, $M$ is also Hausdorff.

Proof. If the action is proper, then the orbit equivalence relation is closed since it is the image of $G \times M$ in $M \times M$, and so $M/G$ is Hausdorff. The second part is left as an exercise. \qed

We also have the following properties (see Bourbaki, General Topology [29], Chapter 3, Section 4).

Proposition 19.5. Let $\cdot : G \times M \to M$ be a proper action, with $G$ Hausdorff. For any $x \in M$, let $G \cdot x$ be the orbit of $x$ and let $G_x$ be the stabilizer of $x$. Then:

(a) The map $g \mapsto g \cdot x$ is a proper map from $G$ to $M$.

(b) $G_x$ is compact.

(c) The canonical map from $G/G_x$ to $G \cdot x$ is a homeomorphism.

(d) The orbit $G \cdot x$ is closed in $M$.

If $G$ is locally compact, we have the following characterization of being proper (see Bourbaki, General Topology [29], Chapter 3, Section 4).

Proposition 19.6. If $G$ and $M$ are Hausdorff and $G$ is locally compact, then the action $\cdot : G \times M \to M$ is proper iff for all $x, y \in M$, there exist some open sets, $V_x$ and $V_y$ in $M$, with $x \in V_x$ and $y \in V_y$, so that the closure $\overline{K}$ of the set $K = \{g \in G \mid g \cdot V_x \cap V_y \neq \emptyset\}$, is compact in $G$.

In particular, if $G$ has the discrete topology, the above condition holds iff the sets $\{g \in G \mid g \cdot V_x \cap V_y \neq \emptyset\}$ are finite. Also, if $G$ is compact, then $\overline{K}$ is automatically compact, so every compact group acts properly.

If $M$ is locally compact, we have the following characterization of being proper (see Bourbaki, General Topology [29], Chapter 3, Section 4).

Proposition 19.7. Let $\cdot : G \times M \to M$ be a continuous action, with $G$ and $M$ Hausdorff. For any compact subset $K$ of $M$ we have:

(a) The set $G_K = \{g \in G \mid g \cdot K \cap K \neq \emptyset\}$ is closed.

(b) If $M$ is locally compact, then the action is proper iff $G_K$ is compact for every compact subset $K$ of $M$.

In the special case where $G$ is discrete (and $M$ is locally compact), condition (b) says that the action is proper iff $G_K$ is finite. We use this criterion to show that the action $\cdot : \mathbb{Z} \times \mathbb{R} \to \mathbb{R}$ given by $n \cdot x = 2^n x$ is not proper. Note that $\mathbb{R}$ is locally compact. Take $K = \{0, 1\}$, a set which is clearly compact in $\mathbb{R}$. Then $n \cdot K = \{0, 2^n\}$ and $G_K = \mathbb{Z}$, which
is not compact or finite in \( \mathbb{R} \). Intuitively, proper actions on manifolds involve translations, rotations, and constrained expansions. The action \( n \cdot x = 2^n x \) provides too much dilation on \( \mathbb{R} \) to be a proper action.

**Remark:** If \( G \) is a Hausdorff topological group and if \( H \) is a subgroup of \( G \), then it can be shown that the action of \( G \) on \( G/H \) \( ((g_1, g_2 H) \mapsto g_1 g_2 H) \) is proper iff \( H \) is compact in \( G \).

**Definition 19.3.** An action \( \cdot : G \times M \to M \) is **free** if for all \( g \in G \) and all \( x \in M \), if \( g \neq 1 \) then \( g \cdot x \neq x \).

An equivalent way to state that an action \( \cdot : G \times M \to M \) is free is as follows. For every \( g \in G \), let \( \tau_g : M \to M \) be the diffeomorphism of \( M \) given by

\[
\tau_g(x) = g \cdot x, \quad x \in M.
\]

Then the action \( \cdot : G \times M \to M \) is free iff for all \( g \in G \), if \( g \neq 1 \) then \( \tau_g \) has no fixed point. Another equivalent statement is that for every \( x \in M \), the stabilizer \( G_x \) of \( x \) is reduced to the trivial group \( \{1\} \). For example, the action of \( \text{SO}(3) \) on \( S^2 \) given by Example 5.2 of Section 5.2 is not free since any rotation of \( S^2 \) fixes the two points of the rotation axis.

If \( H \) is a subgroup of \( G \), obviously \( H \) acts freely on \( G \) (by multiplication on the left or on the right).

There is a stronger version of the results that we are going to state next that involves the notion of principal bundle. Since this notion is not discussed until Section 28.5, we state weaker versions not dealing with principal bundles. The weaker version that does not mention principal bundles is usually stated for left actions; for instance, in Lee [117] (Chapter 9, Theorem 9.16). We formulate both a left and a right version.

**Theorem 19.8.** Let \( M \) be a smooth manifold, \( G \) be a Lie group, and let \( \cdot : G \times M \to M \) be a left (resp. right) smooth action which is proper and free. Then, the canonical projection \( \pi : G \to G/H \) is a submersion (which means that \( d\pi_g \) is surjective for all \( g \in G \)), and there is a unique manifold structure on \( G/H \) with this property.

Theorem 19.8 has some interesting corollaries. Because a closed subgroup \( H \) of a Lie group \( G \) is a Lie group, and because the action of a closed subgroup is free and proper, we get the following result (proofs can also be found in Bröcker and tom Dieck [31] (Chapter I, Section 4) and in Duistermaat and Kolk [64] (Chapter 1, Section 11)). This is the result we use to verify reductive homogeneous spaces are indeed manifolds.

**Theorem 19.9.** If \( G \) is a Lie group and \( H \) is a closed subgroup of \( G \), then the canonical projection \( \pi : G \to G/H \) is a submersion (which means that \( d\pi_g \) is surjective for all \( g \in G \)), and there is a unique manifold structure on \( G/H \) with this property.
In the special case where $G$ acts transitively on $M$, for any $x \in M$, if $G_x$ is the stabilizer of $x$, then with $H = G_x$, Theorem 19.9 shows that there is a manifold structure on $G/H$ such $\pi: G \to G/H$ is a submersion.

Actually, $G/H$ is diffeomorphic to $M$, as shown by the following theorem whose proof can be found in Lee [117] (Chapter 9, Theorem 9.24)

**Theorem 19.10.** Let $\cdot: G \times M \to M$ be a smooth transitive action of a Lie group $G$ on a smooth manifold $M$ (so that $M$ is a homogeneous space). For any $x \in M$, if $G_x$ is the stabilizer of $x$ and if we write $H = G_x$, then the map $\pi_x: G/H \to M$ given by

$$\pi_x(gH) = g \cdot x$$

is a diffeomorphism and an equivariant map (with respect to the action of $G$ on $G/H$ and the action of $G$ on $M$).

The proof of Theorem 19.10 is not particularly difficult. It relies on technical properties of equivariant maps that we have not discussed. We refer the reader to the excellent account in Lee [117] (Chapter 9).

By Theorem 19.9 and Theorem 19.10, every homogeneous space $M$ (with a smooth $G$-action) is equivalent to a manifold $G/H$ as above. This is an important and very useful result that reduces the study of homogeneous spaces to the study of coset manifolds of the form $G/H$ where $G$ is a Lie group and $H$ is a closed subgroup of $G$.

Here is a simple example of Theorem 19.9. Let $G = \text{SO}(3)$ and

$$H = \left\{ M \in \text{SO}(3) \mid M = \begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix}, \ S \in \text{SO}(2) \right\}.$$

The right action $\cdot: \text{SO}(3) \times H \to \text{SO}(3)$ given by the matrix multiplication

$$g \cdot h = gh, \quad g \in \text{SO}(3), \ h \in H,$$

yields the left cosets $gH$, and the orbit space $\text{SO}(3)/\text{SO}(2)$ which by Theorem 19.9 and Theorem 19.10 is diffeomorphic to $S^2$.

### 19.3 Riemannian Submersions and Coverings Induced by Group Actions

The purpose of this section is to equip the orbit space $M/G$ of Theorem 19.8 with the inner product structure of a Riemannian manifold. Because we provide a different proof for why reductive homogenous manifolds are Riemannian manifolds, namely Proposition 19.19, this section is not necessary for understanding the material in Section 19.4 and may be skipped on the first reading.
If \((N, h)\) is a Riemannian manifold and if \(G\) is a Lie group acting by isometries on \(N\), which means that for every \(g \in G\), the diffeomorphism \(\tau_g : N \rightarrow N\) is an isometry \(((d\tau_g)_p : T_p N \rightarrow T_{\tau_g(p)} N\) is an isometry for all \(p \in M\)), then \(\pi : N \rightarrow N/G\) can be made into a Riemannian submersion.

**Theorem 19.11.** Let \((N, h)\) be a Riemannian manifold and let \(\cdot : G \times N \rightarrow N\) be a smooth, free and proper action, with \(G\) a Lie group acting by isometries of \(N\). Then, there is a unique Riemannian metric \(g\) on \(M = N/G\) such that \(\pi : N \rightarrow M\) is a Riemannian submersion.

**Sketch of proof.** We follow Gallot, Hulin, Lafontaine [73] (Chapter 2, Proposition 2.28). Pick any \(x \in M = N/G\), and any \(u, v \in T_x M\). For any \(p \in \pi^{-1}(x)\), there exist unique lifts \(\overline{u}, \overline{v} \in H_p\) such that \(d\pi_p(\overline{u}) = u\) and \(d\pi_p(\overline{v}) = v\). Set
\[
g_x(u, v) = h_p(\overline{u}, \overline{v}),
\]
which makes \((T_x M, g_x)\) isometric to \((H_p, h_p)\). We need to check that \(g_x\) does not depend on the choice of \(p\) in the fibre \(\pi^{-1}(x)\), and that \((g_x)\) is a smooth family. We check the first property (for the second property, see Gallot, Hulin, Lafontaine [73]). If \(\pi(q) = \pi(p)\), then there is some \(g \in G\) such that \(\tau_g(p) = q\), and \((d\tau_g)_p\) induces an isometry between \(H_p\) and \(H_q\) which commutes with \(\pi\). Therefore, \(g_x\) does not depend on the choice of \(p \in \pi^{-1}(x)\).

As an example, take \(N = S^{2n+1}\), where \(N\) is isomorphic to the subspace of \(\mathbb{C}^{n+1}\) given by
\[
\Sigma^n = \left\{ (z_1, z_2, \ldots, z_{n+1}) \in \mathbb{C}^{n+1} \mid \sum_{i=1}^{n+1} z_i \overline{z_i} = 1 \right\}.
\]
The group \(G = S^1 = \text{SU}(1)\) acts by isometries on \(S^{2n+1}\) by complex multiplication. In other words, given \(p \in \Sigma^n\) and \(e^{i\theta} \in \text{SU}(1)\),
\[
e^{i\theta} \cdot p = (e^{i\theta} z_1, e^{i\theta} z_2, \ldots, e^{i\theta} z_{n+1}) \in \Sigma^n.
\]
Since the action of \(G\) on \(N\) is free and proper, Theorem 19.11 and Example 5.8 imply that obtain the Riemann submersion \(\pi : S^{2n+1} \rightarrow \mathbb{CP}^n\). If we pick the canonical metric on \(S^{2n+1}\), by Theorem 19.11, we obtain a Riemannian metric on \(\mathbb{CP}^n\) known as the *Fubini–Study metric*. Using Proposition 14.7, it is possible to describe the geodesics of \(\mathbb{CP}^n\); see Gallot, Hulin, Lafontaine [73] (Chapter 2).

Another situation where a group action yields a Riemannian submersion is the case where a transitive action is reductive, considered in the next section.

We now consider the case of a smooth action \(\cdot : G \times M \rightarrow M\), where \(G\) is a discrete group (and \(M\) is a manifold). In this case, we will see that \(\pi : M \rightarrow M/G\) is a Riemannian covering map.
Assume $G$ is a discrete group. By Proposition 19.6, the action $\cdot : G \times M \to M$ is proper iff for all $x, y \in M$, there exist some open sets, $V_x$ and $V_y$ in $M$, with $x \in V_x$ and $y \in V_y$, so that the set $K = \{ g \in G \mid g \cdot V_x \cap V_y \neq \emptyset \}$ is finite. By Proposition 19.7, the action $\cdot : G \times M \to M$ is proper iff $G_K = \{ g \in G \mid g \cdot K \cap K \neq \emptyset \}$ is finite for every compact subset $K$ of $M$.

It is shown in Lee [117] (Chapter 9) that the above conditions are equivalent to the conditions below.

**Proposition 19.12.** If $\cdot : G \times M \to M$ is a smooth action of a discrete group $G$ on a manifold $M$, then this action is proper iff

(i) For every $x \in M$, there is some open subset $V$ with $x \in V$ such that $gV \cap V \neq \emptyset$ for only finitely many $g \in G$.

(ii) For all $x, y \in M$, if $y \notin G \cdot x$ ($y$ is not in the orbit of $x$), then there exist some open sets $V, W$ with $x \in V$ and $y \in W$ such that $gV \cap W = 0$ for all $g \in G$.

The following proposition gives necessary and sufficient conditions for a discrete group to act freely and properly often found in the literature (for instance, O’Neill [138], Berger and Gostiaux [20], and do Carmo [60], but beware that in this last reference Hausdorff separation is not required!).

**Proposition 19.13.** If $X$ is a locally compact space and $G$ is a discrete group, then a smooth action of $G$ on $X$ is free and proper iff the following conditions hold:

(i) For every $x \in X$, there is some open subset $V$ with $x \in V$ such that $gV \cap V = \emptyset$ for all $g \in G$ such that $g \neq 1$.

(ii) For all $x, y \in X$, if $y \notin G \cdot x$ ($y$ is not in the orbit of $x$), then there exist some open sets $V, W$ with $x \in V$ and $y \in W$ such that $gV \cap W = 0$ for all $g \in G$.

**Proof.** Condition (i) of Proposition 19.13 implies condition (i) of Proposition 19.12, and condition (ii) is the same in Proposition 19.13 and Proposition 19.12. If (i) holds, then the action must be free since if $g \cdot x = x$, then $gV \cap V \neq \emptyset$, which implies that $g = 1$.

Conversely, we just have to prove that the conditions of Proposition 19.12 imply condition (i) of Proposition 19.13. By (i) of Proposition 19.12, there is some open subset $U$ containing $x$ and a finite number of elements of $G$, say $g_1, \ldots, g_m$, with $g_i \neq 1$, such that $g_iU \cap U \neq \emptyset$, $i = 1, \ldots, m$.

Since our action is free and $g_i \neq 1$, we have $g_i \cdot x \neq x$, so by Hausdorff separation, there exist some open subsets $W_i, W'_i$, with $x \in W_i$ and $g_i \cdot x \in W'_i$, such that $W_i \cap W'_i = \emptyset$, $i = 1, \ldots, m$. Then, if we let

$$V = W \cap \left( \bigcap_{i=1}^{m} (W_i \cap g_i^{-1}W'_i) \right),$$

we see that $V \cap g_i V = \emptyset$, and since $V \subseteq W$, we also have $V \cap gV = \emptyset$ for all other $g \in G$. \qed
Remark: The action of a discrete group satisfying the properties of Proposition 19.13 is often called “properly discontinuous.” However, as pointed out by Lee ([117], just before Proposition 9.18), this term is self-contradictory since such actions are smooth, and thus continuous!

Then we have the following useful result.

**Theorem 19.14.** Let \( N \) be a smooth manifold and let \( G \) be discrete group acting smoothly, freely and properly on \( N \). Then, there is a unique structure of smooth manifold on \( N/G \) such that the projection map \( \pi: N \to N/G \) is a covering map.

For a proof, see Gallot, Hulin, Lafontaine [73] (Theorem 1.88) or Lee [117] (Theorem 9.19).

Real projective spaces are illustrations of Theorem 19.14. Indeed, if \( N \) is the unit \( n \)-sphere \( S^n \subseteq \mathbb{R}^{n+1} \) and \( G = \{ I, -I \} \), where \(-I\) is the antipodal map, then the conditions of Proposition 19.13 are easily checked (since \( S^n \) is compact), and consequently the quotient

\[
\mathbb{RP}^n = S^n / G
\]

is a smooth manifold and the projection map \( \pi: S^n \to \mathbb{RP}^n \) is a covering map. The fiber \( \pi^{-1}([x]) \) of every point \([x] \in \mathbb{RP}^n \) consists of two antipodal points: \( x, -x \in S^n \).

The next step is to see how a Riemannian metric on \( N \) induces a Riemannian metric on the quotient manifold \( N/G \). The following theorem is the Riemannian version of Theorem 19.14.

**Theorem 19.15.** Let \( (N, h) \) be a Riemannian manifold and let \( G \) be discrete group acting smoothly, freely and properly on \( N \), and such that the map \( x \mapsto \sigma \cdot x \) is an isometry for all \( \sigma \in G \). Then there is a unique structure of Riemannian manifold on \( M = N/G \) such that the projection map \( \pi: N \to M \) is a Riemannian covering map.

**Proof sketch.** For a complete proof see Gallot, Hulin, Lafontaine [73] (Proposition 2.20). To define a Riemannian metric \( g \) on \( M = N/G \) we need to define an inner product \( g_p \) on the tangent space \( T_p M \) for every \( p \in M \). Pick any \( q_1 \in \pi^{-1}(p) \) in the fibre of \( p \). Because \( \pi \) is a Riemannian covering map, it is a local diffeomorphism, and thus \( d\pi_{q_1}: T_{q_1}N \to T_p M \) is an isometry. Then, given any two tangent vectors \( u, v \in T_p M \), we define their inner product \( g_p(u, v) \) by

\[
g_p(u, v) = h_{q_1}(d\pi_{q_1}^{-1}(u), d\pi_{q_1}^{-1}(v)).
\]

Now, we need to show that \( g_p \) does not depend on the choice of \( q_1 \in \pi^{-1}(p) \). So, let \( q_2 \in \pi^{-1}(p) \) be any other point in the fibre of \( p \). By definition of \( M = N/G \), we have \( q_2 = g \cdot q_1 \) for some \( g \in G \), and we know that the map \( f: q \mapsto g \cdot q \) is an isometry of \( N \). Now, since \( \pi = \pi \circ f \) we have

\[
d\pi_{q_1} = d\pi_{q_2} \circ df_{q_1},
\]

and since \( d\pi_{q_1}: T_{q_1}N \to T_p M \) and \( d\pi_{q_2}: T_{q_2}N \to T_p M \) are isometries, we get

\[
d\pi_{q_2}^{-1} = df_{q_1} \circ d\pi_{q_1}^{-1}.
\]
But $df_{q_1} : T_{q_1}N \to T_{q_2}N$ is also an isometry, so

$$h_{q_2}(d\pi^{-1}_{q_2}(u), d\pi^{-1}_{q_2}(v)) = h_{q_1}(d\pi^{-1}_{q_1}(u), d\pi^{-1}_{q_1}(v)).$$

Therefore, the inner product $g_p$ is well defined on $T_pM$.

Theorem 19.15 implies that every Riemannian metric $g$ on the sphere $S^n$ induces a Riemannian metric $\tilde{g}$ on the projective space $\mathbb{R}P^n$, in such a way that the projection $\pi : S^n \to \mathbb{R}P^n$ is a Riemannian covering. In particular, if $U$ is an open hemisphere obtained by removing its boundary $S^{n-1}$ from a closed hemisphere, then $\pi$ is an isometry between $U$ and its image $\mathbb{R}P^n - \pi(S^{n-1}) \approx \mathbb{R}P^n - \mathbb{R}P^{n-1}$.

In summary, given a Riemannian manifold $N$ and a group $G$ acting on $N$, Theorem 19.11 gives us a method for obtaining a Riemannian manifold $N/G$ such that $\pi : N \to N/G$ is a Riemannian submersion ($\cdot : G \times N \to N$ is a free and proper action and $G$ acts by isometries). Theorem 19.15 gives us a method for obtaining a Riemannian manifold $N/G$ such that $\pi : N \to N/G$ is a Riemannian covering ($\cdot : G \times N \to N$ is a free and proper action of a discrete group $G$ acting by isometries).

In the next section we show that Riemannian submersions arise from a reductive homogeneous space.

### 19.4 Reductive Homogeneous Spaces

If $\cdot : G \times M \to M$ is a smooth action of a Lie group $G$ on a manifold $M$, then a certain class of Riemannian metrics on $M$ is particularly interesting. Recall that for every $g \in G$, $\tau_g : M \to M$ is the diffeomorphism of $M$ given by

$$\tau_g(p) = g \cdot p, \quad \text{for all } p \in M.$$

If $M = G$ and $G$ acts on itself (on the left) by left multiplication, then $\tau_g = L_g$ for all $g \in G$, as defined earlier in Section 15.1. Thus, the left multiplications $\tau_g$ generalize left multiplications in a group.

**Definition 19.4.** Given a smooth action $\cdot : G \times M \to M$, a metric $\langle -, - \rangle$ on $M$ is $G$-invariant if $\tau_g$ is an isometry for all $g \in G$; that is, for all $p \in M$, we have

$$\langle d(\tau_g)p(u), d(\tau_g)p(v)\rangle_{\tau_g(p)} = \langle u, v \rangle_p \quad \text{for all } u, v \in T_pM.$$

If the action is transitive, then for any fixed $p_0 \in M$ and for every $p \in M$, there is some $g \in G$ such that $p = g \cdot p_0$, so it is sufficient to require that $d(\tau_g)p_0$ be an isometry for every $g \in G$.

From now on we are dealing with a smooth transitive action $\cdot : G \times M \to M$, and for any given $p_0 \in M$, if $H = G_{p_0}$ is the stabilizer of $p_0$, then $M$ is diffeomorphic to $G/H$. 

The existence of $G$-invariant metrics on $G/H$ depends on properties of a certain representation of $H$ called the isotropy representation (see Proposition 19.17). The isotropy representation is equivalent to another representation $\text{Ad}^{G/H}: H \rightarrow \text{GL}(\mathfrak{g}/\mathfrak{h})$ of $H$ involving the quotient algebra $\mathfrak{g}/\mathfrak{h}$.

This representation is too complicated to deal with, so we consider the more tractable situation where the Lie algebra $\mathfrak{g}$ of $G$ factors as a direct sum

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m},$$

for some subspace $\mathfrak{m}$ of $\mathfrak{g}$ such that $\text{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$ for all $h \in H$, where $\mathfrak{h}$ is the Lie algebra of $H$. Then, $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$ and the representation $\text{Ad}^{G/H}: H \rightarrow \text{GL}(\mathfrak{g}/\mathfrak{h})$ becomes the restriction $\text{Ad}: H \rightarrow \mathfrak{m}$ of $\text{Ad}$ to $\mathfrak{m}$. In this situation there is an isomorphism between $T_{p_0}M \cong T_o(G/H)$ and $\mathfrak{m}$ (where $o$ denotes the point in $G/H$ corresponding to the coset $H$). It is also the case that if $H$ is “nice” (for example, compact), then $M = G/H$ will carry $G$-invariant metrics, and that under such metrics, the projection $\pi: G \rightarrow G/H$ is a Riemannian submersion.

In order to proceed it is necessary to express the derivative $d\pi_1: \mathfrak{g} \rightarrow T_o(G/H)$ of the projection map $\pi: G \rightarrow G/H$ in terms of certain vector fields. This is a special case of a process in which an action $\cdot: G \times M \rightarrow M$ associates a vector field $X^*$ on $M$ to every vector $X \in \mathfrak{g}$ in the Lie algebra of $G$.

**Definition 19.5.** Given a smooth action $\varphi: G \times M \rightarrow M$ of a Lie group on a manifold $M$, for every $X \in \mathfrak{g}$, we define the vector field $X^*$ (or $X_M$) on $M$ called an action field or infinitesimal generator of the action corresponding to $X$, by

$$X^*(p) = \left. \frac{d}{dt}(\exp(tX) \cdot p) \right|_{t=0}, \quad p \in M.$$

For a fixed $X \in \mathfrak{g}$, the map $t \mapsto \exp(tX)$ is a curve through 1 in $G$, so the map $t \mapsto \exp(tX) \cdot p$ is a curve through $p$ in $M$, and $X^*(p)$ is the tangent vector to this curve at $p$.

For example, in the case of the adjoint action $\text{Ad}: G \times \mathfrak{g} \rightarrow \mathfrak{g}$, for every $X \in \mathfrak{g}$, we have

$$X^*(Y) = [X,Y],$$

so $X^* = \text{ad}(X)$.

For any $p_0 \in M$, there is a diffeomorphism $G/G_{p_0} \rightarrow G \cdot p_0$ onto the orbit $G \cdot p_0$ of $p_0$ viewed as a manifold, and it is not hard to show that for any $p \in G \cdot p_0$, we have an isomorphism

$$T_p(G \cdot p_0) = \{X^*(p) \mid X \in \mathfrak{g}\};$$

see Marsden and Ratiu [121] (Chapter 9, Section 9.3). It can also be shown that the Lie algebra $\mathfrak{g}_p$ of the stabilizer $G_p$ of $p$ is given by

$$\mathfrak{g}_p = \{X \in \mathfrak{g} \mid X^*(p) = 0\}.$$
The following technical proposition will be needed. It is shown in Marsden and Ratiu [121] (Chapter 9, Proposition 9.3.6 and lemma 9.3.7).

**Proposition 19.16.** Given a smooth action \( \varphi: G \times M \to M \) of a Lie group on a manifold \( M \), the following properties hold:

1. For every \( X \in \mathfrak{g} \), we have
   \[
   (\text{Ad}_g X)^* = \tau_{g^{-1}}^* X^* = (\tau_g)_* X^*, \quad \text{for every } g \in G;
   \]
   Here, \( \tau_{g^{-1}}^* \) is the pullback associated with \( \tau_{g^{-1}} \), and \( (\tau_g)_* \) is the push-forward associated with \( \tau_g \).

2. The map \( X \mapsto X^* \) from \( \mathfrak{g} \) to \( \mathfrak{X}(M) \) is a Lie algebra anti-homomorphism, which means that
   \[
   [X^*, Y^*] = -[X, Y]^* \quad \text{for all } X, Y \in \mathfrak{g}.
   \]

If the metric on \( M \) is \( G \)-invariant (that is, every \( \tau_g \) is an isometry of \( M \)), then the vector field \( X^* \) is a Killing vector field on \( M \) for every \( X \in \mathfrak{g} \).

Given a pair \( (G, H) \), where \( G \) is a Lie group and \( H \) is a closed subgroup of \( G \), it turns out that there is a criterion for the existence of some \( G \)-invariant metric on the homogeneous space \( G/H \) in terms of a certain representation of \( H \) called the isotropy representation. Let us explain what this representation is.

Recall that \( G \) acts on the left on \( G/H \) via
\[
g_1 \cdot (g_2 H) = g_1 g_2 H, \quad g_1, g_2 \in G.
\]
For any \( g_1 \in G \), the diffeomorphism \( \tau_{g_1}: G/H \to G/H \) is left coset multiplication, given by
\[
\tau_{g_1}(g_2 H) = g_1 \cdot (g_2 H) = g_1 g_2 H.
\]
Denote the point in \( G/H \) corresponding to the coset \( 1H = H \) by \( o \). Then, we have a homomorphism
\[
\chi^{G/H}: H \to \text{GL}(T_o(G/H)),
\]
given by
\[
\chi^{G/H}(h) = (d\tau_h)_o, \quad \text{for all } h \in H.
\]
The homomorphism \( \chi^{G/H} \) is called the isotropy representation of the homogeneous space \( G/H \). It is a representation of the group \( H \), and since we can view \( H \) as the isotropy group (the stabilizer) of the element \( o \in G/H \) corresponding to the coset \( H \), it makes sense to call it the isotropy representation. It is not easy to deal with the isotropy representation directly. Fortunately, the isotropy representation is equivalent to another representation \( \text{Ad}^{G/H}: H \to \text{GL}(\mathfrak{g}/\mathfrak{h}) \) obtained from the representation \( \text{Ad}: G \to \text{GL}(\mathfrak{g}) \) by a quotient process that we now describe.
Recall that $\text{Ad}_{g_1}(g_2) = g_1 g_2 g_1^{-1}$ for all $g_1, g_2 \in G$, and that the canonical projection $\pi: G \to G/H$ is given by $\pi(g) = gH$. Then, following O’Neill [138] (see Proposition 22, Chapter 11), observe that $\tau_h \circ \pi = \pi \circ \text{Ad}_h$ for all $h \in H$, since $h \in H$ implies that $h^{-1}H = H$, so for all $g \in G$,

$$(\tau_h \circ \pi)(g) = hgH = hgh^{-1}H = (\pi \circ \text{Ad}_h)(g).$$

By taking derivatives at 1, we get

$$(d\tau_h)_o \circ d\pi_1 = d\pi_1 \circ \text{Ad}_h,$$

which is equivalent to the commutativity of the diagram

$$\begin{array}{ccc}
\mathfrak{g} & \xrightarrow{\text{Ad}_h} & \mathfrak{g} \\
\downarrow d\pi_1 & & \downarrow d\pi_1 \\
T_o(G/H) & \xrightarrow{(dr_h)_o} & T_o(G/H).
\end{array}$$

For any $X \in \mathfrak{g}$, we can express $d\pi_1(X)$ in terms of the vector field $X^*$ introduced in Definition 19.5. Indeed, to compute $d\pi_1(X)$, we can use the curve $t \mapsto \exp(tX)$, and we have

$$d\pi_1(X) = \left. \frac{d}{dt}(\pi(\exp(tX))) \right|_{t=0} = \left. \frac{d}{dt}(\exp(tX)H) \right|_{t=0} = X^*_o.$$ 

For every $X \in \mathfrak{h}$, since the curve $t \mapsto \exp(tX)H$ in $G/H$ has the constant value $o$, we see that

$$\text{Ker } d\pi_1 = \mathfrak{h},$$

and thus, $d\pi_1$ factors through $\mathfrak{g}/\mathfrak{h}$ as $d\pi_1 = \pi_{\mathfrak{g}/\mathfrak{h}} \circ \varphi$, where $\pi_{\mathfrak{g}/\mathfrak{h}}: \mathfrak{g} \to \mathfrak{g}/\mathfrak{h}$ is the quotient map and $\varphi: \mathfrak{g}/\mathfrak{h} \to T_o(G/H)$ is the isomorphism given by the First Isomorphism Theorem. Since $\text{Ad}_h$ is an isomorphism, the kernel of the map $\pi_{\mathfrak{g}/\mathfrak{h}} \circ \text{Ad}_h$ is $\mathfrak{h}$, and by the First Isomorphism Theorem there is a unique map $\text{Ad}_{\mathfrak{g}/\mathfrak{h}}^{G/H}: \mathfrak{g}/\mathfrak{h} \to \mathfrak{g}/\mathfrak{h}$ such that

$$\pi_{\mathfrak{g}/\mathfrak{h}} \circ \text{Ad}_h = \text{Ad}_{\mathfrak{g}/\mathfrak{h}}^{G/H} \circ \pi_{\mathfrak{g}/\mathfrak{h}}$$

making the following diagram commute:

$$\begin{array}{ccc}
\mathfrak{g} & \xrightarrow{\text{Ad}_h} & \mathfrak{g} \\
\downarrow \pi_{\mathfrak{g}/\mathfrak{h}} & & \downarrow \pi_{\mathfrak{g}/\mathfrak{h}} \\
\mathfrak{g}/\mathfrak{h} & \xrightarrow{\text{Ad}_{\mathfrak{g}/\mathfrak{h}}^{G/H}} & \mathfrak{g}/\mathfrak{h}.
\end{array}$$
Then we have the following diagram in which the outermost rectangle commutes and the upper rectangle commutes:

\[
\begin{array}{ccc}
g & \xrightarrow{\text{Ad}_h} & g \\
p_{g/h} & & p_{g/h} \\
g/h & \xrightarrow{\text{Ad}^{G/H}_h} & g/h \\
\phi & & \phi \\
T_o(G/H) & \xrightarrow{(d\tau)_o} & T_o(G/H).
\end{array}
\]

Since \(p_{g/h}\) is surjective, it follows that the lower rectangle commutes; that is

\[
\begin{array}{ccc}
g/h & \xrightarrow{\text{Ad}^{G/H}_h} & g/h \\
\phi & & \phi \\
T_o(G/H) & \xrightarrow{(d\tau)_o} & T_o(G/H)
\end{array}
\]

commutes. Observe that \(\text{Ad}^{G/H}_h\) is a linear isomorphism of \(g/h\) for every \(h \in H\), so that the map \(\text{Ad}^{G/H}_h : H \to \text{GL}(g/h)\) is a representation of \(H\). This proves the first part of the following proposition.

**Proposition 19.17.** Let \((G, H)\) be a pair where \(G\) is a Lie group and \(H\) is a closed subgroup of \(G\). The following properties hold:

1. The representations \(\chi^{G/H}_H : H \to \text{GL}(T_o(G/H))\) and \(\text{Ad}^{G/H}_H : H \to \text{GL}(g/h)\) are equivalent; this means that for every \(h \in H\), we have the commutative diagram

\[
\begin{array}{ccc}
g/h & \xrightarrow{\text{Ad}^{G/H}_h} & g/h \\
\phi & & \phi \\
T_o(G/H) & \xrightarrow{(d\tau)_o} & T_o(G/H)
\end{array}
\]

where the isomorphism \(\phi : g/h \to T_o(G/H)\) and the quotient map \(\text{Ad}^{G/H}_h : g/h \to g/h\) are defined as above.

2. The homogeneous space \(G/H\) has some \(G\)-invariant metric iff the closure of \(\text{Ad}^{G/H}_H(H)\) is compact in \(\text{GL}(g/h)\). Furthermore, this metric is unique up to a scalar if the isotropy representation is irreducible.
We just proved the first part, which is Proposition 2.40 of Gallot, Hulin, Lafontaine [73] (Chapter 2, Section A). The proof of the second part is very similar to the proof of Theorem 17.5; see Gallot, Hulin, Lafontaine [73] (Chapter 2, Theorem 2.42).

The representation $\text{Ad}^{G/H} : H \to \text{GL}(g/h)$ which involves the quotient algebra $g/h$ is hard to deal with. To make things more tractable, it is natural to assume that $g$ splits as a direct sum $g = h \oplus m$, for some well-behaved subspace $m$ of $g$, so that $g/h$ is isomorphic to $m$.

**Definition 19.6.** Let $(G, H)$ be a pair where $G$ is a Lie group and $H$ is a closed subgroup of $G$. We say that the homogeneous space $G/H$ is *reductive* if there is some subspace $m$ of $g$ such that

$$g = h \oplus m,$$

and

$$\text{Ad}_h(m) \subseteq m \quad \text{for all } h \in H.$$

See Figure 19.1.

![Figure 19.1: A schematic illustration of a reductive homogenous manifold. Note that $g = h \oplus m$ and that $T_o(M) \cong m$ via $d\pi_1$.](image)
Observe that unlike $\mathfrak{h}$, which is a Lie subalgebra of $\mathfrak{g}$, the subspace $\mathfrak{m}$ is not necessarily closed under the Lie bracket, so in general it is not a Lie algebra. Also, since $\mathfrak{m}$ is finite-dimensional and since $\text{Ad}_h$ is an isomorphism, we actually have $\text{Ad}_h(\mathfrak{m}) = \mathfrak{m}$.

Definition 19.6 allows us to deal with $\mathfrak{g}/\mathfrak{h}$ in a tractable manner, but does not provide any means of defining a metric on $G/H$. We would like to define $G$-invariant metrics on $G/H$ and key property of a reductive spaces is that there is a criterion for the existence of a $G$-invariant metrics on $G/H$ in terms of $\text{Ad}(H)$-invariant inner products on $\mathfrak{m}$.

Since $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$, by the reasoning just before Proposition 19.17, the map $d\pi_1: \mathfrak{g} \to T_o(G/H)$ restricts to an isomorphism between $\mathfrak{m}$ and $T_o(G/H)$ (where $o$ denotes the point in $G/H$ corresponding to the coset $H$). The representation $\text{Ad}^{G/H}: H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ becomes the representation $\text{Ad}: H \to \mathbf{GL}(\mathfrak{m})$, where $\text{Ad}_h(X)$ is the restriction of $\text{Ad}_h$ to $\mathfrak{m}$ for every $h \in H$.

We also know that for any $X \in \mathfrak{g}$, we can express $d\pi_1(X)$ in terms of the vector field $X^*$ introduced in Definition 19.5 by

$$d\pi_1(X) = X^*_o,$$

and that

$$\text{Ker }d\pi_1 = \mathfrak{h}.$$

Thus, the restriction of $d\pi_1$ to $\mathfrak{m}$ is an isomorphism onto $T_o(G/H)$, given by $X \mapsto X^*_o$. Also, for every $X \in \mathfrak{g}$, since $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, we can write $X = X_h + X_m$, for some unique $X_h \in \mathfrak{h}$ and some unique $X_m \in \mathfrak{m}$, and

$$d\pi_1(X) = d\pi_1(X_m) = X^*_m.$$

We use the isomorphism $d\pi_1$ to transfer any inner product $\langle - ,- \rangle_\mathfrak{m}$ on $\mathfrak{m}$ to an inner product $\langle -,- \rangle$ on $T_o(G/H)$, and vice-versa, by stating that

$$\langle X,Y \rangle_\mathfrak{m} = \langle X^*_o,Y^*_o \rangle, \quad \text{for all } X,Y \in \mathfrak{m};$$

that is, by declaring $d\pi_1$ to be an isometry between $\mathfrak{m}$ and $T_o(G/H)$. See Figure 19.1.

If the metric on $G/H$ is $G$-invariant, then the map $p \mapsto \exp(tX) \cdot p = \exp(tX)aH$ (with $p = aH \in G/H, a \in G$) is an isometry of $G/H$ for every $t \in \mathbb{R}$, so $X^*$ is a Killing vector field.

**Proposition 19.18.** Let $(G,H)$ be a pair of Lie groups defining a reductive homogeneous space $M = G/H$, with reductive decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$. The following properties hold:

1. The isotropy representation $\chi^{G/H}: H \to \mathbf{GL}(T_o(G/H))$ is equivalent to the representation $\text{Ad}: H \to \mathbf{GL}(\mathfrak{m})$ (where $\text{Ad}_h$ is restricted to $\mathfrak{m}$ for every $h \in H$): this means
that for every \( h \in H \), we have the commutative diagram

\[
\begin{array}{ccc}
m & \xrightarrow{\text{Ad}_h} & m \\
\downarrow{d\pi_1} & & \downarrow{d\pi_1} \\
T_o(G/H) & \xrightarrow{(d\tau_h)_o} & T_o(G/H),
\end{array}
\]

where \( d\pi_1 : m \rightarrow T_o(G/H) \) is the isomorphism induced by the canonical projection \( \pi : G \rightarrow G/H \).

(2) By making \( d\pi_1 \) an isometry between \( m \) and \( T_o(G/H) \) (as explained above), there is a one-to-one correspondence between \( G \)-invariant metrics on \( G/H \) and \( \text{Ad}(H) \)-invariant inner products on \( m \) (inner products \( \langle -, - \rangle_m \) such that

\[ \langle u, v \rangle_m = \langle \text{Ad}_h(u), \text{Ad}_h(v) \rangle_m, \quad \text{for all } h \in H \text{ and all } u, v \in m. \]

(3) The homogeneous space \( G/H \) has some \( G \)-invariant metric iff the closure of \( \text{Ad}(H) \) is compact in \( \text{GL}(m) \). Furthermore, if the representation \( \text{Ad} : H \rightarrow \text{GL}(m) \) is irreducible, then such a metric is unique up to a scalar. In particular, if \( H \) is compact, then a \( G \)-invariant metric on \( G/H \) always exists.

Proof. Part (1) follows immediately from the fact that \( \text{Ad}_h(m) \subseteq m \) for all \( h \in H \) and from the identity

\[ (d\tau_h)_o \circ d\pi_1 = d\pi_1 \circ \text{Ad}_h, \]

which was proved just before Proposition 19.17. Part (2) is proved in O'Neill [138] (Chapter 11, Proposition 22), Arvanitoyeorgos [11] (Chapter 5, Proposition 5.1), and Ziller [180] (Chapter 6, Lemma 6.22). Since the proof is quite informative, we provide it. First, assume that the metric on \( G/H \) is \( G \)-invariant. By restricting this \( G \)-invariant metric to the tangent space at \( o \), we will show the existence of a metric on \( m \) obeys the property of \( \text{Ad}(H) \) invariance. For every \( h \in H \), the map \( \tau_h \) is an isometry of \( G/H \), so in particular we have

\[ \langle (d\tau_h)_o(X^*_o), (d\tau_h)_o(Y^*_o) \rangle = \langle X^*_o, Y^*_o \rangle, \quad \text{for all } X, Y \in m. \]

However, the commutativity of the diagram in (1) can be expressed as

\[ (d\tau_h)_o(X^*_o) = (\text{Ad}_h(X))^*_o, \]

so we get

\[ \langle (\text{Ad}_h(X))^*_o, (\text{Ad}_h(Y))^*_o \rangle = \langle X^*_o, Y^*_o \rangle, \]

which is equivalent to

\[ \langle \text{Ad}_h(X), \text{Ad}_h(Y) \rangle_m = \langle X, Y \rangle_m, \quad \text{for all } X, Y \in m. \]
19.4. REDUCTIVE HOMOGENEOUS SPACES

Conversely, assume we have an inner product $\langle -, - \rangle_m$ on $m$ which is $\text{Ad}(H)$-invariant. The proof strategy is as follows: place the metric on $T_o(G/H)$ and then use the maps $\tau_g : G/H \to G/H$ to transfer this metric around $G/H$ in a fashion that is consistent with the notion of $G$-invariance. The condition of $\text{Ad}(H)$-invariance ensures that this construction of the metric on $G/H$ is well defined.

First we transfer this metric on $T_o(G/H)$ using the isomorphism $d\pi_1$ between $m$ and $T_o(G/H)$. Since $(d\tau_o)_o : T_o(G/H) \to T_o(G/H)$ is a linear isomorphism with inverse $(d\tau^{-1}_o)_p$, for any $p = aH$, we define a metric on $G/H$ as follows: for every $p \in G/H$, for any coset representative $aH$ of $p$, set

$$\langle u, v \rangle_p = \langle (d\tau^{-1}_o)_p(u), (d\tau^{-1}_o)_p(v) \rangle_o, \quad \text{for all } u, v \in T_p(G/H).$$

We need to show that the above does not depend on the representative $aH$ chosen for $p$. Here is where we make use of the Ad$(H)$-invariant condition. By reversing the computation that we just made, each map $(d\tau_h)_o$ is an isometry of $T_o(G/H)$ If $bH$ is another representative for $p$, so that $aH = bH$, then $b^{-1}a = h$ for some $h \in H$, so $b^{-1} = ha^{-1}$, and we have

$$\langle (d\tau^{-1}_h)_p(u), (d\tau^{-1}_h)_p(v) \rangle_o = \langle (d\tau_h)_o((d\tau^{-1}_h)_p(u)), (d\tau_h)_o((d\tau^{-1}_h)_p(v)) \rangle_o$$

$$= \langle (d\tau^{-1}_o)_p(u), (d\tau^{-1}_o)_p(v) \rangle_o,$$

since $(d\tau_h)_o$ is an isometry. Since $G$ is a principal $H$-bundle over $G/H$ (see Theorem 19.9), for every $p \in G/H$, there is a local trivialization $\varphi_o : \pi^{-1}(U_o) \to U_\alpha \times H$, where $U_\alpha$ is some open subset in $G/H$ containing $p$, so smooth local sections over $U_\alpha$ exist (for example, pick some $h \in H$ and define $s : U_\alpha \to \pi^{-1}(U_o)$ by $s(q) = \varphi^{-1}_o(q, h)$, for all $q \in U_\alpha$). Given any smooth local section $s$ over $U_\alpha$ (as $s(q) \in G$ and $q = \pi(s(q)) = s(q)H$), we have

$$\langle u, v \rangle_q = \langle (d\tau_{s(q^{-1})})_q(u), (d\tau_{s(q^{-1})})_q(v) \rangle_o, \quad \text{for all } q \in U_\alpha \text{ and all } u, v \in T_q(G/H),$$

which shows that the resulting metric on $G/H$ is smooth. By definition, the metric that we just defined is $G$-invariant.

Part (3) is shown in Gallot, Hulin, Lafontaine [73] (Chapter 2, Theorem 2.42). \qed

At this stage, we have a mechanism to equip $G/H$ with a Riemannian metric from an inner product $m$ which has the special property of being $\text{Ad}(H)$-invariant, but this mechanism does not provide a Riemannian metric on $G$. The construction of a Riemannian metric on $G$ can be done by extending the $\text{Ad}(H)$-invariant metric on $m$ to all of $g$, and using the bijective correspondence between left-invariant metrics on a Lie group $G$, and inner products on its Lie algebra $g$ given by Proposition 17.1.

**Proposition 19.19.** Let $(G, H)$ be a pair of Lie groups defining a reductive homogeneous space $M = G/H$, with reductive decomposition $g = h \oplus m$. If $m$ has some $\text{Ad}(H)$-invariant inner product $\langle -, - \rangle_m$, for any inner product $\langle -, - \rangle_g$ on $g$ extending $\langle -, - \rangle_m$ such that $h$ and $m$ are orthogonal, if we give $G$ the left-invariant metric induced by $\langle -, - \rangle_g$, then the map $\pi : G \to G/H$ is a Riemannian submersion.
CHAPTER 19. MANIFOLDS ARISING FROM GROUP ACTIONS

Proof. (After O’Neill [138] (Chapter 11, Lemma 24). The map $\pi: G \to G/H$ is clearly a smooth submersion. For condition (2) of Definition 14.2, for all $a, b \in G$, since

$$\tau_a(\pi(b)) = \tau_a(bH) = abH = L_a(b)H = \pi(L_a(b)),$$

we have

$$\tau_a \circ \pi = \pi \circ L_a,$$

and by taking derivatives at 1, we get

$$d(\tau_a)_o \circ d\pi_1 = d\pi_a \circ (dL_a)_1.$$

The horizontal subspace at $a \in G$ is $\mathcal{H}_a = (dL_a)_1(m)$, and since the metric on $G$ is left-invariant, $(dL_a)_1$ is an isometry; the map $d(\tau_a)_o$ is an isometry because the metric on $G/H$ is $G$-invariant, and $d\pi_1$ is an isometry between $m$ and $T_o(G/H)$ by construction, so

$$d\pi_a = (d\tau_a)_o \circ d\pi_1 \circ (dL_a^{-1})_a$$

is an isometry between $\mathcal{H}_a$ and $T_p(G/H)$, where $p = aH$. \qed

By Proposition 14.7, a Riemannian submersion carries horizontal geodesics to geodesics.

19.5 Examples of Reductive Homogeneous Spaces

We now apply the theory of Propositions 19.18 and 19.19 to construct a family of reductive homogeneous spaces, the Stiefel manifolds $S(k, n)$. We first encountered the Stiefel manifolds in Section 5.3. For any $n \geq 1$ and any $k$ with $1 \leq k \leq n$, let $S(k, n)$ be the set of all orthonormal $k$-frames, where an orthonormal $k$-frame is a $k$-tuples of orthonormal vectors $(u_1, \ldots, u_k)$ with $u_i \in \mathbb{R}^n$. Recall that $\text{SO}(n)$ acts transitively on $S(k, n)$ via the action

$$R \cdot (u_1, \ldots, u_k) = (Ru_1, \ldots, Ru_k).$$

and that the stabilizer of this action is

$$H = \left\{ \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \biggm| R \in \text{SO}(n-k) \right\}.$$ 

Theorem 19.10 implies that $S(k, n) \cong G/H$, with $G = \text{SO}(n)$ and $H \cong \text{SO}(n-k)$. Observe that the points of $G/H \cong S(k, n)$ are the cosets $QH$, with $Q \in \text{SO}(n)$; that is, the equivalence classes $[Q]$, with the equivalence relation on $\text{SO}(n)$ given by

$$Q_1 \equiv Q_2 \text{ iff } Q_2 = Q_1 \tilde{R}, \text{ for some } \tilde{R} \in H.$$
If we write $Q = [Y \ Y_{\perp}]$, where $Y$ consists of the first $k$ columns of $Q$ and $Y_{\perp}$ consists of the last $n - k$ columns of $Q$, it is clear that $[Q]$ is uniquely determined by $Y$. In fact, if $P_{n,k}$ denotes the projection matrix consisting of the first $k$ columns of the identity matrix $I_n$, 

$$P_{n,k} = \begin{pmatrix} I_k \\ 0_{n-k,k} \end{pmatrix},$$

for any $Q = [Y \ Y_{\perp}]$, the unique representative $Y$ of the equivalence class $[Q]$ is given by

$$Y = QP_{n,k}.$$ 

Furthermore $Y_{\perp}$ is characterized by the fact that $Q = [Y \ Y_{\perp}]$ is orthogonal, namely, $YY^T + Y_{\perp}Y_{\perp}^T = I$.

Define

$$\mathfrak{h} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \mid S \in \mathfrak{so}(n-k) \right\}, \quad \mathfrak{m} = \left\{ \begin{pmatrix} T & -A^T \\ A & 0 \end{pmatrix} \mid T \in \mathfrak{so}(k), A \in M_{n-k,k}(\mathbb{R}) \right\}.$$ 

Clearly $\mathfrak{g} = \mathfrak{so}(n) = \mathfrak{h} \oplus \mathfrak{m}$. For $h \in H$ with $h = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix}$, note that $h^{-1} = \begin{pmatrix} I & 0 \\ 0 & R^T \end{pmatrix}$. Given any $X \in \mathfrak{m}$ with $X = \begin{pmatrix} T & -A^T \\ A & 0 \end{pmatrix}$, we see that

$$hXh^{-1} = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} T & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & R^T \end{pmatrix} = \begin{pmatrix} T & -A^TR^T \\ RA & 0 \end{pmatrix} \in \mathfrak{m},$$

which implies that $\text{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$. Therefore Definition 19.6 shows that $S(k, n) \cong G/H$ is a reductive homogenous manifold with $\mathfrak{g}/\mathfrak{h} \cong \mathfrak{m}$.

Since $H \cong \text{SO}(n-k)$ is compact, Proposition 19.18 guarantees the existence of a $G$-invariant metric on $G/H$, which in turn ensures the existence of an $\text{Ad}(H)$-invariant metric on $\mathfrak{m}$. We construct such a metric by using the Killing form on $\mathfrak{so}(n)$. We know that the Killing form on $\mathfrak{so}(n)$ is given by $B(X, Y) = (n - 2)\text{tr}(XY)$. Now observe that if take $\begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \in \mathfrak{h}$ and $\begin{pmatrix} T & -A^T \\ A & 0 \end{pmatrix} \in \mathfrak{m}$, then

$$\text{tr} \left( \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} T & -A^T \\ A & 0 \end{pmatrix} \right) = \text{tr} \left( \begin{pmatrix} 0 & 0 \\ SA & 0 \end{pmatrix} \right) = 0.$$ 

Furthermore, it is clear that $\dim(\mathfrak{m}) = \dim(\mathfrak{g}) - \dim(\mathfrak{h})$, so $\mathfrak{m}$ is the orthogonal complement of $\mathfrak{h}$ with respect to the Killing form. If $X, Y \in \mathfrak{m}$, with

$$X = \begin{pmatrix} S & -A^T \\ A & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} T & -B^T \\ B & 0 \end{pmatrix},$$

...
observe that
\[
\begin{align*}
\text{tr} \left( \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} T & -B^\top \\ B & 0 \end{pmatrix} \right) &= \text{tr} \left( \begin{pmatrix} ST & A^\top B \\ AT & -AB^\top \end{pmatrix} \right) = \text{tr}(ST) - 2\text{tr}(A^\top B),
\end{align*}
\]
and since \( S^\top = -S \), we have
\[
\text{tr}(ST) - 2\text{tr}(A^\top B) = -\text{tr}(S^\top T) - 2\text{tr}(A^\top B),
\]
so we define an inner product on \( \mathfrak{m} \) by
\[
\langle X, Y \rangle = -\frac{1}{2} \text{tr}(XY) = \frac{1}{2} \text{tr}(X^\top Y) = \frac{1}{2} \text{tr}(S^\top T) + \text{tr}(A^\top B).
\]
We give \( \mathfrak{h} \) the same inner product. For \( X, Y \in \mathfrak{m} \) as defined above, and \( h = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \in H \), we have
\[
\begin{align*}
\text{Ad}_h(X) &= hXh^{-1} = \begin{pmatrix} S & -A^\top R^\top \\ RA & 0 \end{pmatrix}, \\
\text{Ad}_h(Y) &= hYh^{-1} = \begin{pmatrix} T & -B^\top T^\top \\ TB & 0 \end{pmatrix}.
\end{align*}
\]
Thus
\[
\begin{align*}
\text{tr}(\text{Ad}_h(X)\text{Ad}_h(Y)) &= \text{tr} \left( \begin{pmatrix} ST - A^\top B & -SB^\top R^\top \\ TRA & -RAB^\top R^\top \end{pmatrix} \right) \\
&= \text{tr}(ST) - \text{tr}(A^\top B) - \text{tr}(RAB^\top R^\top) \\
&= \text{tr}(ST) - \text{tr}(A^\top B) - \text{tr}(AB^\top R) \\
&= \text{tr}(ST) - \text{tr}(A^\top B) - \text{tr}(AB^\top) \\
&= \text{tr}(ST) - 2\text{tr}(A^\top B) = \text{tr}(XY),
\end{align*}
\]
and this shows that the inner product defined on \( \mathfrak{m} \) is \( \text{Ad}(H) \)-invariant.

Observe that there is a bijection between the space \( \mathfrak{m} \) of \( n \times n \) matrices of the form
\[
X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix}
\]
and the set of \( n \times k \) matrices of the form
\[
\Delta = \begin{pmatrix} S \\ A \end{pmatrix},
\]
but the inner product given by
\[
\langle \Delta_1, \Delta_2 \rangle = \text{tr}(\Delta_1^\top \Delta_2)
\]
yields

$$\langle \Delta_1, \Delta_2 \rangle = \text{tr}(S^T T) + \text{tr}(A^T B),$$

without the factor $1/2$ in front of $S^T T$. These metrics are different.

The vector space $m$ is the tangent space $T_o S(k, n)$ to $S(k, n)$ at $o = [H]$, the coset of the point corresponding to $H$. For any other point $[Q] \in G/H \cong S(k, n)$, the tangent space $T_{[Q]} S(k, n)$ is given by

$$T_{[Q]} S(k, n) = \left\{ Q \begin{pmatrix} S & -A^T \\ A & 0 \end{pmatrix} \mid S \in \mathfrak{so}(k), A \in M_{n-k,k}(\mathbb{R}) \right\}. $$

Using the decomposition $Q = [Y Y_\perp]$, where $Y$ consists of the first $k$ columns of $Q$ and $Y_\perp$ consists of the last $n-k$ columns of $Q$, we have

$$[Y Y_\perp] \begin{pmatrix} S & -A^T \\ A & 0 \end{pmatrix} = [YS + Y_\perp A - YA^T].$$

If we write $X = YS + Y_\perp A$, since $Y$ and $Y_\perp$ are parts of an orthogonal matrix, we have $Y_\perp^T Y = 0$, $Y^T Y = I$, and $Y_\perp^T Y_\perp = I$, so we can recover $A$ from $X$ and $Y_\perp$ and $S$ from $X$ and $Y$, by

$$Y_\perp^T X = Y_\perp^T (YS + Y_\perp A) = A,$$

and

$$Y^T X = Y^T (YS + Y_\perp A) = S.$$ 

Since $A = Y_\perp^T X$, we also have $A^T A = X^T Y_\perp Y_\perp^T X = X^T (I - YY^T) X$.

Therefore, given $Q = [Y Y_\perp]$, the matrices

$$[Y Y_\perp] \begin{pmatrix} S & -A^T \\ A & 0 \end{pmatrix}$$

are in one-to-one correspondence with the $n \times k$ matrices of the form $YS + Y_\perp A$. Since $Y$ describes an element of $S(k, n)$, we can say that the tangent vectors to $S(k, n)$ at $Y$ are of the form

$$X = YS + Y_\perp A, \quad S \in \mathfrak{so}(k), A \in M_{n-k,k}(\mathbb{R}).$$

Since $[Y Y_\perp]$ is an orthogonal matrix, we get $Y^T X = S$, which shows that $Y^T X$ is skew-symmetric. Conversely, since the columns of $[Y Y_\perp]$ form an orthonormal basis of $\mathbb{R}^n$, every $n \times k$ matrix $X$ can be written as

$$X = (Y Y_\perp) \begin{pmatrix} S \\ A \end{pmatrix} = YS + Y_\perp A,$$

where $S \in M_{k,k}$ and $A \in M_{n-k,k}(\mathbb{R})$, and if $Y^T X$ is skew-symmetric, then $S = Y^T X$ is also skew-symmetric. Therefore, the tangent vectors to $S(k, n)$ at $Y$ are the vectors $X \in M_{n,k}(\mathbb{R})$
such that $Y^\top X$ is skew-symmetric. This is the description given in Edelman, Arias and Smith [65].

Another useful observation is that if $X = YS + Y_\perp A$ is a tangent vector to $S(k, n)$ at $Y$, then the square norm $\langle X, X \rangle$ (in the canonical metric) is given by

$$\langle X, X \rangle = \text{tr}\left(X^\top \left(I - \frac{1}{2}YY^\top\right)X\right).$$

Indeed, we have

$$X^\top \left(I - \frac{1}{2}YY^\top\right)X = (S^\top Y^\top + A^\top Y_\perp^\top)\left(I - \frac{1}{2}YY^\top\right)(YS + Y_\perp A)$$

$$= \left(\frac{1}{2}S^\top Y^\top + A^\top Y_\perp^\top\right)YS + A^\top Y_\perp A + \frac{1}{2}S^\top Y^\top Y_\perp A + A^\top Y_\perp^\top YS$$

$$= \langle X, X \rangle.$$

By polarization we find that the canonical metric is given by

$$\langle X_1, X_2 \rangle = \text{tr}\left(X_1^\top \left(I - \frac{1}{2}YY^\top\right)X_2\right).$$

In that paper it is also observed that because $Y_\perp$ has rank $n - k$ (since $Y_\perp^\top Y_\perp = I$), for every $(n - k) \times k$ matrix $A$, there is some $n \times k$ matrix $C$ such that $A = Y_\perp^\top C$ (every column of $A$ must be a linear combination of the $n - k$ columns of $Y_\perp$, which are linearly independent). Thus, we have

$$YS + Y_\perp A = YS + Y_\perp Y_\perp^\top C = YS + (I - YY^\top)C.$$

In order to describe the geodesics of $S(k, n) \cong G/H$, we will need the additional requirement of naturally reductiveness which is defined in the next section.

### 19.6 Naturally Reductive Homogeneous Spaces

When $M = G/H$ is a reductive homogeneous space that has a $G$-invariant metric, it is possible to give an expression for $(\nabla_{X^*} Y^*)_o$ (where $X^*$ and $Y^*$ are the vector fields corresponding to $X, Y \in \mathfrak{m}$).
If \( X^*, Y^*, Z^* \) are the Killing vector fields associated with \( X, Y, Z \in \mathfrak{m} \), then by Proposition 14.9, we have
\[
X^*\langle Y^*, Z^* \rangle = \langle [X^*, Y^*], Z^* \rangle + \langle Y^*, [X^*, Z^*] \rangle \\
Y^*\langle X^*, Z^* \rangle = \langle [Y^*, X^*], Z^* \rangle + \langle X^*, [Y^*, Z^*] \rangle \\
Z^*\langle X^*, Y^* \rangle = \langle [Z^*, X^*], Y^* \rangle + \langle X^*, [Z^*, Y^*] \rangle.
\]
Using the Koszul formula (see Proposition 11.8),
\[
2\langle \nabla_X Y^*, Z^* \rangle = X^*\langle (Y^*, Z^*) \rangle + Y^*\langle (X^*, Z^*) \rangle - Z^*\langle (X^*, Y^*) \rangle \\
- \langle Y^*, [X^*, Z^*] \rangle - \langle X^*, [Y^*, Z^*] \rangle - \langle Z^*, [Y^*, X^*] \rangle,
\]
we obtain
\[
2\langle \nabla_X Y^*, Z^* \rangle = \langle [X^*, Y^*], Z^* \rangle + \langle [X^*, Z^*], Y^* \rangle + \langle [Y^*, Z^*], X^* \rangle.
\]
Since \([X^*, Y^*] = -[X, Y]^*\), we obtain
\[
2\langle \nabla_X Y^*, Z^* \rangle = -\langle [X, Y]^*, Z^* \rangle - \langle [X, Z]^*, Y^* \rangle - \langle [Y, Z]^*, X^* \rangle.
\]

The problem is that the vector field \( \nabla_X Y^* \) is not necessarily of the form \( W^* \) for some \( W \in \mathfrak{g} \). However, we can find its value at \( o \). By evaluating at \( o \) and using the fact that \( X_o^* = (X_o^*)_o \) for any \( X \in \mathfrak{g} \), we obtain
\[
2\langle \langle \nabla_X Y^* \rangle_o, Z_o^* \rangle = -\langle ([X, Y]^*_m)_o, Z_o^* \rangle - \langle ([X, Z]^*_m)_o, Y_o^* \rangle - \langle ([Y, Z]^*_m)_o, X_o^* \rangle.
\]
Hence
\[
2\langle \langle \nabla_X Y^* \rangle_o, Z_o^* \rangle + \langle ([X, Y]^*_m)_o, Z_o^* \rangle = \langle ([Z, X]^*_m)_o, Y_o^* \rangle + \langle ([Z, Y]^*_m)_o, X_o^* \rangle,
\]
and consequently,
\[
\langle \nabla_X Y^* \rangle_o = -\frac{1}{2}\langle [X, Y]^*_m \rangle_o + U(X, Y)_o,
\]
where \([X, Y]^*_m \) is the component of \([X, Y] \) on \( \mathfrak{m} \) and \( U(X, Y) \) is determined by
\[
2\langle U(X, Y), Z \rangle = \langle [Z, X]^*_m, Y \rangle + \langle X, [Z, Y]^*_m \rangle,
\]
for all \( Z \in \mathfrak{m} \). Here, we are using the isomorphism \( X \mapsto X_o^* \) between \( \mathfrak{m} \) and \( T_o(G/H) \) and the fact that the inner product on \( \mathfrak{m} \) is chosen so that \( \mathfrak{m} \) and \( T_o(G/H) \) are isometric.

Since the term \( U(X, Y)_o \) clearly complicates matters, it is natural to make the following definition, which is equivalent to requiring that \( U(X, Y) = 0 \) for all \( X, Y \in \mathfrak{m} \).

**Definition 19.7.** A homogeneous space \( G/H \) is naturally reductive if it is reductive with some reductive decomposition \( \mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m} \), it has a \( G \)-invariant metric, and if
\[
\langle [X, Z]^*_m, Y \rangle = \langle X, [Z, Y]^*_m \rangle, \quad \text{for all } X, Y, Z \in \mathfrak{m}.
\]
If \( G/H \) is naturally reductive, then
\[
(\nabla_X Y^*)_o = -\frac{1}{2}([X,Y]_m)_o.
\]
Since \( G/H \) has a \( G \)-invariant metric, \( X^*, Y^* \) are Killing vector fields on \( G/H \).

We can now find the geodesics on a naturally reductive homogeneous that has a \( G \)-invariant metric. Indeed, if \( M = (G,H) \) is a reductive homogeneous space and \( M \) has a \( G \)-invariant metric, then there is an \( \text{Ad}(H) \)-invariant inner product \( \langle -,- \rangle_m \) on \( M \). Pick any inner product \( \langle -,- \rangle_h \) on \( h \), and define an inner product on \( g = h \oplus m \) by setting \( h \) and \( m \) to be orthogonal. Then, we get a left-invariant metric on \( G \) for which the elements of \( h \) are vertical vectors and the elements of \( m \) are horizontal vectors.

Observe that in this situation, the condition for being naturally reductive extends to left-invariant vector fields on \( G \) induced by vectors in \( m \). Since \( (dL_g)_1 : g \rightarrow T_g G \) is a linear isomorphism for all \( g \in G \), the direct sum decomposition \( g = h \oplus m \) yields a direct sum decomposition \( T_g G = (dL_g)_1(h) \oplus (dL_g)_1(m) \). Given a left-invariant vector field \( X^L \) induced by a vector \( X \in g \), if \( X = X^h + X^m \) is the decomposition of \( X \) onto \( h \oplus m \), we obtain a decomposition
\[
X^L = X^L_h + X^L_m,
\]
into a left-invariant vector field \( X^L_h \in h^L \) and a left-invariant vector field \( X^L_m \in m^L \), with
\[
X^L_h(g) = (dL_g)_1(X^h), \quad X^L_m(g) = (dL_g)_1(X^m).
\]
Since the \( (dL_g)_1 \) are isometries, if \( h \) and \( m \) are orthogonal, so are \( (dL_g)_1(h) \) and \( (dL_g)_1(m) \), and so \( X^L_h \) and \( X^L_m \) are orthogonal vector fields.

Since \( [X^L,Y^L] = [X,Y]^L \), we have \( [X^L,Y^L]_m(g) = [X,Y]^L_m(g) = (dL_g)_1([X,Y]_m) \), so if \( X^L,Y^L,Z^L \) are the left-invariant vector fields induced by \( X,Y,Z \in m \), since the metric on \( G \) is left-invariant, for any \( g \in G \), we have
\[
\langle [X^L,Z^L]_m(g), Y^L(g) \rangle = \langle (dL_g)_1([X,Z]_m), (dL_g)_1(Y) \rangle = \langle [X,Z]_m, Y \rangle.
\]
Similarly, we have
\[
\langle X^L(g), [Z^L,Y^L]_m(g) \rangle = \langle X, [Z,Y]_m \rangle.
\]
Therefore, if the condition for being naturally reductive holds, namely
\[
\langle [X,Z]_m, Y \rangle = \langle X, [Z,Y]_m \rangle, \quad \text{for all} \ X,Y,Z \in m,
\]
then a similar condition holds for left-invariant vector fields:
\[
\langle [X^L,Z^L]_m, Y^L \rangle = \langle X^L, [Z^L,Y^L]_m \rangle, \quad \text{for all} \ X^L,Y^L,Z^L \in m^L.
\]

Recall that the left action of \( G \) on \( G/H \) is given by \( g_1 \cdot g_2 H = g_1 g_2 H \), and that \( o \) denotes the coset \( 1H \).
Proposition 19.20. If $M = G/H$ is a naturally reductive homogeneous space, for every $X \in \mathfrak{m}$, the geodesic $\gamma_{d\pi_1(X)}$ through $o$ is given by

$$\gamma_{d\pi_1(X)}(t) = \pi \circ \exp(tX) = \exp(tX) \cdot o, \quad \text{for all } t \in \mathbb{R}.$$ 

Proof. As explained earlier, since there is a $G$-invariant metric on $G/H$, we can construct a left-invariant metric $\langle -, - \rangle$ on $G$ such that its restriction to $\mathfrak{m}$ is $\text{Ad}(H)$-invariant, and such that $\mathfrak{h}$ and $\mathfrak{m}$ are orthogonal. The curve $\alpha(t) = \exp(tX)$ is horizontal in $G$, since it is an integral curve of the horizontal vector field $X^L \in \mathfrak{m}^L$. By 14.7, the Riemannian submersion $\pi$ carries horizontal geodesics in $G$ to geodesics in $G/H$. Thus, it suffices to show that $\alpha$ is a geodesic in $G$. Following O'Neill (O'Neill [138], Chapter 11, Proposition 25), we prove that

$$\nabla_{X^L} Y^L = \frac{1}{2} [X^L, Y^L], \quad X, Y \in \mathfrak{m}.$$ 

As noted in Section 17.3, since the metric on $G$ is left-invariant, the Koszul formula reduces to

$$2 \langle \nabla_{X^L} Y^L, Z^L \rangle = \langle [X^L, Y^L], Z^L \rangle - \langle [Y^L, Z^L], X^L \rangle + \langle [Z^L, X^L], Y^L \rangle;$$

that is

$$2 \langle \nabla_{X^L} Y^L, Z^L \rangle = \langle [X^L, Y^L], Z^L \rangle + \langle [Z^L, Y^L], X^L \rangle - \langle [X^L, Z^L], Y^L \rangle, \quad \text{for all } X, Y, Z \in \mathfrak{g}.$$ 

Since $\langle -, - \rangle$ and $\mathfrak{m}$ are $\text{Ad}(H)$-invariant, as in the proof of Proposition 17.8, for all $a \in H$,

$$\langle \text{Ad}_a(X), \text{Ad}_a(Y) \rangle = \langle X, Y \rangle, \quad \text{for all } X, Y \in \mathfrak{m},$$

so the function $a \mapsto \langle \text{Ad}_a(X), \text{Ad}_a(Y) \rangle$ is constant, and by taking the derivative with $a = \exp(tZ)$ at $t = 0$, we get

$$\langle [X, Z], Y \rangle = \langle X, [Z, Y] \rangle, \quad X, Y \in \mathfrak{m}, \quad Z \in \mathfrak{h}.$$ 

Since the metric on $G$ is left-invariant, as in the proof of Proposition 17.8, by applying $(dL_g)_e$ to $X, Y, Z$, we obtain

$$\langle [X^L, Z^L], Y^L \rangle = \langle X^L, [Z^L, Y^L] \rangle, \quad X, Y \in \mathfrak{m}, \quad Z \in \mathfrak{h}.$$ 

The natural reductivity condition is

$$\langle [X^L, Z^L]_m, Y^L \rangle = \langle X^L, [Z^L, Y^L]_m \rangle \quad \text{for all } X, Y, Z \in \mathfrak{m}.$$ 

Also recall that $\mathfrak{h}$ and $\mathfrak{m}$ are orthogonal. Let us now consider the Koszul formula for $X, Y \in \mathfrak{m}$ and $Z \in \mathfrak{g}$. If $Z \in \mathfrak{m}$, then by (m), the last two terms cancel out. Similarly, if $Z \in \mathfrak{h}$, then by (h), the last two terms cancel out. Therefore,

$$2 \langle \nabla_{X^L} Y^L, Z^L \rangle = \langle [X^L, Y^L], Z^L \rangle \quad \text{for all } X \in \mathfrak{g},$$
which shows that
\[ \nabla_{X^L}Y^L = \frac{1}{2}[X^L,Y^L], \quad X, Y \in \mathfrak{m}. \]
To finish the proof, the above formula implies that
\[ \nabla_{X^L}X^L = 0, \]
but since \( \alpha \) is a one-parameter group, \( \alpha' = X^L \), which shows that \( \alpha \) is indeed a geodesic.

If \( \gamma \) is any geodesic through \( o \) with initial condition \( X^*_o = d\pi_1(X) \) \( (X \in \mathfrak{m}) \), then the curve \( t \mapsto \exp(tX) \cdot o \) is also a geodesic through \( o \) with the same initial condition, so \( \gamma \) must coincide with this curve. \( \square \)

Proposition 19.20 shows that the geodesics in \( G/H \) are given by the orbits of the one-parameter groups \( (t \mapsto \exp tX) \) generated by the members of \( \mathfrak{m} \).

We can also obtain a formula for the geodesic through every point \( p = gH \in G/H \). Recall from Definition 19.5 that the vector field \( X^* \) associated with a vector \( X \in \mathfrak{m} \) is given by
\[ X^*(p) = \frac{d}{dt}(\exp(tX) \cdot p) \bigg|_{t=0}, \quad p \in G/H. \]
We have an isomorphism between \( \mathfrak{m} \) and \( T_o(G/H) \) given by \( X \mapsto X^*_o \). Furthermore, \( (\tau_g)_* \) induces an isomorphism between \( T_o(G/H) \) and \( T_p(G/H) \). By Proposition 19.16, we have
\[ (\text{Ad}_gX)^* = (\tau_g)_*X^*, \]
so the isomorphism from \( \mathfrak{m} \) to \( T_p(G/H) \) is given by
\[ X \mapsto (\text{Ad}_gX)^*_p. \]
It follows that the geodesic through \( p \) with initial velocity \( (\text{Ad}_gX)^*_p \) is given by
\[ t \mapsto \exp(t\text{Ad}_gX) \cdot p. \]
Since \( \exp(t\text{Ad}_gX) = g \exp(tX)g^{-1} \) and \( g^{-1} \cdot p = o \), the geodesic through \( p = gH \) with initial velocity \( (\text{Ad}_gX)^*_p = (\tau_g)_*X^*_p \) is given by
\[ t \mapsto g \exp(tX) \cdot o. \]

An important corollary of Proposition 19.20 is that naturally reductive homogeneous spaces are complete. Indeed, the one-parameter group \( t \mapsto \exp(tX) \) is defined for all \( t \in \mathbb{R} \).

One can also figure out a formula for the sectional curvature (see O’Neill [138], Chapter 11, Proposition 26). Under the identification of \( \mathfrak{m} \) and \( T_o(G/H) \) given by the restriction of \( d\pi_1 \) to \( \mathfrak{m} \), we have
\[ \langle R(X,Y)X,Y \rangle = \frac{1}{4}([X,Y]_m,[X,Y]_m) + \langle [[X,Y]_h,X]_m,Y \rangle, \quad \text{for all } X, Y \in \mathfrak{m}. \]

Conditions on a homogeneous space that ensure that such a space is naturally reductive are obviously of interest. Here is such a condition.
Proposition 19.21. Let $M = G/H$ be a homogeneous space with $G$ a connected Lie group, assume that $\mathfrak{g}$ admits an $\text{Ad}(G)$-invariant inner product $\langle -, - \rangle$, and let $\mathfrak{m} = \mathfrak{h}^\perp$ be the orthogonal complement of $\mathfrak{h}$ with respect to $\langle -, - \rangle$. Then, the following properties hold:

1. The space $G/H$ is reductive with respect to the decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$.

2. Under the $G$-invariant metric induced by $\langle -, - \rangle$, the homogeneous space $G/H$ is naturally reductive.

3. The sectional curvature is determined by

$$\langle R(X, Y)X, Y \rangle = \frac{1}{4} \langle [X, Y]_\mathfrak{m}, [X, Y]_\mathfrak{m} \rangle + \langle [X, Y]_\mathfrak{h}, [X, Y]_\mathfrak{h} \rangle.$$ 

Sketch of proof. Since $H$ is closed under $\text{Ad}_h$ for every $h \in H$, by taking the derivative at 1 we see that $\mathfrak{h}$ is closed under $\text{Ad}_h$ for all $h \in H$. In fact, since $\text{Ad}_h$ is an isomorphism, we have $\text{Ad}_h(\mathfrak{h}) = \mathfrak{h}$. Since $\mathfrak{m} = \mathfrak{h}^\perp$, we can show that $\mathfrak{m}$ is also closed under $\text{Ad}_h$. If $u \in \mathfrak{m} = \mathfrak{h}^\perp$, then

$$\langle u, v \rangle = 0 \quad \text{for all } v \in \mathfrak{h}.$$ 

Since the inner product $\langle -, - \rangle$ is $\text{Ad}(G)$-invariant, for any $h \in H$ we get

$$\langle \text{Ad}_h(u), \text{Ad}_h(v) \rangle = 0 \quad \text{for all } v \in \mathfrak{h}.$$ 

Since $\text{Ad}_h(\mathfrak{h}) = \mathfrak{h}$, the above means that

$$\langle \text{Ad}_h(u), w \rangle = 0 \quad \text{for all } w \in \mathfrak{h},$$

proving that $\text{Ad}_h(u) \in \mathfrak{h}^\perp = \mathfrak{m}$. Therefore $\text{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$ for all $a \in H$.

To prove (2), since $\langle -, - \rangle$ is $\text{Ad}(G)$-invariant, for all $a \in G$, we have

$$\langle \text{Ad}_a(X), \text{Ad}_a(Y) \rangle = \langle X, Y \rangle, \quad \text{for all } X, Y \in \mathfrak{m},$$

so for $a = \exp(tZ)$ with $Z \in \mathfrak{m}$, by taking derivatives at $t = 0$, we get

$$\langle [X, Z], Y \rangle = \langle X, [Z, Y] \rangle, \quad X, Y, Z \in \mathfrak{m}.$$ 

However, since $\mathfrak{m}$ and $\mathfrak{h}$ are orthogonal, the above implies that

$$\langle [X, Z]_\mathfrak{m}, Y \rangle = \langle X, [Z, Y]_\mathfrak{m} \rangle, \quad X, Y, Z \in \mathfrak{m},$$

which is the natural reductivity condition.

Part (3) is proved in Kobayashi and Nomizu [107] (Chapter X, Theorem 3.5).
Recall a Lie group \( G \) is said to be semisimple if its Lie algebra \( g \) is semisimple. From Theorem 17.22, a Lie algebra \( g \) is semisimple iff its Killing form \( B \) is nondegenerate, and from Theorem 17.24, a connected Lie group \( G \) is compact and semisimple iff its Killing form \( B \) is negative definite. By Proposition 17.21, the Killing form is \( \text{Ad}(G) \)-invariant. Thus, for any connected compact semisimple Lie group \( G \), for any constant \( c > 0 \), the bilinear form \(-cB\) is an \( \text{Ad}(G) \)-invariant inner product on \( g \). Then, as a corollary of Proposition 19.21, we obtain the following result.

\[\text{Proposition 19.22.}\]

Let \( M = G/H \) be a homogeneous space such that \( G \) is a connected compact semisimple group. Then, under any inner product \( \langle -, - \rangle \) on \( g \) given by \(-cB\), where \( B \) is the Killing form of \( g \) and \( c > 0 \) is any positive real, the space \( G/H \) is naturally reductive with respect to the decomposition \( g = h \oplus m \), where \( m = h^\perp \) be the orthogonal complement of \( h \) with respect to \( \langle -, - \rangle \). The sectional curvature is non-negative.

A homogeneous space as in Proposition 19.22 is called a normal homogeneous space.

19.7 Examples of Naturally Reductive Homogeneous Spaces

Since \( \text{SO}(n) \) is connected, semisimple, and compact for \( n \geq 3 \), the Stiefel manifolds \( S(k, n) \cong \text{SO}(n)/\text{SO}(n-k) \) described in Section 19.5 are reductive spaces which satisfy the assumptions of Proposition 19.22 (with an inner product induced by a scalar factor of \(-1/2\) of the Killing form on \( \text{SO}(n) \)). Therefore, Stiefel manifolds \( S(k, n) \) are naturally reductive homogeneous spaces for \( n \geq 3 \) (under the reduction \( g = h \oplus m \) induced by the Killing form).

Another class of naturally reductive homogeneous spaces is the Grassmannian manifolds \( G(k, n) \) which may obtained via a refinement of the Stiefel manifold \( S(k, n) \). Given any \( n \geq 1 \), for any \( k \), with \( 0 \leq k \leq n \), let \( G(k, n) \) be the set of all linear \( k \)-dimensional subspaces of \( \mathbb{R}^n \), where the \( k \)-dimensional subspace \( U \) of \( \mathbb{R} \) is spanned by \( k \) linearly independent vectors \( u_1, \ldots, u_k \) in \( \mathbb{R}^n \); write \( U = \text{span}(u_1, \ldots, u_k) \). In Section 5.3 we have shown that the action

\[ \cdot : \text{SO}(n) \times G(k, n) \to G(k, n) \]

\[ R \cdot U = \text{span}(Ru_1, \ldots, Ru_k). \]

is well-defined, transitive, and has the property that stabilizer of \( U \) is the set of matrices in \( \text{SO}(n) \) with the form

\[ R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix}, \]

where \( S \in \text{O}(k), T \in \text{O}(n-k) \) and \( \det(S) \det(T) = 1 \). We denote this group by \( S(\text{O}(k) \times \text{O}(n-k)) \). Since \( \text{SO}(n) \) is a connected, compact semi-simple Lie group whenever \( n \geq 3 \), Proposition 19.22 implies that

\[ G(k, n) \cong \text{SO}(n)/S(\text{O}(k) \times \text{O}(n-k)) \]
is a naturally reductive homogeneous manifold whenever \( n \geq 3 \).

If \( n = 2 \), then \( \text{SO}(2) \) is an abelian group, and thus not semisimple. However, in this case, \( G(1,2) = \mathbb{RP}(1) \cong \text{SO}(2)/S(O(1) \times O(1)) \cong \text{SO}(2)/O(1) \), and \( S(1,2) = S^1 \cong \text{SO}(2)/\text{SO}(1) \cong S(1,2) \). These are special cases of symmetric spaces discussed in Section 19.9. In the first case, \( H = S(O(1) \times O(1)) \), and in the second case, \( H = \text{SO}(1) \). In both cases,

\[
\mathfrak{h} = (0),
\]

and we can pick

\[
\mathfrak{m} = \mathfrak{so}(2),
\]

which is trivially \( \text{Ad}(H) \)-invariant. In Section 19.9, we show that the inner product on \( \mathfrak{so}(2) \) given by

\[
\langle X,Y \rangle = \text{tr}(X^\top Y)
\]

is \( \text{Ad}(H) \)-invariant, and with the induced metric, \( \mathbb{RP}(1) \) and \( S^1 \cong \text{SO}(2) \) are examples of naturally reductive homogeneous spaces which are also symmetric spaces.

For \( n \geq 3 \), we have \( S(1,n) = S^{n-1} \) and \( S(n-1,n) = \text{SO}(n) \), which are symmetric spaces. On the other hand, \( S(k,n) \) it is not a symmetric space if \( 2 \leq k \leq n - 2 \). A justification is given in Section 19.10.

To construct yet another class of naturally reductive homogeneous spaces known as the oriented Grassmannian \( G^0(k,n) \), we consider the set of \( k \)-dimensional oriented subspaces of \( \mathbb{R}^n \). An oriented \( k \)-subspace is a \( k \)-dimensional subspace \( W \) together with the choice of a basis \( (u_1, \ldots, u_k) \) determining the orientation of \( W \). Another basis \( (v_1, \ldots, v_k) \) of \( W \) is positively oriented if \( \det(f) > 0 \), where \( f \) is the unique linear map \( f \) such that \( f(u_i) = v_i \), \( i = 1, \ldots, k \). The set of of \( k \)-dimensional oriented subspaces of \( \mathbb{R}^n \) is denoted by \( G^0(k,n) \).

The action of \( \text{SO}(n) \) on \( G(k,n) \) is readily adjusted to become a transitive action \( G^0(k,n) \). By a reasoning similar to the one used in the case where \( \text{SO}(n) \) acts on \( G(k,n) \), we find that the stabilizer of the oriented subspace \( (e_1, \ldots, e_k) \) is the set of orthogonal matrices of the form

\[
\begin{pmatrix}
Q & 0 \\
0 & R
\end{pmatrix},
\]

where \( Q \in \text{SO}(k) \) and \( R \in \text{SO}(n-k) \), because this time, \( Q \) has to preserve the orientation of the subspace spanned by \( (e_1, \ldots, e_k) \). Thus, the isotropy group is isomorphic to

\[
\text{SO}(k) \times \text{SO}(n-k).
\]

It follows from Proposition 19.22 that

\[
G^0(k,n) \cong \text{SO}(n)/\text{SO}(k) \times \text{SO}(n-k)
\]

is a naturally reductive homogeneous space whenever \( n \geq 3 \). Furthermore, since \( G^0(1,2) \cong \text{SO}(2)/\text{SO}(1) \times \text{SO}(1) \cong \text{SO}(2)/\text{SO}(1) \cong S(1,2) \), the same reasoning that shows why \( S(1,2) \) is a symmetric space explains why \( G^0(1,2) \cong S^1 \) is also a symmetric space.
Since the Grassmann manifolds $G(k, n)$ and the oriented Grassmann manifolds $G^0(k, n)$ have more structure (they are symmetric spaces), in this section we restrict our attention to the Stiefel manifolds $S(k, n)$. The Grassmannian manifolds $G(k, n)$ and $G^0(k, n)$ are discussed in Section 19.9.

Stiefel manifolds have been presented as reductive homogeneous spaces in Section 19.5, but since they are also naturally reductive, we can describe their geodesics.

By Proposition 19.20, the geodesic through $o$ with initial velocity $X = \left( \begin{array}{cc} S & -A^\top \\ A & 0 \end{array} \right)$ is given by

$$\gamma(t) = \exp\left( t \left( \begin{array}{cc} S & -A^\top \\ A & 0 \end{array} \right) \right) P_{n,k}.$$ 

This is not a very explicit formula. It is possible to do better, see later in this section for details.

Let us consider the case where $k = n - 1$, which is simpler.

If $k = n - 1$, then $n - k = 1$, so $S(n - 1, n) = \text{SO}(n)$, $H \cong \text{SO}(1) = \{1\}$, $\mathfrak{h} = (0)$ and $\mathfrak{m} = \text{so}(n)$. The inner product on $\text{so}(n)$ is given by

$$\langle X, Y \rangle = -\frac{1}{2} \text{tr}(XY) = \frac{1}{2} \text{tr}(X^\top Y), \quad X, Y \in \text{so}(n).$$

Every matrix $X \in \text{so}(n)$ is a skew-symmetric matrix, and we know that every such matrix can be written as $X = P^\top D P$, where $P$ is orthogonal and where $D$ is a block diagonal matrix whose blocks are either a 1-dimensional block consisting of a zero, of a $2 \times 2$ matrix of the form

$$D_j = \begin{pmatrix} 0 & -\theta_j \\ \theta_j & 0 \end{pmatrix},$$

with $\theta_j > 0$. Then, $e^X = P^\top e^D P = P^\top \Sigma P$, where $\Sigma$ is a block diagonal matrix whose blocks are either a 1-dimensional block consisting of a 1, of a $2 \times 2$ matrix of the form

$$D_j = \begin{pmatrix} \cos \theta_j & -\sin \theta_j \\ \sin \theta_j & \cos \theta_j \end{pmatrix}.$$ 

We also know that every matrix $R \in \text{SO}(n)$ can be written as

$$R = e^X,$$

for some matrix $X \in \text{so}(n)$ as above, with $0 < \theta_j \leq \pi$. Then, we can give a formula for the distance $d(I, Q)$ between the identity matrix and any matrix $Q \in \text{SO}(n)$. Since the geodesics from $I$ through $Q$ are of the form

$$\gamma(t) = e^{tx} \quad \text{with} \quad e^x = Q,$$
and since the length \( L(\gamma) \) of the geodesic from \( I \) to \( e^X \) is

\[
L(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle^{\frac{1}{2}} dt,
\]

we have

\[
d(I, Q) = \min_{X|e^X = Q} \int_0^1 \langle (e^{tX})', (e^{tX})' \rangle^{\frac{1}{2}} dt
\]

\[
= \min_{X|e^X = Q} \int_0^1 \langle Xe^{tX}, Xe^{tX} \rangle^{\frac{1}{2}} dt
\]

\[
= \min_{X|e^X = Q} \int_0^1 \left( \frac{1}{2} \text{tr} \left( (e^{tX})^\top X^\top X e^{tX} \right) \right)^{\frac{1}{2}} dt
\]

\[
= \min_{X|e^X = Q} \int_0^1 \left( \frac{1}{2} \text{tr} \left( X^\top X e^{tX} e^{tX}^\top \right) \right)^{\frac{1}{2}} dt
\]

\[
= \min_{X|e^X = Q} \left( \frac{1}{2} \text{tr} \left( X^\top X \right) \right)^{\frac{1}{2}}
\]

\[
= (\theta_1^2 + \cdots + \theta_m^2)^{\frac{1}{2}},
\]

where \( \theta_1, \ldots, \theta_m \) are the angles associated with the eigenvalues \( e^{\pm i\theta_1}, \ldots, e^{\pm i\theta_m} \) of \( Q \) distinct from 1, and with \( 0 < \theta_j \leq \pi \). Therefore,

\[
d(I, Q) = (\theta_1^2 + \cdots + \theta_m^2)^{\frac{1}{2}},
\]

and if \( Q, R \in \text{SO}(n) \), then

\[
d(Q, R) = (\theta_1^2 + \cdots + \theta_m^2)^{\frac{1}{2}},
\]

where \( \theta_1, \ldots, \theta_m \) are the angles associated with the eigenvalues \( e^{\pm i\theta_1}, \ldots, e^{\pm i\theta_m} \) of \( Q^{-1}R = Q^\top R \) distinct from 1, and with \( 0 < \theta_j \leq \pi \).

**Remark:** Since \( X^\top = -X \), the square distance \( d(I, Q)^2 \) can also be expressed as

\[
d(I, Q)^2 = -\frac{1}{2} \min_{X|e^X = Q} \text{tr}(X^2),
\]

or even (with some abuse of notation, since log is multi-valued) as

\[
d(I, Q)^2 = -\frac{1}{2} \min \text{tr}((\log Q)^2).
\]
In the other special case where \( k = 1 \), we have \( S(1, n) = S^{n-1}, H \cong \text{SO}(n-1) \),

\[
\mathfrak{h} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \middle| \ S \in \mathfrak{so}(n-1) \right\},
\]

and

\[
\mathfrak{m} = \left\{ \begin{pmatrix} 0 & -u^\top \\ u & 0 \end{pmatrix} \middle| \ u \in \mathbb{R}^{n-1} \right\}.
\]

Therefore, there is a one-to-one correspondence between \( \mathfrak{m} \) and \( \mathbb{R}^{n-1} \). Given any \( Q \in \text{SO}(n) \), the equivalence class \([Q]\) of \( Q \) is uniquely determined by the first column of \( Q \), and we view it as a point on \( S^{n-1} \).

If we let \( \|u\| = \sqrt{u^\top u} \), we leave it as an exercise to prove that for any

\[
X = \begin{pmatrix} 0 & -u^\top \\ u & 0 \end{pmatrix},
\]

we have

\[
e^{tX} = \begin{pmatrix} \cos(\|u\| t) & -\sin(\|u\| t) \frac{u^\top}{\|u\|^2} \\ \sin(\|u\| t) \frac{u}{\|u\|} & \mathbb{I} + (\cos(\|u\| t) - 1) \frac{uu^\top}{\|u\|^2} \end{pmatrix}.
\]

Consequently (under the identification of \( S^{n-1} \) with the first column of matrices \( Q \in \text{SO}(n) \)), the geodesic \( \gamma \) through \( e_1 \) (the column vector corresponding to the point \( o \in S^{n-1} \)) with initial tangent vector \( u \) is given by

\[
\gamma(t) = \begin{pmatrix} \cos(\|u\| t) \\ \sin(\|u\| t) \frac{u}{\|u\|} \end{pmatrix} = \cos(\|u\| t)e_1 + \sin(\|u\| t)\frac{u}{\|u\|},
\]

where \( u \in \mathbb{R}^{n-1} \) is viewed as the vector in \( \mathbb{R}^n \) whose first component is 0. Then, we have

\[
\gamma'(t) = \|u\| \left( -\sin(\|u\| t)e_1 + \cos(\|u\| t)\frac{u}{\|u\|} \right),
\]

and we find that the length \( L(\gamma)(\theta) \) of the geodesic from \( e_1 \) to the point

\[
p(\theta) = \gamma(\theta) = \cos(\|u\| \theta)e_1 + \sin(\|u\| \theta)\frac{u}{\|u\|},
\]

is given by

\[
L(\gamma)(\theta) = \int_0^\theta \langle \gamma'(t), \gamma'(t) \rangle^{\frac{1}{2}} dt = \theta \|u\|.
\]

Since

\[
\langle e_1, p(\theta) \rangle = \cos(\theta \|u\|),
\]
we see that for a unit vector $u$ and for any angle $\theta$ such that $0 \leq \theta \leq \pi$, the length of the geodesic from $e_1$ to $p(\theta)$ can be expressed as
\[
L(\gamma)(\theta) = \theta = \arccos(\langle e_1, p \rangle);
\]
that is, the angle between the unit vectors $e_1$ and $p$. This is a generalization of the distance between two points on a circle.

Geodesics can also be determined in the general case where $2 \leq k \leq n - 2$; we follow Edelman, Arias and Smith [65], with one change because some point in that paper requires some justification which is not provided.

Given a point $[Y Y_{\perp}] \in S(k, n)$, and given any tangent vector $X = YS + Y_{\perp}A$, we need to compute
\[
\gamma(t) = [Y Y_{\perp}] \exp \left( t \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \right) P_{n,k}.
\]
We can compute this exponential if we replace the matrix by a more “regular matrix,” and for this, we use a QR-decomposition of $A$. Let
\[
A = U \begin{pmatrix} R \\ 0 \end{pmatrix}
\]
be a QR-decomposition of $A$, with $U$ an orthogonal $(n - k) \times (n - k)$ matrix and $R$ an upper triangular $k \times k$ matrix. We can write $U = [U_1 U_2]$, where $U_1$ consists of the first $k$ columns on $U$ and $U_2$ of the last $n - 2k$ columns of $U$ (if $2k \leq n$). We have
\[
A = U_1 R,
\]
and we can write
\[
\begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & U_1 \end{pmatrix} \begin{pmatrix} S & -R^\top \\ R & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & U_1^\top \end{pmatrix}.
\]
Then, we have
\[
\gamma(t) = [Y Y_{\perp}] \begin{pmatrix} I & 0 \\ 0 & U_1 \end{pmatrix} \exp \left( t \begin{pmatrix} S & -R^\top \\ R & 0 \end{pmatrix} \right) \begin{pmatrix} I & 0 \\ 0 & U_1^\top \end{pmatrix} P_{n,k}
\]
\[
= [Y Y_{\perp}U_1] \exp \left( t \begin{pmatrix} S & -R^\top \\ R & 0 \end{pmatrix} \right) \begin{pmatrix} I & 0 \\ 0 & U_1^\top \end{pmatrix} P_{n,k}
\]
\[
= [Y Y_{\perp}U_1] \exp \left( t \begin{pmatrix} S & -R^\top \\ R & 0 \end{pmatrix} \right) I_k.
\]
This is essentially the formula given in Section 2.4.2 of Edelman, Arias and Smith [65], except for the term $Y_1 U_1$. To explain the difference, observe that since $A = U_1 R$, we have $Y_{\perp}A = Y_{\perp}U_1 R$, but $A = Y_1^\top X$ so $Y_{\perp}A = (I - YY^\top)X$, and thus
\[
(I - YY^\top)X = Y_{\perp}U_1 R.
\]
If we write $Q = Y U_1$, then we have
\[ Q^\top Q = U_1^\top Y_1^\top Y_1 U_1 = I, \]
since $Y_1^\top Y_1 = I$ and $U_1^\top U_1 = I$. Therefore,
\[ (I - YY^\top)X = QR \]
is a compact QR-decomposition of $(I - YY^\top)X$.

The problem is that the QR decomposition is not unique in general. Thus, it seems to us that one has to use a QR-decomposition of $A$. Furthermore, given a QR-decomposition of $(I - YY^\top)X$,
\[ (I - YY^\top)X = QR, \]
since $(I - YY^\top)X = Y_1 A$, we get
\[ A = Y_1^\top QR. \]
But,
\[ (Y_1^\top Q)^\top Y_1^\top Q = Q^\top Y_1 Y_1^\top Q, \]
and there is no reason why this term should be equal to $I$. In conclusion, it does not appear that a QR-decomposition of $(I - YY^\top)X$ yields a QR-decomposition of $A$, a fact that seems to be implicitly used in Edelman, Arias and Smith [65]. Instead, a QR-decomposition of $A$ yields a QR-decomposition of $(I - YY^\top)X$. In any case, there are efficient algorithms to compute the exponential of the $2k \times 2k$ matrix
\[ t \begin{pmatrix} S & -R^\top \\ R & 0 \end{pmatrix}. \]

Since by Proposition 14.7(1), the length of the geodesic $\gamma$ from $o$ to $p = e^{sX} \cdot o$ is the same as the the length of the geodesic $\nabla$ in $G$ from 1 to $e^{sX}$, for any $X \in \mathfrak{m}$, we can easily compute the length $L(\gamma)(s)$ of the geodesic $\gamma$ from $o$ to $p = e^{sX} \cdot o$.

Indeed, for any
\[ X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \in \mathfrak{m}, \]
we know that the geodesic (in $G$) from 1 with initial velocity $X$ is $\nabla(t) = e^{tX}$, so we have
\[ L(\gamma)(s) = \langle (e^{tX})', (e^{tX})' \rangle^{1/2} dt, \]
but we already did this computation and found that
\[ (L(\gamma)(s))^2 = s^2 \left( \frac{1}{2} \text{tr}(X^\top X) \right) = s^2 \left( \frac{1}{2} \text{tr}(S^\top S) + \text{tr}(A^\top A) \right). \]
We can compute these traces using the eigenvalues of $S$ and the singular values of $A$. If $\pm i\theta_1, \ldots, \pm i\theta_m$ are the nonzero eigenvalues of $S$ and $\sigma_1, \ldots, \sigma_k$ are the singular values of $A$, then

$$L(\gamma)(s) = s(\theta_1^2 + \cdots + \theta_m^2 + \sigma_1^2 + \cdots + \sigma_k^2)^{\frac{1}{2}}.$$

We conclude this section with a proposition that shows that under certain conditions, $G$ is determined by $m$ and $H$. A point $p \in M = G/H$ is called a pole if the exponential map at $p$ is a diffeomorphism. The following proposition is proved in O’Neill [138] (Chapter 11, Lemma 27).

**Proposition 19.23.** If $M = G/H$ is a naturally reductive homogeneous space, then for any pole $o \in M$, there is a diffeomorphism $m \times H \cong G$ given by the map $(X, h) \mapsto (\exp(X))h$.

Next, we will see that there exists a large supply of naturally reductive homogeneous spaces: symmetric spaces.

**19.8 A Glimpse at Symmetric Spaces**

There is an extensive theory of symmetric spaces and our goal is simply to show that the additional structure afforded by an involutive automorphism of $G$ yields spaces that are naturally reductive. The theory of symmetric spaces was entirely created by one person, Élie Cartan, who accomplished the tour de force of giving a complete classification of these spaces using the classification of semisimple Lie algebras that he had obtained earlier. One of the most complete exposition is given in Helgason [88]. O’Neill [138], Petersen [140], Sakai [150] and Jost [99] have nice and more concise presentations. Ziller [180] is also an excellent introduction, and Borel [23] contains a fascinating historical account.

Until now, we have denoted a homogeneous space by $G/H$, but when dealing with symmetric spaces, it is customary to denote the closed subgroup of $G$ by $K$ rather than $H$.

Given a homogeneous space $G/K$, the new ingredient is that we have an automorphism $\sigma$ of $G$ such that $\sigma \neq \text{id}$ and $\sigma^2 = \text{id}$ called an involutive automorphism of $G$. Let $G^\sigma$ be the set of fixed points of $\sigma$, the subgroup of $G$ given by

$$G^\sigma = \{g \in G \mid \sigma(g) = g\},$$

and let $G_0^\sigma$ be the identity component of $G^\sigma$ (the connected component of $G^\sigma$ containing 1). If $G_0^\sigma \subseteq K \subseteq G^\sigma$, then we can consider the $+1$ and $-1$ eigenspaces of $d\sigma_1: g \to g$, given by

$$\mathfrak{k} = \{X \in \mathfrak{g} \mid d\sigma_1(X) = X\}$$

$$\mathfrak{m} = \{X \in \mathfrak{g} \mid d\sigma_1(X) = -X\}.$$

**Definition 19.8.** An involutive automorphism of $G$ satisfying $G_0^\sigma \subseteq K \subseteq G^\sigma$ is called a Cartan involution. The map $d\sigma_1$ is often denoted by $\theta$. 
The following proposition will be needed later.

**Proposition 19.24.** Let \( \sigma \) be an involutive automorphism of \( G \) and let \( \mathfrak{k} \) and \( \mathfrak{m} \) be the \(+1\) and \(-1\) eigenspaces of \( d\sigma_1 : \mathfrak{g} \to \mathfrak{g} \). Then, for all \( X \in \mathfrak{m} \) and all \( Y \in \mathfrak{k} \), we have

\[
B(X, Y) = 0,
\]

where \( B \) is the Killing form of \( \mathfrak{g} \).

**Proof.** By Proposition 17.21, \( B \) is invariant under automorphisms of \( \mathfrak{g} \). Since \( \theta = d\sigma_1 : \mathfrak{g} \to \mathfrak{g} \) is an automorphism and since \( \mathfrak{m} \) and \( \mathfrak{k} \) are eigenspaces of \( \theta \) for the eigenvalues \(-1\) and \(+1\) respectively, we have

\[
B(X, Y) = B(\theta(X), \theta(Y)) = B(-X, Y) = -B(X, Y),
\]

so \( B(X, Y) = 0 \). \( \Box \)

Remarkably, \( \mathfrak{k} \) and \( \mathfrak{m} \) yield a reductive decomposition of \( G/K \).

**Proposition 19.25.** Given a homogeneous space \( G/K \) with a Cartan involution \( \sigma \) (\( \mathcal{G}_0^\sigma \subseteq K \subseteq \mathcal{G}^\sigma \)), if \( \mathfrak{k} \) and \( \mathfrak{m} \) are defined as above, then

1. \( \mathfrak{k} \) is indeed the Lie algebra of \( K \).
2. We have a reductive decomposition

\[
\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m},
\]

and

\[
[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}, \quad [\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k}.
\]

3. We have \( \text{Ad}(K)(\mathfrak{m}) \subseteq \mathfrak{m} \); in particular \( [\mathfrak{k}, \mathfrak{m}] \subseteq \mathfrak{m} \).

The proof of Proposition 19.25 is not particularly difficult and can be found in O’Neill [138] (Chapter 11) and Ziller [180] (Chapter 6).

If we also assume that \( G \) is connected and that \( \mathcal{G}_0^\sigma \) is compact, then we obtain the following remarkable result proved in O’Neill [138] (Chapter 11) and Ziller [180] (Chapter 6).

**Theorem 19.26.** Let \( G \) be a connected Lie group and let \( \sigma : G \to G \) be an automorphism such that \( \sigma^2 = \text{id} \), \( \sigma \neq \text{id} \) (an involutive automorphism), and \( \mathcal{G}_0^\sigma \) is compact. For every compact subgroup \( K \) of \( G \), if \( \mathcal{G}_0^\sigma \subseteq K \subseteq \mathcal{G}^\sigma \), then \( G/K \) has \( G \)-invariant metrics, and for every such metric \( G/K \) is a naturally reductive space with reductive decomposition \( \mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m} \) given by the \(+1\) and \(-1\) eigenspaces of \( d\sigma_1 \). For every \( p \in G/K \), there is an isometry \( s_p : G/K \to G/K \) such that \( s_p(p) = p \), \( d(s_p)_p = -\text{id} \), and

\[
s_p \circ \pi = \pi \circ \sigma,
\]
as illustrated in the diagram below:

\[
G \xrightarrow{\sigma} G \\
\pi \downarrow \quad \downarrow \pi \\
G/K \xrightarrow{s_p} G/K.
\]

Observe that since \(\sigma\) is a Cartan involution, by Proposition 19.25, we have

\[[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k},\]

so \([X, Z], [Z, Y] \in \mathfrak{k}\) for all \(X, Y, Z \in \mathfrak{m}\), and since \(\mathfrak{k} \cap \mathfrak{m} = (0)\), we have \([X, Z]_{\mathfrak{m}} = [Z, Y]_{\mathfrak{m}} = 0\), which implies the natural reductivity condition

\[
\langle [X, Z]_{\mathfrak{m}}, Y \rangle = \langle X, [Z, Y]_{\mathfrak{m}} \rangle, \quad \text{for all } X, Y, Z \in \mathfrak{m}.
\]

**Definition 19.9.** A triple \((G, K, \sigma)\) satisfying the assumptions of Theorem 19.26 is called a symmetric pair.\(^3\)

A triple \((G, K, \sigma)\) as above defines a special kind of naturally homogeneous space \(G/K\) known as a symmetric space.

If \(M\) is a connected Riemannian manifold, for any \(p \in M\), an isometry \(s_p\) such that \(s_p(p) = p\) and \(d(s_p)_p = -\text{id}\) is a called a global symmetry at \(p\).

**Definition 19.10.** A connected Riemannian manifold \(M\) for which there is a global symmetry for every point of \(M\) is called a symmetric space.

Theorem 19.26 implies that the naturally reductive homogeneous space \(G/K\) defined by a symmetric pair \((G, K, \sigma)\) is a symmetric space.

It can be shown that a global symmetry \(s_p\) reverses geodesics at \(p\) and that \(s_p^2 = \text{id}\), so \(s_p\) is an involution. It should be noted that although \(s_p \in \text{Isom}(M)\), the isometry \(s_p\) does not necessarily lie in \(\text{Isom}(M)_0\).

The following facts are proved in O’Neill [138] (Chapters 9 and 11), Ziller [180] (Chapter 6), and Sakai [150] (Chapter IV).

Every symmetric space is complete, and \(\text{Isom}(M)\) acts transitively on \(M\). In fact the identity component \(\text{Isom}(M)_0\) acts transitively on \(M\). As a consequence, every symmetric space is a homogeneous space of the form \(\text{Isom}(M)_0/K\), where \(K\) is the isotropy group of any chosen point \(p \in M\) (it turns out that \(K\) is compact). The symmetry \(s_p\) gives rise to a Cartan involution \(\sigma\) of \(G = \text{Isom}(M)_0\) defined so that

\[
\sigma(g) = s_p \circ g \circ s_p \quad g \in G.
\]

\(^3\)Once again we fall victims of tradition. A symmetric pair is actually a triple!
Then, we have
\[ G_0^\sigma \subseteq K \subseteq G^\sigma. \]
Therefore, every symmetric space \( M \) is presented by a symmetric pair \((\text{Isom}(M)_0, K, \sigma)\).

However, beware that in the presentation of the symmetric space \( M = G/K \) given by a symmetric pair \((G, K, \sigma)\), the group \( G \) is not necessarily equal to \( \text{Isom}(M)_0 \). Thus, we do not have a one-to-one correspondence between symmetric spaces and symmetric pairs. From our point of view, this does not matter since we are more interested in getting symmetric spaces from the data \((G, K, \sigma)\). By abuse of terminology (and notation), we refer to the homogeneous space \( G/K \) defined by a symmetric pair \((G, K, \sigma)\) as the symmetric space \((G, K, \sigma)\).

Since the homogeneous space \( G/K \) defined by a symmetric pair \((G, K, \sigma)\) is naturally reductive and has a \( G \)-invariant metric, by Proposition 19.20, its geodesics coincide with the one-parameter groups (they are given by the Lie group exponential).

The Levi-Civita connection on a symmetric space depends only on the Lie bracket on \( \mathfrak{g} \). Indeed, we have the following formula proved in Ziller [180] (Chapter 6).

**Proposition 19.27.** Given any symmetric space \( M \) defined by the triple \((G, K, \sigma)\), for any \( X \in \mathfrak{m} \) and vector field \( Y \) on \( M \cong G/K \), we have
\[
(\nabla_X Y)_o = [X^*, Y^*]_o.
\]

**Proof.** If \( X^*, Z^* \) are the Killing vector fields induced by any \( X, Z \in \mathfrak{m} \), by the Koszul formula,
\[
2\langle \nabla_X Y, Z^* \rangle = X^*(\langle Y, Z^* \rangle) + Y(\langle X^*, Z^* \rangle) - Z^*(\langle X^*, Y \rangle) - \langle Y, [X^*, Z^*] \rangle - \langle X^*, [Y, Z^*] \rangle - \langle Z^*, [Y, X^*] \rangle.
\]

Since \( X^* \) and \( Z^* \) are Killing vector fields, we have
\[
X^* \langle Y, Z^* \rangle = \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle,
\]
\[
Z^* \langle X^*, Y \rangle = \langle [Z^*, X^*], Y \rangle + \langle X^*, [Z^*, Y] \rangle,
\]
and because the Levi-Civita connection is symmetric and torsion-free,
\[
Y \langle X^*, Z^* \rangle = \langle \nabla_Y X^*, Z^* \rangle + \langle X^*, \nabla_Y Z^* \rangle,
\]
\[
\langle Z^*, \nabla_Y X^* \rangle = \langle Z^*, \nabla_X Y \rangle - \langle Z^*, [X^*, Y] \rangle,
\]
so we get
\[
Y \langle X^*, Z^* \rangle = \langle \nabla_Y Z^*, X^* \rangle + \langle \nabla_X Y, Z^* \rangle - \langle [X^*, Y], Z^* \rangle.
\]

Plugging these expressions in the Koszul formula, we get
\[
2\langle \nabla_X Y, Z^* \rangle = \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle + \langle \nabla_Y Z^*, X^* \rangle + \langle \nabla_X Y, Z^* \rangle - \langle [X^*, Y], Z^* \rangle - \langle [Z^*, X^*], Y \rangle - \langle X^*, [Z^*, Y] \rangle - \langle Y, [X^*, Z^*] \rangle - \langle X^*, [Y, Z^*] \rangle - \langle Z^*, [Y, X^*] \rangle
\]
\[
= \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle + \langle \nabla_X Y, Z^* \rangle + \langle \nabla_Y Z^*, X^* \rangle.
\]
and thus,
\[
\langle \nabla_{X^*}Y^*, Z^* \rangle = \langle [X^*, Y^*], Z^* \rangle + \langle Y^*, [X^*, Z^*] \rangle + \langle \nabla_Y Z^*, X^* \rangle \\
= \langle [X^*, Y^*], Z^* \rangle - \langle Y^*, [X, Z]^* \rangle + \langle \nabla_Y Z^*, X^* \rangle.
\]

Therefore, evaluating at \( o \) and using the fact that \([X, Z]^*_o = ([X, Z]^* m)_o\), we have
\[
\langle (\nabla_{X^*}Y^*)_o, Z^*_o \rangle = \langle [X^*, Y^*]_o, Z^*_o \rangle - \langle Y_o, ([X, Z]^*_m)_o \rangle + \langle (\nabla_Y Z^*)_o, X^*_o \rangle.
\]

Since \([m, m] \subseteq \mathfrak{k} \) and \( m \cap \mathfrak{k} = (0) \), we have \([X, Z]^*_m = 0 \), so \( \langle Y_o, ([X, Z]^*_m)_o \rangle = 0 \).

Since \( Y_o \in T_o(G/H) \), there is some \( W \in m \) such that \( Y_o = W^*_o \), so
\[
(\nabla_Y Z^*)_o = (\nabla_{Y_o} Z^*)_o = (\nabla_{W^*_o} Z^*)_o = (\nabla_{W^*} Z^*)_o.
\]

Furthermore, since a symmetric space is naturally reductive, we showed earlier (see just before Definition 19.7) that
\[
(\nabla_{W^*} Z^*)_o = -\frac{1}{2}([W, Z]^*_m)_o,
\]
and since \([m, m] \subseteq \mathfrak{k} \), and \( m \cap \mathfrak{k} = (0) \), we have \([W, Z]^*_m = 0 \), which implies that
\[
(\nabla_{W^*} Z^*)_o = 0.
\]

Therefore, \( (\nabla_Y Z^*)_o = 0 \), so \( \langle (\nabla_{X^*}Y^*)_o, Z^*_o \rangle = \langle [X^*, Y^*]_o, Z^*_o \rangle \) for all \( Z \in m \), and we conclude that
\[
(\nabla_{X^*}Y^*)_o = [X^*, Y^*]_o,
\]
as claimed. \( \square \)

Another nice property of symmetric space is that the curvature formulae are quite simple. If we use the isomorphism between \( m \) and \( T_o(G/K) \) induced by the restriction of \( d\pi_1 \) to \( m \), then for all \( X, Y, Z \in m \) we have:

1. The curvature at \( o \) is given by
\[
R(X, Y)Z = [[X, Y], Z],
\]
or more precisely by
\[
R(d\pi_1(X), d\pi_1(Y))d\pi_1(Z) = d\pi_1([[X, Y], Z]).
\]

In terms of the vector fields \( X^*, Y^*, Z^* \), we have
\[
R(X^*, Y^*)Z^* = [[X, Y], Z]^* = [[X^*, Y^*], Z^*].
\]

2. The sectional curvature \( K(X^*, Y^*) \) at \( o \) is determined by
\[
\langle R(X^*, Y^*)X^*, Y^* \rangle = \langle [[X, Y], X], Y \rangle.
\]
3. The Ricci curvature at \( o \) is given by

\[
\text{Ric}(X^*, X^*) = -\frac{1}{2} B(X, X),
\]

where \( B \) is the Killing form associated with \( g \).

Proof of the above formulae can be found in O’Neill [138] (Chapter 11), Ziller [180] (Chapter 6), Sakai [150] (Chapter IV) and Helgason [88] (Chapter IV, Section 4). However, beware that Ziller, Sakai and Helgason use the opposite of the sign convention that we are using for the curvature tensor (which is the convention used by O’Neill [138], Gallot, Hulin, Lafontaine [73], Milnor [125], and Arvanitoyeorgos [11]). Recall that we define the Riemann tensor by

\[
R(X, Y) = \nabla_{[X,Y]} + \nabla_Y \circ \nabla_X - \nabla_X \circ \nabla_Y,
\]

whereas Ziller, Sakai and Helgason use

\[
R(X, Y) = -\nabla_{[X,Y]} - \nabla_Y \circ \nabla_X + \nabla_X \circ \nabla_Y.
\]

With our convention, the sectional curvature \( K(x, y) \) is determined by \( \langle R(x, y)x, y \rangle \), and the Ricci curvature \( \text{Ric}(x, y) \) as the trace of the map \( v \mapsto R(x, v)y \). With the opposite sign convention, the sectional curvature \( K(x, y) \) is determined by \( \langle R(x, y)y, x \rangle \), and the Ricci curvature \( \text{Ric}(x, y) \) as the trace of the map \( v \mapsto R(v, x)y \). Therefore, the sectional curvature and the Ricci curvature are identical under both conventions (as they should!). In Ziller, Sakai and Helgason, the curvature formula is

\[
R(X^*, Y^*)Z^* = -[[X, Y], Z]^*.
\]

We are now going to see that basically all of the familiar spaces are symmetric spaces.

### 19.9 Examples of Symmetric Spaces

We now apply Theorem 19.26 and construct five families of symmetric spaces. We begin by explaining why the Grassmannian manifold \( G(k, n) \cong \text{SO}(n)/\text{SO}(k) \times \text{SO}(n-k) \) and the oriented Grassmannian manifold \( G^0(k, n) \cong \text{SO}(n)/\text{SO}(k) \times \text{SO}(n-k) \) are special cases of naturally reductive homogeneous space known as symmetric spaces. Readers may find material from Absil, Mahony and Sepulchre [2], especially Chapters 1 and 2, a good complement to our presentation, which uses more advanced concepts (symmetric spaces).

1. **Grassmannians as Symmetric Spaces**

   Let \( G = \text{SO}(n) \) (with \( n \geq 2 \), let

   \[
   I_{k,n-k} = \begin{pmatrix}
   I_k & 0 \\
   0 & -I_{n-k}
   \end{pmatrix},
   \]

   for

   \[
   I_k = \begin{pmatrix}
   1 & & \\
   & \ddots & \\
   & & 1
   \end{pmatrix}, \quad I_{n-k} = \begin{pmatrix}
   0 & & \\
   & \ddots & \\
   & & 0
   \end{pmatrix}.
   \]
where $I_k$ is the $k \times k$-identity matrix, and let $\sigma$ be given by

$$\sigma(P) = I_{k,n-k}PI_{k,n-k}, \quad P \in SO(n).$$

It is clear that $\sigma$ is an involutive automorphism of $G$. Let us find the set $F = G^\sigma$ of fixed points of $\sigma$. If we write

$$P = \begin{pmatrix} Q & U \\ V & R \end{pmatrix}, \quad Q \in M_{k,k}(\mathbb{R}), \ U \in M_{k,n-k}(\mathbb{R}), \ V \in M_{n-k,k}(\mathbb{R}), \ R \in M_{n-k,n-k}(\mathbb{R}),$$

then $P = I_{k,n-k}PI_{k,n-k}$ iff

$$\begin{pmatrix} Q & U \\ V & R \end{pmatrix} = \begin{pmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{pmatrix} \begin{pmatrix} Q & U \\ V & R \end{pmatrix} \begin{pmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{pmatrix}$$

iff

$$\begin{pmatrix} Q & U \\ V & R \end{pmatrix} = \begin{pmatrix} Q & -U \\ -V & R \end{pmatrix},$$

so $U = 0$, $V = 0$, $Q \in O(k)$ and $R \in O(n-k)$. Since $P \in SO(n)$, we conclude that $\det(Q)\det(R) = 1$, so

$$G^\sigma = \left\{ \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \bigg| Q \in O(k), \ R \in O(n-k), \ \det(R)\det(S) = 1 \right\};$$

that is,

$$F = G^\sigma = S(O(k) \times O(n-k)),$$

and

$$G^\sigma_0 = SO(k) \times SO(n-k).$$

Therefore, there are two choices for $K$:

1. $K = SO(k) \times SO(n-k)$, in which case we get the Grassmannian $G^0(k,n)$ of oriented $k$-subspaces.

2. $K = S(O(k) \times O(n-k))$, in which case we get the Grassmannian $G(k,n)$ of $k$-subspaces.

As in the case of Stiefel manifolds, given any $Q \in SO(n)$, the first $k$ columns $Y$ of $Q$ constitute a representative of the equivalence class $[Q]$, but these representatives are not unique; there is a further equivalence relation given by

$$Y_1 \equiv Y_2 \quad \text{iff} \quad Y_2 = Y_1R \quad \text{for some} \ R \in O(k).$$

Nevertheless, it is useful to consider the first $k$ columns of $Q$, given by $QP_{n,k}$, as representative of $[Q] \in G(k,n)$. 

Because $\sigma$ is a linear map, its derivative $d\sigma$ is equal to $\sigma$, and since $\mathfrak{so}(n)$ consists of all skew-symmetric $n \times n$ matrices, the $+1$-eigenspace is given by
\[ \mathfrak{k} = \left\{ \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix} \mid S \in \mathfrak{so}(k), T \in \mathfrak{so}(n-k) \right\}, \]
and the $-1$-eigenspace by
\[ \mathfrak{m} = \left\{ \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \mid A \in M_{n-k,k}(\mathbb{R}) \right\}. \]
Thus, $\mathfrak{m}$ is isomorphic to $M_{n-k,k}(\mathbb{R}) \cong \mathbb{R}^{(n-k)k}$. By using the equivalence provided by Proposition 19.18 (1), we can show that the isotropy representation is given by
\[ \text{Ad}((Q,R))A = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} Q^T & 0 \\ 0 & R^T \end{pmatrix} = \begin{pmatrix} 0 & -QA^T R^T \\ RAQ^T & 0 \end{pmatrix} = RAQ^T, \]
where $(Q,R)$ represents an element of $S(O(k) \times O(n-k))$, and $A$ represents an element of $\mathfrak{m}$. It can be shown that this representation is irreducible iff $(k,n) \neq (2,4)$. It can also be shown that if $n \geq 3$, then $G^0(k,n)$ is simply connected, $\pi_1(G(k,n)) = \mathbb{Z}_2$, and $G^0(k,n)$ is a double cover of $G(n,k)$.

An $\text{Ad}(K)$-invariant inner product on $\mathfrak{m}$ is given by
\[ \left\langle \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix}, \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \right\rangle = -\frac{1}{2} \text{tr} \left( \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \right) = \text{tr}(AB^T) = \text{tr}(A^T B). \]
We also give $\mathfrak{g}$ the same inner product. Then, we immediately check that $\mathfrak{k}$ and $\mathfrak{m}$ are orthogonal.

In the special case where $k = 1$, we have $G^0(1,n) = S^{n-1}$ and $G(1,n) = \mathbb{RP}^{n-1}$, and then the $\text{SO}(n)$-invariant metric on $S^{n-1}$ (resp. $\mathbb{RP}^{n-1}$) is the canonical one.

For any point $[Q] \in G(k,n)$ with $Q \in \text{SO}(n)$, if we write $Q = [Y \ Y_\perp]$, where $Y$ denotes the first $k$ columns of $Q$ and $Y_\perp$ denotes the last $n-k$ columns of $Q$, the tangent vectors $X \in T_{[Q]}G(k,n)$ are of the form
\[ X = [Y \ Y_\perp \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} = [Y_\perp A \ -YA^T], \quad A \in M_{n-k,k}(\mathbb{R}). \]
Consequently, there is a one-to-one correspondence between matrices $X$ as above and $n \times k$ matrices of the form $X' = Y_\perp A$, for any matrix $A \in M_{n-k,k}(\mathbb{R})$. As noted in Edelman, Arias and Smith [65], because the spaces spanned by $Y$ and $Y_\perp$ form an orthogonal direct sum in $\mathbb{R}^n$, there is a one-to-one correspondence between $n \times k$ matrices of the form $Y_\perp A$ for any matrix $A \in M_{n-k,k}(\mathbb{R})$, and matrices $X' \in M_{n,k}(\mathbb{R})$ such that
\[ Y^T X' = 0. \]
This second description of tangent vectors to $G(k, n)$ at $[Y]$ is sometimes more convenient. The tangent vectors $X' \in M_{n,k}(\mathbb{R})$ to the Stiefel manifold $S(k, n)$ at $Y$ satisfy the weaker condition that $Y^\top X'$ is skew-symmetric.

Indeed, the tangent vectors at $Y$ to the Stiefel manifold $S(k, n)$ are of the form

$$YS + Y_\perp A,$$

with $S$ skew-symmetric, and since the Grassmanian $G(k, n)$ is obtained from the Stiefel manifold $S(k, n)$ by forming the quotient under the equivalence $Y_1 \equiv Y_2$ iff $Y_2 = Y_1 R$, for some $R \in O(k)$, the contribution $YS$ is a vertical tangent vector at $Y$ in $S(k, n)$, and thus the horizontal tangent vector is $Y_\perp A$; these vectors can be viewed as tangent vectors at $[Y]$ to $G(k, n)$.

Given any $X \in \mathfrak{m}$ of the form

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix},$$

the geodesic starting at $o$ is given by

$$\gamma(t) = \exp(tX) \cdot o.$$ 

Thus, we need to compute

$$\exp(tX) = \exp\begin{bmatrix} 0 & -tA^\top \\ tA & 0 \end{bmatrix}.$$ 

This can be done using SVD.

Since $G(k, n)$ and $G(n - k, n)$ are isomorphic, without loss of generality, assume that $2k \leq n$. Then, let

$$A = U \begin{pmatrix} \Sigma & \\ 0_{n-2k,k} \end{pmatrix} V^\top$$

be an SVD for $A$, with $U$ a $(n - k) \times (n - k)$ orthogonal matrix, $\Sigma$ a diagonal $k \times k$ matrix, and $V$ a $k \times k$ orthogonal matrix. Since we assumed that $k \leq n - k$, we can write

$$U = [U_1 \ U_2],$$

with $U_1$ is a $(n - k) \times k$ matrix and $U_2$ an $(n - k) \times (n - 2k)$ matrix. Then, from

$$A = [U_1 \ U_2] \begin{pmatrix} \Sigma & \\ 0_{n-2k,k} \end{pmatrix} V^\top = U_1 \Sigma V^\top,$$

we get

$$\begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix} = \begin{pmatrix} V & 0 & 0 \\ 0 & U_1 & U_2 \end{pmatrix} \begin{pmatrix} 0 & -\Sigma & 0 \\ \Sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} V^\top & 0 \\ 0 & U_1^\top \\ 0 & U_2^\top \end{pmatrix}. $$
(where the middle matrix is \( n \times n \)). Since
\[
\begin{pmatrix} V^T & 0 \\ 0 & U^T \end{pmatrix} \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} = \begin{pmatrix} V^T V & 0 \\ 0 & U^T U \end{pmatrix} = I_n,
\]
the \( n \times n \) matrix
\[
R = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} = \begin{pmatrix} V & 0 & 0 \\ 0 & U_1 & U_2 \end{pmatrix}
\]
is orthogonal, so we have
\[
\exp(tX) = \exp\left( t \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \right) = R \exp\left( \begin{pmatrix} 0 & -t \Sigma & 0 \\ t \Sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right) R^T.
\]
Then the computation of the middle exponential proceeds just as in the case where \( \Sigma \) is a scalar, so we get
\[
\exp\left( \begin{pmatrix} 0 & -t \Sigma & 0 \\ t \Sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right) = \begin{pmatrix} \cos t \Sigma & -\sin t \Sigma & 0 \\ \sin t \Sigma & \cos t \Sigma & 0 \\ 0 & 0 & I \end{pmatrix},
\]
so
\[
\exp(tX) = \exp\left( t \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix} \right) = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} \cos t \Sigma & -\sin t \Sigma & 0 \\ \sin t \Sigma & \cos t \Sigma & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} V^T & 0 \\ 0 & U^T \end{pmatrix}.
\]
Now, \( \exp(tX)P_{n,k} \) is certainly a representative of the equivalence class of \([\exp(tX)]\), so as a \( n \times k \) matrix, the geodesic through \( o \) with initial velocity
\[
X = \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix}
\]
(with \( A \) any \((n-k) \times k\) matrix with \( n-k \geq k \)) is given by
\[
\gamma(t) = \begin{pmatrix} V & 0 \\ 0 & U_1 \end{pmatrix} \begin{pmatrix} \cos t \Sigma & -\sin t \Sigma & 0 \\ \sin t \Sigma & \cos t \Sigma & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} V^T \\ 0 \end{pmatrix},
\]
where \( A = U_1 \Sigma V^T \), a compact SVD of \( A \).

Remark: Because symmetric spaces are geodesically complete, we get an interesting corollary. Indeed, every equivalence class \([Q] \in G(k, n)\) possesses some representative of the form \( e^X \) for some \( X \in \mathfrak{m} \), so we conclude that for every orthogonal matrix \( Q \in \text{SO}(n) \), there exist some orthogonal matrices \( V, \tilde{V} \in O(k) \) and \( U, \tilde{U} \in O(n-k) \), and some diagonal matrix \( \Sigma \) with nonnegatives entries, so that
\[
Q = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} \cos \Sigma & -\sin \Sigma & 0 \\ \sin \Sigma & \cos \Sigma & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} (\tilde{V})^T & 0 \\ 0 & (\tilde{U})^T \end{pmatrix}.
\]
19.9. EXAMPLES OF SYMMETRIC SPACES

This is an instance of the CS-decomposition; see Golub and Van Loan [76]. The matrices $\cos \Sigma$ and $\sin \Sigma$ are actually diagonal matrices of the form

$$
\cos \Sigma = \text{diag}(\cos \theta_1, \ldots, \cos \theta_k) \quad \text{and} \quad \sin \Sigma = \text{diag}(\sin \theta_1, \ldots, \sin \theta_k),
$$

so we may assume that $0 \leq \theta_i \leq \pi/2$, because if $\cos \theta_i$ or $\sin \theta_i$ is negative, we can change the sign of the $i$th row of $V$ (resp. the sign of the $i$-th row of $U$) and still obtain orthogonal matrices $U'$ and $V'$ that do the job. One should also observe that the first $k$ columns of $Q$ are

$$
Y = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} \cos \Sigma \\ \sin \Sigma \\ 0 \end{pmatrix} (\tilde{V})^\top,
$$

and that the matrix

$$
V(\cos \Sigma)(\tilde{V})^\top
$$

is an SVD for the matrix $P_{n,k}^T Y$, which consists of the first $k$ rows of $Y$. Now, it is known that $(\theta_1, \ldots, \theta_k)$ are the principal angles (or Jordan angles) between the subspaces spanned the first $k$ columns of $I_n$ and the subspace spanned by the columns of $Y$ (see Golub and vanLoan [76]). Recall that given two $k$-dimensional subspaces $U$ and $V$ determined by two $n \times k$ matrices $Y_1$ and $Y_2$ of rank $k$, the principal angles $\theta_1, \ldots, \theta_k$ between $U$ and $V$ are defined recursively as follows: Let $U_1 = U$, $V_1 = V$, let

$$
\cos \theta_1 = \max_{u \in U, v \in V} \frac{\langle u, v \rangle}{\|u\|_2 \|v\|_2},
$$

let $u_1 \in U$ and $v_1 \in V$ be any two unit vectors such that $\cos \theta_1 = \langle u_1, v_1 \rangle$, and for $i = 2, \ldots, k$, if $U_i = U_{i-1} \cap \{u_{i-1}\}^\perp$ and $V_i = V_{i-1} \cap \{v_{i-1}\}^\perp$, let

$$
\cos \theta_i = \max_{u \in U_i, v \in V_i} \frac{\langle u, v \rangle}{\|u\|_2 \|v\|_2},
$$

and let $u_i \in U_i$ and $v_i \in V_i$ be any two unit vectors such that $\cos \theta_i = \langle u_i, v_i \rangle$.

The vectors $u_i$ and $v_i$ are not unique, but it is shown in Golub and van Loan [76] that $(\cos \theta_1, \ldots, \cos \theta_k)$ are the singular values of $Y_1^T Y_2$ (with $0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k \leq \pi/2$).

We can also determine the length $L(\gamma)(s)$ of the geodesic $\gamma(t)$ from $o$ to $p = e^{sX} \cdot o$, for any $X \in \mathfrak{m}$, with

$$
X = \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix}.
$$

Since by Proposition 14.7(1), the length of the geodesic $\gamma$ from $o$ to $p = e^{sX} \cdot o$ is the same as the the length of the geodesic $\gamma$ in $G$ from $1$ to $e^{sX}$, for any $X \in \mathfrak{m}$, the computation from Section 19.7 remains valid, and we obtain

$$
(L(\gamma)(s))^2 = (L(\gamma)(s))^2 = s^2 \left( \frac{1}{2} \text{tr}(X^T X) \right) = s^2 \text{tr}(A^T A).
$$
Then, if $\theta_1, \ldots, \theta_k$ are the singular values of $A$, we get

$$L(\gamma)(s) = s(\theta_1^2 + \cdots + \theta_k^2)^{1/2}. $$

In view of the above discussion regarding principal angles, we conclude that if $Y_1$ consists of the first $k$ columns of an orthogonal matrix $Q_1$ and if $Y_2$ consists of the first $k$ columns of an orthogonal matrix $Q_2$ then the distance between the subspaces $[Q_1]$ and $[Q_2]$ is given by

$$d([Q_1], [Q_2]) = (\theta_1^2 + \cdots + \theta_k^2)^{1/2},$$

where $(\cos \theta_1, \ldots, \cos \theta_k)$ are the singular values of $Y_1^\top Y_2$ (with $0 \leq \theta_i \leq \pi/2$); the angles $(\theta_1, \ldots, \theta_k)$ are the principal angles between the spaces $[Q_1]$ and $[Q_2]$.

In Golub and van Loan, a different distance between subspaces is defined, namely

$$d_{p2}( [Q_1], [Q_2]) = \|Y_1 Y_1^\top - Y_2 Y_2^\top \|_2. $$

If we write $\Theta = \text{diag}(\theta_1, \ldots, \theta_k)$, then it is shown that

$$d_{p2}( [Q_1], [Q_2]) = \|\sin \Theta\|_\infty = \max_{1 \leq i \leq k} \sin \theta_i. $$

This metric is derived by embedding the Grassmannian in the set of $n \times n$ projection matrices of rank $k$, and then using the 2-norm. Other metrics are proposed in Edelman, Arias and Smith [65].

We leave it to the brave readers to compute $\langle [[X,Y],[X,Y]] \rangle$, where

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix},$$

and check that

$$\langle [[X,Y],[X,Y]] \rangle = \langle BA^\top - AB^\top, BA^\top - AB^\top \rangle + \langle A^\top B - B^\top A, A^\top B - B^\top A \rangle,$$

which shows that the sectional curvature is nonnegative. When $k = 1$ (or $k = n - 1$), which corresponds to $\mathbb{R}P^{n-1}$ (or $S^{n-1}$), we get a metric of constant positive curvature.

2. Symmetric Positive Definite Matrices

Recall that the space $\text{SPD}(n)$ of symmetric positive definite matrices ($n \geq 2$) appears as the homogeneous space $\text{GL}^+(n, \mathbb{R})/\text{SO}(n)$, under the action of $\text{GL}^+(n, \mathbb{R})$ on $\text{SPD}(n)$ given by

$$A \cdot S = ASA^\top.$$

Write $G = \text{GL}^+(n, \mathbb{R})$, $K = \text{SO}(n)$, and choose the Cartan involution $\sigma$ given by

$$\sigma(S) = (S^\top)^{-1}.$$
It is immediately verified that 
\[ G^\sigma = \text{SO}(n), \]
and that the derivative \( \theta = d\sigma_1 \) of \( \sigma \) is given by
\[ \theta(S) = -S^T, \quad S \in \text{M}_n(\mathbb{R}), \]
since \( \mathfrak{gl}^+(n) = \mathfrak{gl}(n) = \text{M}_n(\mathbb{R}) \). It follows that \( \mathfrak{k} = \text{so}(n) \), and \( \mathfrak{m} = \text{S}(n) \), the vector space of symmetric matrices. We define an \( \text{Ad(\text{SO}(n))} \)-invariant inner product on \( \mathfrak{gl}^+(n) \) by
\[ \langle X,Y \rangle = \text{tr}(X^T Y). \]
If \( X \in \mathfrak{m} \) and \( Y \in \mathfrak{k} = \text{so}(n) \), then
\[ \langle X,Y \rangle = \text{tr}(X^T Y) = \text{tr}((X^T Y)^\top) = \text{tr}(Y X^T) = -\text{tr}(X^T Y) = -\langle X,Y \rangle, \]
so \( \langle X,Y \rangle = 0 \). Thus, we have
\[ \langle X,Y \rangle = \begin{cases} -\text{tr}(XY) & \text{if } X,Y \in \mathfrak{k} \\ \text{tr}(XY) & \text{if } X,Y \in \mathfrak{m} \\ 0 & \text{if } X \in \mathfrak{m}, \ Y \in \mathfrak{k}. \end{cases} \]

We leave it as an exercise (see Petersen [140], Chapter 8, Section 2.5) to show that
\[ \langle [[X,Y],X],Y \rangle = -\text{tr}([[X,Y]^\top[X,Y]]), \quad \text{for all } X,Y \in \mathfrak{m}. \]
This shows that the sectional curvature is nonpositive. It can also be shown that the isotropy representation is given by
\[ \chi_A(X) = AXA^{-1} = AXA^\top, \]
for all \( A \in \text{SO}(n) \) and all \( X \in \mathfrak{m} \).

Recall that the exponential \( \exp : \text{S}(n) \to \text{SPD}(n) \) is a bijection. Then, given any \( S \in \text{SPD}(n) \), there is a unique \( X \in \mathfrak{m} \) such that \( S = e^X \), and the unique geodesic from \( I \) to \( S \) is given by
\[ \gamma(t) = e^{tX}. \]
Let us try to find the length \( L(\gamma) = d(I,S) \) of this geodesic. As in Section 19.7, we have
\[ L(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle^{\frac{1}{2}} dt, \]
but this time, \( X \in \mathfrak{m} \) is symmetric and the geodesic is unique, so we have
\[ L(\gamma) = \int_0^1 \langle (e^{tX})', (e^{tX})' \rangle^{\frac{1}{2}} dt = \int_0^1 \langle X e^{tX}, X e^{tX} \rangle^{\frac{1}{2}} dt = \int_0^1 \langle \text{tr}((e^{tX})^\top X^\top e^{tX}) \rangle^{\frac{1}{2}} dt = \int_0^1 \langle \text{tr}(X^2 e^{2tX}) \rangle^{\frac{1}{2}} dt. \]
Since $X$ is a symmetric matrix, we can write
\[ X = P^T \Lambda P, \]
with $P$ orthogonal and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, a real diagonal matrix, and we have
\[
\text{tr}(X^2 e^{2tX}) = \text{tr}(P^T \Lambda^2 PP^T e^{2t\Lambda} P)
= \text{tr}(\Lambda^2 e^{2t\Lambda})
= \lambda_1^2 e^{2t\lambda_1} + \cdots + \lambda_n^2 e^{2t\lambda_n}.
\]
Therefore,
\[
d(I, S) = L(\gamma) = \int_0^1 \left( \lambda_1^2 e^{2\lambda_1 t} + \cdots + \lambda_n^2 e^{2\lambda_n t} \right)^{\frac{1}{2}} dt.
\]
Actually, since $S = e^X$ and $S$ is SPD, $\lambda_1, \ldots, \lambda_n$ are the logarithms of the eigenvalues $\sigma_1, \ldots, \sigma_n$ of $X$, so we have
\[
d(I, S) = L(\gamma) = \int_0^1 \left( (\log \sigma_1)^2 e^{2\log \sigma_1 t} + \cdots + (\log \sigma_n)^2 e^{2\log \sigma_n t} \right)^{\frac{1}{2}} dt.
\]
Unfortunately, there doesn’t appear to be a closed form formula for this integral.

The symmetric space $\text{SPD}(n)$ contains an interesting submanifold, namely the space of matrices $S$ in $\text{SPD}(n)$ such that $\det(S) = 1$. This the symmetric space $\text{SL}(n, \mathbb{R})/\text{SO}(n)$, which we suggest denoting by $\text{SSPD}(n)$. For this space, $\mathfrak{g} = \mathfrak{sl}(n)$, and the reductive decomposition is given by
\[
\mathfrak{t} = \mathfrak{so}(n), \quad \mathfrak{m} = \mathfrak{S}(n) \cap \mathfrak{sl}(n).
\]
Now, recall that the Killing form on $\mathfrak{g}(n)$ is given by
\[
B(X, Y) = 2n \text{tr}(XY) - 2 \text{tr}(X) \text{tr}(Y).
\]
On $\mathfrak{sl}(n)$, the Killing form is $B(X, Y) = 2n \text{tr}(XY)$, and it is proportional to the inner product
\[
\langle X, Y \rangle = \text{tr}(XY).
\]
Therefore, we see that the restriction of the Killing form of $\mathfrak{sl}(n)$ to $\mathfrak{m} = \mathfrak{S}(n) \cap \mathfrak{sl}(n)$ is positive definite, whereas it is negative definite on $\mathfrak{t} = \mathfrak{so}(n)$. The symmetric space $\text{SSPD}(n) \cong \text{SL}(n, \mathbb{R})/\text{SO}(n)$ is an example of a symmetric space of noncompact type. On the other hand, the Grassmannians are examples of symmetric spaces of compact type (for $n \geq 3$). In the next section, we take a quick look at these special types of symmetric spaces.

3. The Hyperbolic Space $\mathcal{H}^+_n(1)$ ⊗

In Section 6.1 we defined the Lorentz group $\text{SO}_0(n, 1)$ as follows: if
\[
J = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},
\]
then a matrix $A \in \mathbb{M}_{n+1}(\mathbb{R})$ belongs to $\text{SO}_0(n, 1)$ iff
\[ A^\top JA = J, \quad \det(A) = +1, \quad a_{n+1,n+1} > 0. \]
In that same section we also defined the hyperbolic space $\mathcal{H}_n^+(1)$ as the sheet of $\mathcal{H}_n(1)$ which contains $(0, \ldots, 0, 1)$ where
\[ \mathcal{H}_n(1) = \{ u = (u, t) \in \mathbb{R}^{n+1} | \|u\|^2 - t^2 = -1 \}. \]
We also showed that the action $\cdot : \text{SO}_0(n, 1) \times \mathcal{H}_n^+(1) \to \mathcal{H}_n^+(1)$ with $A \cdot u = Au$
is a transitive with stabilizer $\text{SO}(n)$ (see Proposition 6.9). Thus $\mathcal{H}_n^+(1)$ arises as the homogeneous space $\text{SO}_0(n, 1)/\text{SO}(n)$.
Since the inverse of $A \in \text{SO}_0(n, 1)$ is $JA^\top J$, the map $\sigma : \text{SO}_0(n, 1) \to \text{SO}_0(n, 1)$ given by
\[ \sigma(A) = JAJ = (A^{-1})^\top \]
is an involutive automorphism of $\text{SO}_0(n, 1)$. Write $G = \text{SO}_0(n, 1), K = \text{SO}(n)$. It is immediately verified that
\[ G^\sigma = \left\{ \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \mid Q \in \text{SO}(n) \right\}, \]
so $G^\sigma \cong \text{SO}(n)$. We have
\[ \mathfrak{so}(n, 1) = \left\{ \begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} \mid B \in \mathfrak{so}(n), u \in \mathbb{R}^n \right\}, \]
and the derivative $\theta : \mathfrak{so}(n, 1) \to \mathfrak{so}(n, 1)$ of $\sigma$ at $I$ is given by
\[ \theta(X) = JXJ = -X^\top. \]
From this we deduce that the $+1$-eigenspace is given by
\[ \mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \mid B \in \mathfrak{so}(n) \right\}, \]
and the $-1$-eigenspace is given by
\[ \mathfrak{m} = \left\{ \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \mid u \in \mathbb{R}^n \right\}, \]
with
\[ \mathfrak{so}(n, 1) = \mathfrak{k} \oplus \mathfrak{m}, \]
a reductive decomposition. We define an \( \text{Ad}(K) \)-invariant inner product on \( \mathfrak{so}(n, 1) \) by
\[
\langle X, Y \rangle = \frac{1}{2} \text{tr}(X^TY).
\]
In fact, on \( \mathfrak{m} \cong \mathbb{R}^n \), we have
\[
\left\langle \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}, \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix} \right\rangle = \frac{1}{2} \text{tr}\left( \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix} \right) = \frac{1}{2} \text{tr}(uv^\top + u^\top v) = u^\top v,
\]
the Euclidean product of \( u \) and \( v \).

As an exercise, the reader should compute \( \langle [[X, Y], X], Y \rangle \), where
\[
X = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix},
\]
and check that
\[
\langle [[X, Y], X], Y \rangle = -\langle uv^\top - vu^\top, uv^\top - vu^\top \rangle,
\]
which shows that the sectional curvature is nonpositive. In fact, \( \mathcal{H}_{n+1}^+ \) has constant negative sectional curvature.

We leave it as an exercise to prove that for \( n \geq 2 \), the Killing form \( B \) on \( \mathfrak{so}(n, 1) \) is given by
\[
B(X, Y) = (n - 1)\text{tr}(XY),
\]
for all \( X, Y \in \mathfrak{so}(n, 1) \). If we write
\[
X = \begin{pmatrix} B_1 & u \\ u^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} B_2 & v \\ v^\top & 0 \end{pmatrix},
\]
then
\[
B(X, Y) = (n - 1)\text{tr}(B_1B_2) + 2(n - 1)u^\top v.
\]
This shows that \( B \) is negative definite on \( \mathfrak{k} \) and positive definite on \( \mathfrak{m} \). This means that the space \( \mathcal{H}^+_{n+1} \) is a symmetric space of noncompact type.

The symmetric space \( \mathcal{H}^+_{n+1} = \text{SO}_0(n, 1)/\text{SO}(n) \) turns out to be dual, as a symmetric space, to \( S^n = \text{SO}(n+1)/\text{SO}(n) \). For the precise notion of duality in symmetric spaces, we refer the reader to O’Neill [138].

4. The Hyperbolic Grassmannian \( G^*(q, p + q) \)

Recall from Section 6.1 that we define \( I_{p,q} \), for \( p, q \geq 1 \), by
\[
I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}.
\]
19.9. EXAMPLES OF SYMMETRIC SPACES

If \( n = p + q \), the matrix \( I_{p,q} \) is associated with the nondegenerate symmetric bilinear form

\[
\varphi_{p,q}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{n} x_j y_j
\]

with associated quadratic form

\[
\Phi_{p,q}((x_1, \ldots, x_n)) = \sum_{i=1}^{p} x_i^2 - \sum_{j=p+1}^{n} x_j^2.
\]

The group \( \text{SO}(p, q) \) is the set of all \( n \times n \)-matrices (with \( n = p + q \))

\[
\text{SO}(p, q) = \{ A \in \text{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}, \ \det(A) = 1 \}.
\]

If we write

\[
A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}, \quad P \in M_p(\mathbb{R}), \quad Q \in M_q(\mathbb{R})
\]

then it is shown in O’Neill [138] (Chapter 9, Lemma 6) that the connected component \( \text{SO}_0(p, q) \) of \( \text{SO}(p, q) \) containing \( I \) is given by

\[
\text{SO}_0(p, q) = \{ A \in \text{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}, \ \det(P) > 0, \ \det(S) > 0 \}.
\]

For both \( \text{SO}(p, q) \) and \( \text{SO}_0(p, q) \), the inverse is given by

\[
A^{-1} = I_{p,q} A^\top I_{p,q}.
\]

This implies that the map \( \sigma : \text{SO}_0(p, q) \to \text{SO}_0(p, q) \) given by

\[
\sigma(A) = I_{p,q} A I_{p,q} = (A^\top)^{-1}
\]

is an involution, and its fixed subgroup \( G^\sigma \) is given by

\[
G^\sigma = \left\{ \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \mid Q \in \text{SO}(p), R \in \text{SO}(q) \right\}.
\]

Thus \( G^\sigma \) is isomorphic to \( \text{SO}(p) \times \text{SO}(q) \).

For \( p, q \geq 1 \), the Lie algebra \( \mathfrak{s}\mathfrak{o}(p, q) \) of \( \text{SO}_0(p, q) \) (and \( \text{SO}(p, q) \) as well) is given by

\[
\mathfrak{s}\mathfrak{o}(p, q) = \left\{ \begin{pmatrix} B & A \\ A^\top & C \end{pmatrix} \mid B \in \mathfrak{s}\mathfrak{o}(p), C \in \mathfrak{s}\mathfrak{o}(q), A \in M_{p,q}(\mathbb{R}) \right\}.
\]

Since \( \theta = d\sigma \) is also given by \( \theta(X) = I_{p,q} X I_{p,q} \), we find that the +1-eigenspace \( \mathfrak{k} \) of \( \theta \) is given by

\[
\mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix} \mid B \in \mathfrak{s}\mathfrak{o}(p), C \in \mathfrak{s}\mathfrak{o}(q) \right\}.
\]
and the $-1$-eigenspace $m$ of $\theta$ is given by

$$m = \left\{ \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix} \mid A \in M_{p,q}(\mathbb{R}) \right\}.$$

Note that $\mathfrak{k}$ is a subalgebra of $\mathfrak{so}(p, q)$ and $\mathfrak{so}(p, q) = \mathfrak{k} \oplus m$.

Write $G = SO_o(p, q)$ and $K = SO(p) \times SO(q)$. We define an $Ad(K)$-invariant inner product on $\mathfrak{so}(p, q)$ by

$$\langle X, Y \rangle = \frac{1}{2} \text{tr}(X^\top Y).$$

Therefore, for $p, q \geq 1$, the coset space $SO_0(p, q)/(SO(p) \times SO(q))$ is a symmetric space. Observe that on $m$, the above inner product is given by

$$\langle X, Y \rangle = -\frac{1}{2} \text{tr}(XY).$$

On the other hand, in the case of $SO(p + q)/(SO(p) \times SO(q))$, on $m$, the inner product is given by

$$\langle X, Y \rangle = \frac{1}{2} \text{tr}(XY).$$

This space can be described explicitly. Indeed, let $G^*(q, p+q)$ be the set of $q$-dimensional subspaces $W$ of $R^n = R^{p+q}$ such that $\Phi_{p,q}$ is negative definite on $W$. Then, we have an obvious matrix multiplication action of $SO_0(p, q)$ on $G^*(q, p+q)$, and it is easy to check that this action is transitive. It is not hard to show that the stabilizer of the subspace spanned by the last $q$ columns of the $(p+q) \times (p+q)$ identity matrix is $SO(p) \times SO(q)$, so the space $G^*(q, p+q)$ is isomorphic to the homogeneous (symmetric) space $SO_0(p, q)/(SO(p) \times SO(q))$. The space $G^*(q, p+q)$ is called the hyperbolic Grassmannian.

Assume that $p + q \geq 3, p, q \geq 1$. Then, it can be shown that the Killing form on $\mathfrak{so}(p, q)$ is given by

$$B(X, Y) = (p + q - 2)\text{tr}(XY),$$

so $\mathfrak{so}(p, q)$ is semisimple. If we write

$$X = \begin{pmatrix} B_1 & A_1 \\ A_1^\top & C_1 \end{pmatrix}, \quad Y = \begin{pmatrix} B_2 & A_2 \\ A_2^\top & C_2 \end{pmatrix},$$

then

$$B(X, Y) = (p + q - 2)(\text{tr}(B_1B_2) + \text{tr}(C_1C_2)) + 2(p + q - 2)A_1^\top A_2.$$

Consequently, $B$ is negative definite on $\mathfrak{k}$ and positive definite on $m$, so $G^*(q, p+q) = SO_0(p, q)/(SO(p) \times SO(q))$ is another example of a symmetric space of noncompact type.

We leave it to the reader to compute $\langle [[[X, Y], X], Y \rangle$, where

$$X = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & B \\ B^\top & 0 \end{pmatrix},$$
and check that
\[
\langle[[X,Y],X],Y\rangle = -(BA^\top - AB^\top, BA^\top - AB^\top) - \langle A^\top B - B^\top A, A^\top B - B^\top A\rangle,
\]
which shows that the sectional curvature is nonpositive. In fact, the above expression is the negative of the expression that we found for the sectional curvature of \(G^0(p,p+q)\). When \(p = 1\) or \(q = 1\), we get a space of constant negative curvature.

The above property is one of the consequences of the fact that the space \(G^*(p,q) = \SO_0(p,q)/\SO(p) \times \SO(q)\) is the symmetric space dual to \(G^0(p,p+q) = \SO(p+q)/\SO(p) \times \SO(q)\), the Grassmannian of oriented \(p\)-planes; see O’Neill [138] (Chapter 11, Definition 37) or Helgason [88] (Chapter V, Section 2).

5. Compact Lie Groups

If \(H\) be a compact Lie group, then \(G = H \times H\) acts on \(H\) via
\[
(h_1, h_2) \cdot h = h_1^h h_2^{-1}.
\]
The stabilizer of \((1, 1)\) is clearly \(K = \Delta H = \{(h, h) \mid h \in H\}\). It is easy to see that the map
\[
(g_1, g_2)K \mapsto g_1 g_2^{-1}
\]
is a diffeomorphism between the coset space \(G/K\) and \(H\) (see Helgason [88], Chapter IV, Section 6). A Cartan involution \(\sigma\) is given by
\[
\sigma(h_1, h_2) = (h_2, h_1),
\]
and obviously \(G^\sigma = K = \Delta H\). Therefore, \(H\) appears as the symmetric space \(G/K\), with \(G = H \times H\), \(K = \Delta H\), and
\[
\mathfrak{g} = \{(X, X) \mid X \in \mathfrak{h}\}, \quad \mathfrak{m} = \{(X, -X) \mid X \in \mathfrak{h}\}.
\]
For every \((h_1, h_2) \in \mathfrak{g}\), we have
\[
(h_1, h_2) = \left(\frac{h_1 + h_2}{2}, \frac{h_1 + h_2}{2}\right) + \left(\frac{h_1 - h_2}{2}, -\frac{h_1 - h_2}{2}\right)
\]
which gives the direct sum decomposition
\[
\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}.
\]
The natural projection \(\pi: H \times H \to H\) is given by
\[
\pi(h_1, h_2) = h_1 h_2^{-1},
\]
which yields \(d\pi_{(1,1)}(X,Y) = X - Y\) (see Helgason [88], Chapter IV, Section 6). It follows that the natural isomorphism \(\mathfrak{m} \to \mathfrak{h}\) is given by
\[
(X, -X) \mapsto 2X.
\]
Given any bi-invariant metric $\langle -, - \rangle$ on $H$, define a metric on $m$ by

$$\langle (X, -X), (Y, -Y) \rangle = 4 \langle X, Y \rangle.$$ 

The reader should check that the resulting symmetric space is isometric to $H$ (see Sakai [150], Chapter IV, Exercise 4).

More examples of symmetric spaces are presented in Ziller [180] and Helgason [88]. To close our brief tour of symmetric spaces, we conclude with a short discussion about the type of symmetric spaces.

### 19.10 Types of Symmetric Spaces

Suppose $(G, K, \sigma)$ ($G$ connected and $K$ compact) presents a symmetric space with Cartan involution $\sigma$, and with

$$g = \mathfrak{k} \oplus m,$$

where $\mathfrak{k}$ (the Lie algebra of $K$) is the eigenspace of $d\sigma_1$ associated with the eigenvalue $+1$ and $m$ is the eigenspace associated with the eigenvalue $-1$. If $B$ is the Killing form of $g$, it turns out that the restriction of $B$ to $\mathfrak{k}$ is always negative semidefinite. However, to guarantee that $B$ is negative definite (that is, $B(Z, Z) = 0$ implies that $Z = 0$) some additional condition is needed.

This condition has to do with the subgroup $N$ of $G$ defined by

$$N = \{ g \in G \mid \tau_g = \text{id} \} = \{ g \in G \mid gaK = aK \text{ for all } a \in G \}.$$

Clearly, $N \subseteq K$, and $N$ is a normal subgroup of both $K$ and $G$. It is not hard to show that $N$ is the largest normal subgroup that $K$ and $G$ have in common (see Ziller [180] (Chapter 6, Section 6.2). We can also describe the subgroup $N$ in a more explicit fashion. We have

$$N = \{ g \in G \mid gaK = aK \text{ for all } a \in G \}
= \{ g \in G \mid a^{-1}gaK = K \text{ for all } a \in G \}
= \{ g \in G \mid a^{-1}ga \in K \text{ for all } a \in G \}.$$ 

**Definition 19.11.** For any Lie group $G$ and any closed subgroup $K$ of $G$, the subgroup $N$ of $G$ given by

$$N = \{ g \in G \mid a^{-1}ga \in K \text{ for all } a \in G \}$$

is called the *ineffective kernel* of the left action of $G$ on $G/K$. The left action of $G$ on $G/K$ is said to be *effective* (or *faithful*) if $N = \{1\}$, *almost effective* if $N$ is a discrete subgroup.

If $K$ is compact, which will be assumed from now on, since a discrete subgroup of a compact group is finite, the action of $G$ on $G/K$ is almost effective if $N$ is finite.
19.10. TYPES OF SYMMETRIC SPACES

For example, the action \( \cdot : \text{SU}(n+1) \times \mathbb{C}P^n \to \mathbb{C}P^n \) of \( \text{SU}(n+1) \) on the (complex) projective space \( \mathbb{C}P^n \) discussed in example (e) of Section 5.2 is almost effective but not effective. It presents \( \mathbb{C}P^n \) as the homogenous manifold

\[
\text{SU}(n+1)/S(\text{U}(1) \times \text{U}(n)) \cong \mathbb{C}P^n.
\]

We leave it as an exercise to the reader to prove that the ineffective kernel of the above action is the finite group

\[
N = \{ \lambda I_{n+1} \mid \lambda^{n+1} = 1, \lambda \in \mathbb{C} \}.
\]

It turns out that the additional requirement needed for the Killing form to be negative definite is that the action of \( G \) on \( G/K \) is almost effective.

The following technical proposition gives a criterion for the left action of \( G \) on \( G/K \) to be almost effective in terms of the Lie algebras \( \mathfrak{g} \) and \( \mathfrak{k} \). This is Proposition 6.27 from Ziller [180].

**Proposition 19.28.** The left action of \( G \) on \( G/K \) (with \( K \) compact) is almost effective iff \( \mathfrak{g} \) and \( \mathfrak{k} \) have no nontrivial ideal in common.

**Proof.** By a previous remark, the effective kernel \( N \) of the left action of \( G \) on \( G/K \) is the largest normal subgroup that \( K \) and \( G \) have in common. To say that \( N \) is finite is equivalent to say that \( N \) is discrete (since \( K \) is compact), which is equivalent to the fact that its Lie algebra \( \mathfrak{n} = (0) \). Since by Theorem 15.13 normal subgroups correspond to ideals, the condition that the largest normal subgroup that \( K \) and \( G \) have in common is finite is equivalent to the condition that \( \mathfrak{g} \) and \( \mathfrak{k} \) have no nontrivial ideal in common.

It is natural to classify symmetric spaces depending on the behavior of \( B \) on \( \mathfrak{m} \).

**Proposition 19.29.** Let \( (G,K,\sigma) \) be a symmetric space (\( K \) compact) with Cartan involution \( \sigma \), and assume that the left action of \( G \) on \( G/K \) is almost effective. If \( B \) is the Killing form of \( \mathfrak{g} \) and \( \mathfrak{k} \neq (0) \), then the restriction of \( B \) to \( \mathfrak{k} \) is negative definite.

**Proof.** (After Ziller [180], Proposition 6.38). The restriction of the Ad-representation of \( G \) to \( K \) yields a representation \( \text{Ad}: K \to \text{GL}(\mathfrak{g}) \). Since \( K \) is compact, by Theorem 17.4 there is an Ad(K)-invariant inner product on \( \mathfrak{g} \). Then, for \( k \in K \), we have

\[
\langle \text{Ad}_k(X), \text{Ad}_k(Y) \rangle = \langle X, Y \rangle, \quad \text{for all } X, Y \in \mathfrak{g},
\]

so for \( k = \exp(tZ) \) with \( Z \in \mathfrak{f} \), by taking derivatives at \( t = 0 \), we get

\[
\langle [X,Z], Y \rangle = \langle X, [Z,Y] \rangle, \quad X, Y \in \mathfrak{g}, Z \in \mathfrak{f},
\]

which can be written as

\[
-\langle [Z,X], Y \rangle = \langle [Z,Y], X \rangle, \quad X, Y \in \mathfrak{g}, Z \in \mathfrak{f}.
\]
Consequently \( \text{ad}(Z) \) is a skew-symmetric linear map on \( \mathfrak{g} \) for all \( Z \in \mathfrak{k} \). But then, \( \text{ad}(Z) \) is represented by a skew symmetric matrix \((a_{ij})\) in any orthonormal basis of \( \mathfrak{g} \), and so

\[
B(Z, Z) = \text{tr}(\text{ad}(Z) \circ \text{ad}(Z)) = -\sum_{i,j=1}^{n} a_{ij}^2 \leq 0.
\]

Next, we need to prove that if \( B(Z, Z) = 0 \), then \( Z = 0 \). This is equivalent to proving that if \( \text{ad}(Z) = 0 \) then \( Z = 0 \). However, \( \text{ad}(Z) = 0 \) means that \([Z, X] = 0\) for all \( X \in \mathfrak{g} \), so \( Z \) belongs to the center of \( \mathfrak{g} \), \( \mathfrak{z}(\mathfrak{g}) = \{ Z \in \mathfrak{g} \mid [Z, X] = 0 \text{ for all } X \in \mathfrak{g} \} \).

It is immediately verified that \( \mathfrak{z}(\mathfrak{g}) \) is an ideal of \( \mathfrak{g} \). But now, \( Z \in \mathfrak{z}(\mathfrak{g}) \cap \mathfrak{k} \), which is an ideal of both \( \mathfrak{g} \) and \( \mathfrak{k} \) by definition of \( \mathfrak{z}(\mathfrak{g}) \), and since the left action of \( G \) on \( G/K \) is almost effective, by Proposition 19.28, the Lie algebras \( \mathfrak{g} \) and \( \mathfrak{k} \) have no nontrivial ideal in common, so \( \mathfrak{z}(\mathfrak{g}) \cap \mathfrak{k} = (0) \), and \( Z = 0 \).

**Definition 19.12.** Let \( M = (G, K, \sigma) \) be a symmetric space \((K \text{ compact})\) with Cartan involution \( \sigma \) and Killing form \( B \), and assume that the left action of \( G \) on \( G/K \) is almost effective. The space \( M \) is said to be of

(1) **Euclidean type** if \( B = 0 \) on \( \mathfrak{m} \).

(2) **Compact type** if \( B \) is negative definite on \( \mathfrak{m} \).

(3) **Noncompact type** if \( B \) is positive definite on \( \mathfrak{m} \).

**Proposition 19.30.** Let \( M = (G, K, \sigma) \) be a symmetric space \((K \text{ compact})\) with Cartan involution \( \sigma \) and Killing form \( B \), and assume that the left action of \( G \) on \( G/K \) is almost effective. The following properties hold:

(1) \( M \) is of Euclidean type iff \([\mathfrak{m}, \mathfrak{m}] = (0)\). In this case, \( M \) has zero sectional curvature.

(2) If \( M \) is of compact type, then \( \mathfrak{g} \) is semisimple and both \( G \) and \( M \) are compact.

(3) If \( M \) is of noncompact type, then \( \mathfrak{g} \) is semisimple and both \( G \) and \( M \) are non-compact.

**Proof.** (1) If \( B \) is zero on \( \mathfrak{m} \), since \( B(\mathfrak{m}, \mathfrak{k}) = 0 \) by Proposition 19.24, we conclude that \( \text{rad}(B) = \mathfrak{m} \) (recall that \( \text{rad}(B) = \{ X \in \mathfrak{g} \mid B(X, Y) = 0 \text{ for all } Y \in \mathfrak{g} \} \)). However, \( \text{rad}(B) \) is an ideal in \( \mathfrak{g} \), so \([\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{m} \), and since \([\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k} \), we deduce that

\[
[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{m} \cap \mathfrak{k} = (0).
\]

Conversely, assume that \([\mathfrak{m}, \mathfrak{m}] = (0)\). Since \( B \) is determined by the quadratic form \( Z \mapsto B(Z, Z) \), it suffices to prove that \( B(Z, Z) = 0 \) for all \( Z \in \mathfrak{m} \). Recall that

\[
B(Z, Z) = \text{tr}(\text{ad}(Z) \circ \text{ad}(Z)).
\]
19.10. TYPES OF SYMMETRIC SPACES

We have

$$(\text{ad}(Z) \circ \text{ad}(Z))(X) = [Z, [Z, X]]$$

for all $X \in \mathfrak{g}$. If $X \in \mathfrak{m}$, then $[Z, X] = 0$ since $Z, X \in \mathfrak{m}$ and $[\mathfrak{m}, \mathfrak{m}] = (0)$, and if $X \in \mathfrak{t}$, then $[Z, X] \in [\mathfrak{m}, \mathfrak{t}] \subseteq \mathfrak{m}$, so $[Z, [Z, X]] = 0$, since $[Z, [Z, X]] \in [\mathfrak{m}, \mathfrak{m}] = (0)$. Since $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{t}$, we proved that

$$(\text{ad}(Z) \circ \text{ad}(Z))(X) = 0 \quad \text{for all } X \in \mathfrak{g},$$

and thus $B(Z, Z) = 0$ on $\mathfrak{m}$, as claimed.

For (2) and (3), we use the fact that $B$ is negative definite on $\mathfrak{t}$, by Proposition 19.29.

(2) Since $B$ is negative definite on $\mathfrak{m}$, it is negative definite on $\mathfrak{g}$, and then by Theorem 17.24 we know that $G$ is semisimple and compact. As $K$ is also compact, $M$ is compact.

(3) Since $B$ is positive definite on $\mathfrak{m}$, it is nondegenerate on $\mathfrak{g}$, and then by Theorem 17.22, $G$ is semisimple. In this case, $G$ is not compact since by Theorem 17.24, $G$ is compact iff $B$ is negative definite. As $G$ is noncompact and $K$ is compact, $M$ is noncompact. \(\square\)

Symmetric spaces of Euclidean type are not that interesting, since they have zero sectional curvature. The Grassmannians $G(k, n)$ and $G^0(k, n)$ are symmetric spaces of compact type, and $\text{SL}(n, \mathbb{R})/\text{SO}(n)$ and $\mathcal{H}^+_n(1) = \text{SO}_0(n, 1)/\text{SO}(n)$ are of noncompact type.

Since $\text{GL}^+(n, \mathbb{R})$ is not semisimple, $\text{SPD}(n) \cong \text{GL}^+(n, \mathbb{R})/\text{SO}(n)$ is not a symmetric space of noncompact type, but it has many similar properties. For example, it has nonpositive sectional curvature and because it is diffeomorphic to $S(n) \cong \mathbb{R}^{n(n-1)/2}$, it is simply connected.

Here is a quick summary of the main properties of symmetric spaces of compact and noncompact types. Proofs can be found in O’Neill [138] (Chapter 11) and Ziller [180] (Chapter 6).

**Proposition 19.31.** Let $M = (G, K, \sigma)$ be a symmetric space ($K$ compact) with Cartan involution $\sigma$ and Killing form $B$, and assume that the left action of $G$ on $G/K$ is almost effective. The following properties hold:

1. If $M$ is of compact type, then $M$ has nonnegative sectional curvature and positive Ricci curvature. The fundamental group $\pi_1(M)$ of $M$ is a finite abelian group.

2. If $M$ is of noncompact type, then $M$ is simply connected, and $M$ has nonpositive sectional curvature and negative Ricci curvature. Furthermore, $M$ is diffeomorphic to $\mathbb{R}^n$ (with $n = \dim(M)$) and $G$ is diffeomorphic to $K \times \mathbb{R}^n$.

There is also an interesting duality between symmetric spaces of compact type and noncompact type, but we will not discuss it here. We refer the reader to O’Neill [138] (Chapter 11), Ziller [180] (Chapter 6), and Helgason [88] (Chapter V, Section 2).
We conclude this section by explaining what the Stiefel manifolds $S(k, n)$ are not symmetric spaces for $2 \leq k \leq n - 2$. This has to do with the nature of the involutions of $\mathfrak{so}(n)$. Recall that the matrices $I_{p,q}$ and $J_n$ are defined by

$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}, \quad J_n = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix},$$

with $2 \leq p + q$ and $n \geq 1$. Observe that $I_{p,q}^2 = I_{p+q}$ and $J_n^2 = -I_{2n}$. It is shown in Helgason [88] (Chapter X, Section 2 and Section 5) that, up to conjugation, the only involutive automorphisms of $\mathfrak{so}(n)$ are given by

1. $\theta(X) = I_{p,q}X I_{p,q}$, in which case the eigenspace $\mathfrak{k}$ of $\theta$ associated with the eigenvalue $+1$ is

$$\mathfrak{k}_1 = \left\{ \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix} \Big| S \in \mathfrak{so}(k), T \in \mathfrak{so}(n - k) \right\}.$$

2. $\theta(X) = -J_n X J_n$, in which case the eigenspace $\mathfrak{k}$ of $\theta$ associated with the eigenvalue $+1$ is

$$\mathfrak{k}_2 = \left\{ \begin{pmatrix} S & -T \\ T & S \end{pmatrix} \Big| S \in \mathfrak{so}(n), T \in \mathfrak{S}(n) \right\}.$$

However, in the case of the Stiefel manifold $S(k, n)$, the Lie subalgebra $\mathfrak{k}$ of $\mathfrak{so}(n)$ associated with $\mathfrak{SO}(n - k)$ is

$$\mathfrak{k} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \Big| S \in \mathfrak{so}(n - k) \right\},$$

and if $2 \leq k \leq n - 2$, then $\mathfrak{k} \neq \mathfrak{k}_1$ and $\mathfrak{k} \neq \mathfrak{k}_2$. Therefore, the Stiefel manifold $S(k, n)$ is not a symmetric space if $2 \leq k \leq n - 2$. This also has to do with the fact that in this case, $\mathfrak{SO}(n - k)$ is not a maximal subgroup of $\mathfrak{SO}(n)$. 
Chapter 20

Construction of Manifolds From Gluing Data ✿

20.1 Sets of Gluing Data for Manifolds

The definition of a manifold given in Chapter 7 assumes that the underlying set \( M \) is already known. However, there are situations where we only have some indirect information about the overlap of the domains \( U_i \) of the local charts defining our manifold \( M \) in terms of the transition functions

\[
\varphi_{ji} : \varphi_i(U_i \cap U_j) \rightarrow \varphi_j(U_i \cap U_j),
\]

but where \( M \) itself is not known. For example, this situation happens when trying to construct a surface approximating a 3D-mesh. If we let \( \Omega_{ij} = \varphi_i(U_i \cap U_j) \) and \( \Omega_{ji} = \varphi_j(U_i \cap U_j) \), then \( \varphi_{ji} \) can be viewed as a “gluing map”

\[
\varphi_{ji} : \Omega_{ij} \rightarrow \Omega_{ji}
\]

between two open subets of \( \Omega_i \) and \( \Omega_j \), respectively.

For technical reasons, it is desirable to assume that the images \( \Omega_i = \varphi_i(U_i) \) and \( \Omega_j = \varphi_j(U_j) \) of distinct charts are disjoint, but this can always be achieved for manifolds. Indeed, the map

\[
\beta : (x_1, \ldots, x_n) \mapsto \left( \frac{x_1}{\sqrt{1 + \sum_{i=1}^{n} x_i^2}}, \ldots, \frac{x_n}{\sqrt{1 + \sum_{i=1}^{n} x_i^2}} \right)
\]

is a smooth diffeomorphism from \( \mathbb{R}^n \) to the open unit ball \( B(0,1) \), with inverse given by

\[
\beta^{-1} : (x_1, \ldots, x_n) \mapsto \left( \frac{x_1}{\sqrt{1 - \sum_{i=1}^{n} x_i^2}}, \ldots, \frac{x_n}{\sqrt{1 - \sum_{i=1}^{n} x_i^2}} \right).
\]

Since \( M \) has a countable basis, using compositions of \( \beta \) with suitable translations, we can make sure that the \( \Omega_i \)'s are mapped diffeomorphically to disjoint open subsets of \( \mathbb{R}^n \).
Remarkably, manifolds can be constructed using the “gluing process” alluded to above from what is often called sets of “gluing data.” In this chapter, we are going to describe this construction and prove its correctness in details, provided some mild assumptions on the gluing data. It turns out that this procedure for building manifolds can be made practical. Indeed, it is the basis of a class of new methods for approximating 3D meshes by smooth surfaces, see Siqueira, Xu and Gallier [163].

Some care must be exercised to ensure that the space obtained by gluing the pieces $\Omega_{ij}$ and $\Omega_{ji}$ is Hausdorff. Some care must also be exercised in formulating the consistency conditions relating the $\varphi_{ji}$’s (the so-called “cocycle condition”). This is because the traditional condition (for example, in bundle theory) has to do with triple overlaps of the $U_i = \varphi_i^{-1}(\Omega_i)$ on the manifold $M$ (see Chapter 28, especially Theorem 28.4), but in our situation, we do not have $M$ nor the parametrization maps $\theta_i = \varphi_i^{-1}$, and the cocycle condition on the $\varphi_{ji}$’s has to be stated in terms of the $\Omega_i$’s and the $\Omega_{ji}$’s.

Note that if the $\Omega_{ij}$ arise from the charts of a manifold, then nonempty triple intersections $U_i \cap U_j \cap U_k$ of domains of charts have images $\varphi_i(U_i \cap U_j \cap U_k)$ in $\Omega_i$, $\varphi_j(U_i \cap U_j \cap U_k)$ in $\Omega_j$, and $\varphi_k(U_i \cap U_j \cap U_k)$ in $\Omega_k$, and since the $\varphi_i$’s are bijective maps, we get

$$\varphi_i(U_i \cap U_j \cap U_k) = \varphi_i(U_i \cap U_j \cap U_i \cap U_k) = \varphi_i(U_i \cap U_j) \cap \varphi_i(U_i \cap U_k) = \Omega_{ij} \cap \Omega_{ik},$$

and similarly

$$\varphi_j(U_i \cap U_j \cap U_k) = \Omega_{ji} \cap \Omega_{jk}, \quad \varphi_k(U_i \cap U_j \cap U_k) = \Omega_{ki} \cap \Omega_{kj},$$

and these sets are related. Indeed, we have

$$\varphi_{ji}(\Omega_{ij} \cap \Omega_{ik}) = \varphi_j \circ \varphi_i^{-1}(\varphi_i(U_i \cap U_j) \cap \varphi_i(U_i \cap U_k))$$

$$= \varphi_j(U_i \cap U_j \cap U_k) = \Omega_{ji} \cap \Omega_{jk},$$

and similar equations relating the other “triple intersections.” In particular,

$$\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik},$$

which implies that

$$\varphi_{ij}^{-1}(\Omega_{ji} \cap \Omega_{jk}) = \varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) \subseteq \Omega_{ik}.$$

This is important, because $\varphi_{ij}^{-1}(\Omega_{ji} \cap \Omega_{jk})$ is the domain of $\varphi_{kj} \circ \varphi_{ji}$ and $\Omega_{ik}$ is the domain of $\varphi_{ki}$, so the condition $\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik}$ implies that the domain of $\varphi_{ki}$ is a subset of the domain of $\varphi_{kj} \circ \varphi_{ji}$. The definition of gluing data given by Grimm and Hughes [81, 82] misses the above condition.

Finding an easily testable necessary and sufficient criterion for the Hausdorff condition appears to be a very difficult problem. We propose a necessary and sufficient condition, but it is not easily testable in general. If $M$ is a manifold, then observe that difficulties may arise when we want to separate two distinct point $p, q \in M$ such that $p$ and $q$ neither belong
to the same open $\theta_i(\Omega_i)$, nor to two disjoint opens $\theta_i(\Omega_i)$ and $\theta_j(\Omega_j)$, but instead to the boundary points in $(\partial(\theta_i(\Omega_{ij})) \cap \theta_i(\Omega_i)) \cup (\partial(\theta_j(\Omega_{ji})) \cap \theta_j(\Omega_j))$. In this case, there are some disjoint open subsets $U_p$ and $U_q$ of $M$ with $p \in U_p$ and $q \in U_q$, and we get two disjoint open subsets $V_x = \theta_i^{-1}(U_p) \subseteq \Omega_i$ and $V_y = \theta_j^{-1}(U_q) \subseteq \Omega_j$ with $\theta_i(x) = p$, $\theta_j(y) = q$, and such that $x \in \partial(\Omega_{ij}) \cap \Omega_i$, $y \in \partial(\Omega_{ji}) \cap \Omega_j$, and no point in $V_y \cap \Omega_{ji}$ is the image of any point in $V_x \cap \Omega_{ij}$ by $\phi_{ji}$. Since $V_x$ and $V_y$ are open, we may assume that they are open balls. This necessary condition turns out to be also sufficient.

With the above motivations in mind, here is the definition of sets of gluing data.

**Definition 20.1.** Let $n$ be an integer with $n \geq 1$ and let $k$ be either an integer with $k \geq 1$ or $k = \infty$. A set of gluing data is a triple $\mathcal{G} = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\phi_{ji})_{(i,j) \in K})$ satisfying the following properties, where $I$ is a (nonempty) countable set:

1. For every $i \in I$, the set $\Omega_i$ is a nonempty open subset of $\mathbb{R}^n$ called a parametrization domain, for short, $p$-domain, and the $\Omega_i$ are pairwise disjoint (i.e., $\Omega_i \cap \Omega_j = \emptyset$ for all $i \neq j$).

2. For every pair $(i,j) \in I \times I$, the set $\Omega_{ij}$ is an open subset of $\Omega_i$. Furthermore, $\Omega_{ii} = \Omega_i$ and $\Omega_{ij} \neq \emptyset$ iff $\Omega_{ji} \neq \emptyset$. Each nonempty $\Omega_{ij}$ (with $i \neq j$) is called a gluing domain.

3. If we let

$$K = \{(i,j) \in I \times I \mid \Omega_{ij} \neq \emptyset\},$$

then $\phi_{ji} : \Omega_{ij} \to \Omega_{ji}$ is a $C^k$ bijection for every $(i,j) \in K$ called a transition function (or gluing function) and the following condition holds:

(c) For all $i, j, k$, if $\Omega_{ji} \cap \Omega_{jk} \neq \emptyset$, then $\phi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik}$, and $\phi_{ki}(x) = \phi_{kj} \circ \phi_{ji}(x)$, for all $x \in \Omega_{ij} \cap \Omega_{ik}$.

Condition (c) is called the cocycle condition.

4. For every pair $(i,j) \in K$, with $i \neq j$, for every $x \in \partial(\Omega_{ij}) \cap \Omega_i$ and every $y \in \partial(\Omega_{ji}) \cap \Omega_j$, there are open balls $V_x$ and $V_y$ centered at $x$ and $y$ so that no point of $V_y \cap \Omega_{ji}$ is the image of any point of $V_x \cap \Omega_{ij}$ by $\phi_{ji}$.

**Remarks.**

1. In practical applications, the index set $I$ is of course finite and the open subsets $\Omega_i$ may have special properties (for example, connected; open simplicies, etc.).

2. We are only interested in the $\Omega_{ij}$’s that are nonempty, but empty $\Omega_{ij}$’s do arise in proofs and constructions, and this is why our definition allows them.

3. Observe that $\Omega_{ij} \subseteq \Omega_i$ and $\Omega_{ji} \subseteq \Omega_j$. If $i \neq j$, as $\Omega_i$ and $\Omega_j$ are disjoint, so are $\Omega_{ij}$ and $\Omega_{ji}$.
(4) The cocycle condition (c) may seem overly complicated but it is actually needed to guarantee the transitivity of the relation $\sim$ defined in the proof of Proposition 20.1. Flawed versions of condition (c) appear in the literature; see the discussion after the proof of Proposition 20.1. The problem is that $\varphi_{kj} \circ \varphi_{ji}$ is a partial function whose domain $\varphi_{ij}^{-1}(\Omega_{\ij} \cap \Omega_{\jk})$ is not necessarily related to the domain $\Omega_{\ik}$ of $\varphi_{\ik}$. To ensure transitivity of $\sim$, we must assert that whenever the composition $\varphi_{kj} \circ \varphi_{ji}$ has a nonempty domain, this domain is contained in the domain $\Omega_{\ik}$ of $\varphi_{\ik}$, and that $\varphi_{kj} \circ \varphi_{ji}$ and $\varphi_{\ik}$ agree in $\varphi_{ij}^{-1}(\Omega_{\ij} \cap \Omega_{\jk})$.

Since the $\varphi_{ij}$ are bijections, condition (c) implies the following conditions:

(a) $\varphi_{ii} = \text{id}_{\Omega_i}$, for all $i \in I$.

(b) $\varphi_{ij} = \varphi_{ji}^{-1}$, for all $(i, j) \in K$.

To get (a), set $i = j = k$. Then, (b) follows from (a) and (c) by setting $k = i$.

(5) If $M$ is a $C^k$ manifold (including $k = \infty$), then using the notation of our introduction, it is easy to check that the open sets $\Omega_i$, $\Omega_{ij}$ and the gluing functions $\varphi_{ij}$ satisfy the conditions of Definition 20.1 (provided that we fix the charts so that the images of distinct charts are disjoint). Proposition 20.1 will show that a manifold can be reconstructed from a set of gluing data.

The idea of defining gluing data for manifolds is not new. André Weil introduced this idea to define abstract algebraic varieties by gluing irreducible affine sets in his book [176] published in 1946. The same idea is well-known in bundle theory and can be found in standard texts such as Steenrod [164], Bott and Tu [24], Morita [133] and Wells [178] (the construction of a fibre bundle from a cocycle is given in Chapter 28, see Theorem 28.4).

The beauty of the idea is that it allows the reconstruction of a manifold $M$ without having prior knowledge of the topology of this manifold (that is, without having explicitly the underlying topological space $M$) by gluing open subsets of $\mathbb{R}^n$ (the $\Omega_i$'s) according to prescribed gluing instructions (namely, glue $\Omega_i$ and $\Omega_j$ by identifying $\Omega_{ij}$ and $\Omega_{ji}$ using $\varphi_{ij}$). This method of specifying a manifold separates clearly the local structure of the manifold (given by the $\Omega_i$'s) from its global structure which is specified by the gluing functions. Furthermore, this method ensures that the resulting manifold is $C^k$ (even for $k = \infty$) with no extra effort since the gluing functions $\varphi_{ij}$ are assumed to be $C^k$.

Grimm and Hughes [81, 82] appear to be the first to have realized the power of this latter property for practical applications, and we wish to emphasize that this is a very significant discovery. However, Grimm [81] uses a condition stronger than our condition (4) to ensure that the resulting space is Hausdorff. The cocycle condition in Grimm and Hughes [81, 82] is also not strong enough to ensure transitivity of the relation $\sim$. We will come back to these points after the proof of Proposition 20.1.

Working with overlaps of open subsets of the parameter domain makes it much easier to enforce smoothness conditions compared to the traditional approach with splines where the
parameter domain is subdivided into *closed* regions, and where enforcing smoothness along boundaries is much more difficult.

Let us show that a set of gluing data defines a $C^k$ manifold in a natural way.

**Proposition 20.1.** For every set of gluing data $G = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})$, there is an $n$-dimensional $C^k$ manifold $M_G$ whose transition functions are the $\varphi_{ji}$’s.

**Proof.** Define the binary relation $\sim$ on the disjoint union $\bigsqcup_{i \in I} \Omega_i$ of the open sets $\Omega_i$ as follows: For all $x, y \in \bigsqcup_{i \in I} \Omega_i$,

$$x \sim y \iff (\exists (i, j) \in K)(x \in \Omega_{ij}, y \in \Omega_{ji}, y = \varphi_{ji}(x)).$$

Note that if $x \sim y$ and $x \neq y$, then $i \neq j$, as $\varphi_{ii} = \text{id}$. But then, as $x \in \Omega_{ij} \subseteq \Omega_i$,

$$y \in \Omega_{ji} \subseteq \Omega_j$$

and $\Omega_i \cap \Omega_j = \emptyset$ when $i \neq j$, if $x \sim y$ and $x, y \in \Omega_i$, then $x = y$. We claim that $\sim$ is an equivalence relation. This follows easily from the cocycle condition. Clearly, condition 3a of Definition 20.1 ensures reflexivity, while condition 3b ensures symmetry. To check transitivity, assume that $x \sim y$ and $y \sim z$. Then, there are some $i, j, k$ such that (i) $x \in \Omega_{ij}$, $y \in \Omega_{ji} \cap \Omega_{jk}$, $z \in \Omega_{kj}$, and (ii) $y = \varphi_{ji}(x)$ and $z = \varphi_{kj}(y)$. Consequently, $\Omega_{ji} \cap \Omega_{jk} \neq \emptyset$ and $x \in \varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$, so by 3c, we get $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik} \subseteq \Omega_{ik}$. So, $\varphi_{ki}(z)$ is defined and by 3c again, $\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x) = z$, i.e., $x \sim z$, as desired.

Since $\sim$ is an equivalence relation, let

$$M_G = \left(\bigsqcup_{i \in I} \Omega_i\right) / \sim$$

be the quotient set and let $p: \bigsqcup_{i \in I} \Omega_i \to M_G$ be the quotient map, with $p(x) = [x]$, where $[x]$ denotes the equivalence class of $x$. Also, for every $i \in I$, let $i_i: \Omega_i \to \bigsqcup_{i \in I} \Omega_i$ be the natural injection and let

$$\tau_i = p \circ i_i: \Omega_i \to M_G.$$

Since we already noted that if $x \sim y$ and $x, y \in \Omega_i$, then $x = y$, we can conclude that every $\tau_i$ is injective. We give $M_G$ the coarsest topology that makes the bijections, $\tau_i: \Omega_i \to \tau_i(\Omega_i)$, into homeomorphisms. Then, if we let $U_i = \tau_i(\Omega_i)$ and $\varphi_i = \tau_i^{-1}$, it is immediately verified that the $(U_i, \varphi_i)$ are charts and that this collection of charts forms a $C^k$ atlas for $M_G$. As there are countably many charts, $M_G$ is second-countable.

To prove that the topology is Hausdorff, we first prove the following:

*Claim.* For all $(i, j) \in I \times I$, we have $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i, j) \in K$ and if so,

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

Assume that $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ and let $[z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$. Observe that $[z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$ iff $z \sim x$ and $z \sim y$, for some $x \in \Omega_i$ and some $y \in \Omega_j$. Consequently, $x \sim y$, which implies that $(i, j) \in K$, $x \in \Omega_{ij}$ and $y \in \Omega_{ji}$. We have $[z] \in \tau_i(\Omega_{ij})$ iff $z \sim x$, for some $x \in \Omega_{ij}$. Then,
either \( i = j \) and \( z = x \) or \( i \neq j \) and \( z \in \Omega_{ji} \), which shows that \( [z] \in \tau_i(\Omega_{ji}) \), and consequently we get \( \tau_i(\Omega_{ij}) \subseteq \tau_j(\Omega_{ji}) \). Since the same argument applies by interchanging \( i \) and \( j \), we have that \( \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \), for all \((i, j) \in K\). Furthermore, because \( \Omega_{ij} \subseteq \Omega_i \), \( \Omega_{ji} \subseteq \Omega_j \), and \( \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \), for all \((i, j) \in K\), we also have that \( \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \subseteq \tau_i(\Omega_i) \cap \tau_j(\Omega_j) \), for all \((i, j) \in K\).

For the reverse inclusion, if \( [z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j) \), then we know that there is some \( x \in \Omega_{ij} \) and some \( y \in \Omega_{ji} \) such that \( z \sim x \) and \( z \sim y \), so \( [z] = [x] \in \tau_i(\Omega_{ij}) \) and \( [z] = [y] \in \tau_j(\Omega_{ji}) \), and then we get

\[
\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \subseteq \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) .
\]

This proves that if \( \tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset \), then \((i, j) \in K\) and

\[
\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) .
\]

Finally, assume that \((i, j) \in K\). Then, for any \( x \in \Omega_{ij} \subseteq \Omega_i \), we have \( y = \varphi_{ji}(x) \in \Omega_{ji} \subseteq \Omega_j \) and \( x \sim y \), so that \( \tau_i(x) = \tau_j(y) \), which proves that \( \tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset \). So, our claim is true, and we can use it.

We now prove that the topology of \( M_G \) is Hausdorff. Pick \([x], [y] \in M_G\) with \([x] \neq [y]\), for some \( x \in \Omega_i \) and some \( y \in \Omega_j \). Either \( \tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \emptyset \), in which case, as \( \tau_i \) and \( \tau_j \) are homeomorphisms, \([x]\) and \([y]\) belong to the two disjoint open sets \( \tau_i(\Omega_i) \) and \( \tau_j(\Omega_j) \). If not, then by the Claim, \((i, j) \in K\) and

\[
\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) .
\]

There are several cases to consider:

1. If \( i = j \) then \( x \) and \( y \) can be separated by disjoint opens \( V_x \) and \( V_y \), and as \( \tau_i \) is a homeomorphism, \([x]\) and \([y]\) are separated by the disjoint open subsets \( \tau_i(V_x) \) and \( \tau_j(V_y) \).

2. If \( i \neq j \), \( x \in \Omega_i - \overline{\Omega_{ji}} \) and \( y \in \Omega_j - \overline{\Omega_{ij}} \), then \( \tau_i(\Omega_i - \overline{\Omega_{ij}}) \) and \( \tau_j(\Omega_j - \overline{\Omega_{ij}}) \) are disjoint open subsets separating \([x]\) and \([y]\), where \( \Omega_{ij} \) and \( \Omega_{ji} \) are the closures of \( \Omega_{ij} \) and \( \Omega_{ji} \), respectively.

3. If \( i \neq j \), \( x \in \Omega_{ij} \) and \( y \in \Omega_{ji} \), as \([x] \neq [y]\) and \( x \sim \varphi_{ij}(y) \), then \( x \neq \varphi_{ij}(y) \). We can separate \( x \) and \( \varphi_{ij}(y) \) by disjoint open subsets \( V_x \) and \( V_y \), and \([x]\) and \([y] = [\varphi_{ij}(y)]\) are separated by the disjoint open subsets \( \tau_i(V_x) \) and \( \tau_i(V_{\varphi_{ij}(y)}) \).

4. If \( i \neq j \), \( x \in \partial(\Omega_{ij}) \cap \Omega_i \) and \( y \in \partial(\Omega_{ji}) \cap \Omega_j \), then we use condition 4 of Definition 20.1. This condition yields two disjoint open subsets \( V_x \) and \( V_y \) with \( x \in V_x \) and \( y \in V_y \), such that no point of \( V_x \cap \Omega_{ij} \) is equivalent to any point of \( V_y \cap \Omega_{ji} \), and so \( \tau_i(V_x) \) and \( \tau_j(V_y) \) are disjoint open subsets separating \([x]\) and \([y]\).

Therefore, the topology of \( M_G \) is Hausdorff and \( M_G \) is indeed a manifold. Finally, it is trivial to verify that the transition maps of \( M_G \) are the original gluing functions \( \varphi_{ij} \), since \( \varphi_i = \tau_i^{-1} \) and \( \varphi_{ji} = \varphi_j \circ \varphi_i^{-1} \).

\[ \square \]
20.1. SETS OF GLUING DATA FOR MANIFOLDS

It should be noted that as nice as it is, Proposition 20.1 is a theoretical construction that yields an “abstract” manifold, but does not yield any information as to the geometry of this manifold. Furthermore, the resulting manifold may not be orientable or compact, even if we start with a finite set of $p$-domains.

Here is an example showing that if condition (4) of Definition 20.1 is omitted then we may get non-Hausdorff spaces. Cindy Grimm uses a similar example in her dissertation [81] (Appendix C2, page 126), but her presentation is somewhat confusing because her $\Omega$ appear to be two disjoint copies of the real line in $\mathbb{R}^2$, but these are not open in $\mathbb{R}^2$!

Let $\Omega_1 = (-3, -1), \Omega_2 = (1, 3), \Omega_{12} = (-3, -2)$, $\Omega_{21} = (1, 2)$ and $\varphi_{21}(x) = x + 4$. The resulting space $M$ is a curve looking like a “fork,” and the problem is that the images of $-2$ and 2 in $M$, which are distinct points of $M$, cannot be separated. Indeed, the images of any two open intervals $(-2 - \epsilon, -2 + \epsilon)$ and $(2 - \eta, 2 + \eta)$ (for $\epsilon, \eta > 0$) always intersect, since $(-2 - \min(\epsilon, \eta), -2)$ and $(2 - \min(\epsilon, \eta), 2)$ are identified. Clearly, condition (4) fails.

Cindy Grimm [81] (page 40) uses a condition stronger than our condition (4) to ensure that the quotient, $M_G$ is Hausdorff; namely, that for all $(i, j) \in K$ with $i \neq j$, the quotient $(\Omega_i \sqcup \Omega_j)/ \sim$ should be embeddable in $\mathbb{R}^n$. This is a rather strong condition that prevents obtaining a 2-sphere by gluing two open discs in $\mathbb{R}^2$ along an annulus (see Grimm [81], Appendix C2, page 126).

Grimm uses the following cocycle condition in [81] (page 40) and [82] (page 361):

$$(c') \text{ For all } x \in \Omega_{ij} \cap \Omega_{ik}, \quad \varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x).$$

This condition is not strong enough to imply transitivity of the relation $\sim$, as shown by the following counter-example:

Let $\Omega_1 = (0, 3), \Omega_2 = (4, 5), \Omega_3 = (6, 9), \Omega_{12} = (0, 1), \Omega_{13} = (2, 3), \Omega_{21} = \Omega_{23} = (4, 5), \Omega_{32} = (8, 9), \Omega_{31} = (6, 7), \varphi_{21}(x) = x + 4, \varphi_{32}(x) = x + 4$ and $\varphi_{31}(x) = x + 4$.

Note that the pairwise gluings yield Hausdorff spaces. Obviously, $\varphi_{32} \circ \varphi_{21}(x) = x + 8$, for all $x \in \Omega_{12}$, but $\Omega_{12} \cap \Omega_{13} = \emptyset$. Thus, $0.5 \sim 4.5 \sim 8.5$, and if the relation $\sim$ was transitive, then we would conclude that $0.5 \sim 8.5$. However, the definition of the relation $\sim$ requires that $\varphi_{31}(0.5)$ be defined, which is not the case. Therefore, the relation $\sim$ is not transitive. The problem is that because $\Omega_{12} \cap \Omega_{13} = \emptyset$, condition $(c')$ holds vacuously, but it is not strong enough to ensure that $\varphi_{31}(0.5)$ is defined.

Here is another counter-example in which $\Omega_{12} \cap \Omega_{13} \neq \emptyset$, using a disconnected open $\Omega_2$.

Let $\Omega_1 = (0, 3), \Omega_2 = (4, 5) \cup (6, 7), \Omega_3 = (8, 11), \Omega_{12} = (0, 1) \cup (2, 3), \Omega_{13} = (2, 3), \Omega_{21} = \Omega_{23} = (4, 5) \cup (6, 7), \Omega_{32} = (8, 9) \cup (10, 11), \Omega_{31} = (8, 9), \varphi_{21}(x) = x + 4, \varphi_{32}(x) = x + 2$ on $(6, 7), \varphi_{32}(x) = x + 6$ on $(4, 5), \varphi_{31}(x) = x + 6$.

Note that the pairwise gluings yield Hausdorff spaces. Obviously, $\varphi_{32} \circ \varphi_{21}(x) = x + 6 = \varphi_{31}(x)$ for all $x \in \Omega_{12} \cap \Omega_{13} = (2, 3)$. Thus, $0.5 \sim 4.5 \sim 8.5$, but $0.5 \not\sim 8.5$ since $\varphi_{31}(0.5)$ is
undefined. This time, condition \((c')\) holds and is nontrivial since \(\Omega_{12} \cap \Omega_{13} = (2, 3)\), but it is not strong enough to ensure that \(\varphi_{31}(0.5)\) is defined.

It is possible to give a construction, in the case of a surface, which builds a compact manifold whose geometry is “close” to the geometry of a prescribed 3D-mesh (see Siqueira, Xu and Gallier [163]). Actually, we are not able to guarantee, in general, that the parametrization functions \(\theta_i\) that we obtain are injective, but we are not aware of any algorithm that achieves this.

Given a set of gluing data, \(\mathcal{G} = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ij})_{(i,j) \in K})\), it is natural to consider the collection of manifolds \(M\) parametrized by maps \(\theta_i: \Omega_i \to M\) whose domains are the \(\Omega_i\)'s and whose transitions functions are given by the \(\varphi_{ji}\); that is, such that

\[
\varphi_{ji} = \theta_j^{-1} \circ \theta_i.
\]

We will say that such manifolds are induced by the set of gluing data \(\mathcal{G}\).

The proof of Proposition 20.1 shows that the parametrization maps \(\tau_i\) satisfy the property: \(\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset\) iff \((i, j) \in K\), and if so

\[
\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).
\]

Furthermore, they also satisfy the consistency condition:

\[
\tau_i = \tau_j \circ \varphi_{ji},
\]

for all \((i, j) \in K\). If \(M\) is a manifold induced by the set of gluing data \(\mathcal{G}\), because the \(\theta_i\)'s are injective and \(\varphi_{ji} = \theta_j^{-1} \circ \theta_i\), the two properties stated above for the \(\tau_i\)'s also hold for the \(\theta_i\)'s. We will see in Section 20.2 that the manifold \(M_\mathcal{G}\) is a “universal” manifold induced by \(\mathcal{G}\), in the sense that every manifold induced by \(\mathcal{G}\) is the image of \(M_\mathcal{G}\) by some \(C^k\) map.

Interestingly, it is possible to characterize when two manifolds induced by the same set of gluing data are isomorphic in terms of a condition on their transition functions.

**Proposition 20.2.** Given any set of gluing data \(\mathcal{G} = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ij})_{(i,j) \in K})\), for any two manifolds \(M\) and \(M'\) induced by \(\mathcal{G}\) given by families of parametrizations \((\Omega_i, \theta_i)_{i \in I}\) and \((\Omega_i, \theta_i')_{i \in I}\) respectively, if \(f: M \to M'\) is a \(C^k\) isomorphism, then there are \(C^k\) bijections \(\rho_i: W_{ij} \to W'_{ij}\) for some open subsets \(W_{ij}, W'_{ij} \subseteq \Omega_i\), such that

\[
\varphi'_{ji}(x) = \rho_j \circ \varphi_{ji} \circ \rho_i^{-1}(x), \quad \text{for all } x \in W_{ij},
\]

with \(\varphi_{ji} = \theta_j^{-1} \circ \theta_i\) and \(\varphi'_j = \theta_j'^{-1} \circ \theta_i'.\) Furthermore, \(\rho_i = (\theta_i'^{-1} \circ f \circ \theta_i) \mid W_{ij}\), and if \(\theta_i'^{-1} \circ f \circ \theta_i\) is a bijection from \(\Omega_i\) to itself and \(\theta_i'^{-1} \circ f \circ \theta_i(\Omega_{ij}) = \Omega_{ij}\), for all \(i, j\), then \(W_{ij} = W'_{ij} = \Omega_i\).

**Proof.** The composition \(\theta_i'^{-1} \circ f \circ \theta_i\) is actually a partial function with domain

\[
\text{dom}(\theta_i'^{-1} \circ f \circ \theta_i) = \{ x \in \Omega_i \mid \theta_i(x) \in f^{-1} \circ \theta_i'(\Omega_i) \},
\]
and its "inverse" $\theta_i^{-1} \circ f^{-1} \circ \theta_i'$ is a partial function with domain

$$\text{dom}(\theta_i^{-1} \circ f^{-1} \circ \theta_i') = \{ x \in \Omega_i \mid \theta_i'(x) \in \theta_i(\Omega_i) \}.$$  

The composition $\theta_j^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i'$ is also a partial function, and we let

$$W_{ij} = \Omega_{ij} \cap \text{dom}(\theta_j^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i'), \quad \rho_i = (\theta_i^{-1} \circ f \circ \theta_i) \upharpoonright W_{ij}$$

and $W'_{ij} = \rho_i(W_{ij})$. Observe that $\theta_j \circ \varphi_{ji} = \theta_j \circ \theta_j^{-1} \circ \theta_i = \theta_i$, that is,

$$\theta_i = \theta_j \circ \varphi_{ji}.$$  

Using this, on $W_{ij}$ we get

$$\rho_j \circ \varphi_{ji} \circ \rho_i^{-1} = \theta_j^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ (\theta_i^{-1} \circ f \circ \theta_i)^{-1}$$

$$= \theta_j^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i'$$

$$= \theta_j^{-1} \circ f \circ \theta_i \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i'$$

$$= \theta_i^{-1} \circ \theta_i = \varphi_{ji},$$

as claimed. The last part of the proposition is clear. \hfill $\square$

Proposition 20.2 suggests defining a notion of equivalence on sets of gluing data which yields a converse of this proposition.

**Definition 20.2.** Two sets of gluing data $G = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}(\varphi_{ji})_{(i,j) \in K})$ and $G' = ((\Omega_i)_{i \in I}, (\Omega'_{ij})_{(i,j) \in I \times I}(\varphi'_{ji})_{(i,j) \in K})$ over the same sets of $\Omega_i$'s and $\Omega_{ij}$'s are equivalent iff there is a family of $C^k$ bijections $(\rho_i : \Omega_i \to \Omega_i)_{i \in I}$, such that $\rho_i(\Omega_{ij}) = \Omega'_{ij}$ and

$$\varphi'_{ji}(x) = \rho_j \circ \varphi_{ji} \circ \rho_i^{-1}(x), \quad \text{for all } x \in \Omega_{ij},$$

for all $i, j$.

Here is the converse of Proposition 20.2. It is actually nicer than Proposition 20.2, because we can take $W_{ij} = W'_{ij} = \Omega_i$.

**Proposition 20.3.** If two sets of gluing data $G = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}(\varphi_{ji})_{(i,j) \in K})$ and $G' = ((\Omega_i)_{i \in I}, (\Omega'_{ij})_{(i,j) \in I \times I}(\varphi'_{ji})_{(i,j) \in K})$ are equivalent, then there is a $C^k$ isomorphism $f : M_G \to M_{G'}$ between the manifolds induced by $G$ and $G'$. Furthermore, $f \circ \tau_i = \tau'_i \circ \rho_i$, for all $i \in I$.

**Proof.** Let $f_i : \tau_i(\Omega_i) \to \tau'_i(\Omega_i)$ be the $C^k$ bijection given by

$$f_i = \tau'_i \circ \rho_i \circ \tau_i^{-1},$$
where the $\rho_i : \Omega_i \rightarrow \Omega_i$'s are the maps giving the equivalence of $\mathcal{G}$ and $\mathcal{G}'$. If we prove that $f_i$ and $f_j$ agree on the overlap $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, then the $f_i$ patch and yield a $C^k$ isomorphism $f : M_G \rightarrow M_{G'}$. The conditions of Proposition 20.3 imply that

$$\varphi_{ji} \circ \rho_i = \rho_j \circ \varphi_{ji},$$

and we know that

$$\tau_i' = \tau_j' \circ \varphi_{ji}. \tag{20.3}$$

Consequently, for every $[x] \in \tau_j(\Omega_{ji}) = \tau_i(\Omega_{ij})$ with $x \in \Omega_{ij}$, we have

$$f_j([x]) = \tau_j' \circ \rho_j \circ \tau_j^{-1}([x]) = \tau_j' \circ \rho_j \circ \tau_j^{-1}(\varphi_{ji}(x)) = \tau_j' \circ \varphi_{ji} \circ \rho_i(x) = \tau_i' \circ \rho_i(x) = \tau_i' \circ \rho_i \circ \tau_i^{-1}([x]) = f_i([x]),$$

which shows that $f_i$ and $f_j$ agree on $\tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, as claimed. \hfill $\square$

In the next section, we describe a class of spaces that can be defined by gluing data and parametrization functions $\theta_i$ that are not necessarily injective. Roughly speaking, the gluing data specify the topology and the parametrizations define the geometry of the space. Such spaces have more structure than spaces defined parametrically but they are not quite manifolds. Yet, they arise naturally in practice and they are the basis of efficient implementations of very good approximations of 3D meshes.

### 20.2 Parametric Pseudo-Manifolds

In practice, it is often desirable to specify some $n$-dimensional geometric shape as a subset of $\mathbb{R}^d$ (usually for $d = 3$) in terms of parametrizations which are functions $\theta_i$ from some subset of $\mathbb{R}^n$ into $\mathbb{R}^d$ (usually, $n = 2$). For “open” shapes, this is reasonably well understood, but dealing with a “closed” shape is a lot more difficult because the parametrized pieces should overlap as smoothly as possible, and this is hard to achieve. Furthermore, in practice, the parametrization functions $\theta_i$ may not be injective. Proposition 20.1 suggests various ways of defining such geometric shapes. For the lack of a better term, we will call these shapes, parametric pseudo-manifolds.

**Definition 20.3.** Let $n, k, d$ be three integers with $d > n \geq 1$ and $k \geq 1$ or $k = \infty$. A parametric $C^k$ pseudo-manifold of dimension $n$ in $\mathbb{R}^d$ is a pair $\mathcal{M} = (\mathcal{G}, (\theta_i)_{i \in I})$, where $\mathcal{G} = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})$ is a set of gluing data for some finite set $I$, and each $\theta_i$ is a $C^k$ function $\theta_i : \Omega_i \rightarrow \mathbb{R}^d$ called a parametrization, such that the following property holds:
(C) For all \((i, j) \in K\), we have
\[ \theta_i = \theta_j \circ \varphi_{ji}. \]

For short, we use terminology \textit{parametric pseudo-manifold}. The subset \(M \subseteq \mathbb{R}^d\) given by
\[ M = \bigcup_{i \in I} \theta_i(\Omega_i) \]
is called the \textit{image} of the parametric pseudo-manifold \(M\). When \(n = 2\) and \(d = 3\), we say that \(M\) is a \textit{parametric pseudo-surface}.

Condition (C) obviously implies that
\[ \theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji}), \]
for all \((i, j) \in K\). Consequently, \(\theta_i\) and \(\theta_j\) are consistent parametrizations of the overlap \(\theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji})\). Thus, the shape \(M\) is covered by pieces \(U_i = \theta_i(\Omega_i)\) not necessarily open, with each \(U_i\) parametrized by \(\theta_i\), and where the overlapping pieces \(U_i \cap U_j\), are parametrized consistently. The local structure of \(M\) is given by the \(\theta_i\)'s, and the global structure is given by the gluing data. We recover a manifold if we require the \(\theta_i\) to be bijective and to satisfy the following additional conditions:

(C') For all \((i, j) \in K\),
\[ \theta_i(\Omega_i) \cap \theta_j(\Omega_j) = \theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji}). \]

(C'') For all \((i, j) \notin K\),
\[ \theta_i(\Omega_i) \cap \theta_j(\Omega_j) = \emptyset. \]

Even if the \(\theta_i\)'s are not injective, properties (C') and (C'') would be desirable since they guarantee that \(\theta_i(\Omega_i - \Omega_{ij})\) and \(\theta_j(\Omega_j - \Omega_{ji})\) are parametrized uniquely. Unfortunately, these properties are difficult to enforce. Observe that any manifold induced by \(\mathcal{G}\) is the image of a parametric pseudo-manifold.

Although this is an abuse of language, it is more convenient to call \(M\) a parametric pseudo-manifold, or even a \textit{pseudo-manifold}.

We can also show that the parametric pseudo-manifold \(M\) is the image in \(\mathbb{R}^d\) of the abstract manifold \(M_\mathcal{G}\).

**Proposition 20.4.** Let \(\mathcal{M} = (\mathcal{G}, (\theta_i)_{i \in I})\) be parametric \(C^k\) pseudo-manifold of dimension \(n\) in \(\mathbb{R}^d\), where \(\mathcal{G} = ((\Omega_i)_{i \in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})\) is a set of gluing data for some finite set \(I\). Then, the parametrization maps \(\theta_i\) induce a surjective map \(\Theta: M_\mathcal{G} \to M\) from the abstract manifold \(M_\mathcal{G}\) specified by \(\mathcal{G}\) to the image \(M \subseteq \mathbb{R}^d\) of the parametric pseudo-manifold \(\mathcal{M}\), and the following property holds: For every \(\Omega_i\),
\[ \theta_i = \Theta \circ \tau_i, \]
where the \(\tau_i: \Omega_i \to M_\mathcal{G}\) are the parametrization maps of the manifold \(M_\mathcal{G}\) (see Proposition 20.1). In particular, every manifold \(M\) induced by the gluing data \(\mathcal{G}\) is the image of \(M_\mathcal{G}\) by a map \(\Theta: M_\mathcal{G} \to M\).
Proof. Recall that 

\[ M_G = \left( \coprod_{i \in I} \Omega_i \right) / \sim, \]

where \( \sim \) is the equivalence relation defined so that, for all \( x, y \in \coprod_{i \in I} \Omega_i \),

\[ x \sim y \iff (\exists (i, j) \in K)(x \in \Omega_{ij}, y \in \Omega_{ji}, y = \varphi_{ji}(x)). \]

The proof of Proposition 20.1 also showed that \( \tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset \) iff \( (i, j) \in K \), and so,

\[ \tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}). \]

In particular,

\[ \tau_i(\Omega_i - \Omega_{ij}) \cap \tau_j(\Omega_j - \Omega_{ji}) = \emptyset \]

for all \( (i, j) \in I \times I \) \( (\Omega_{ij} = \Omega_{ji} = \emptyset \) when \( (i, j) \notin K \). These properties with the fact that the \( \tau_i \)’s are injections show that for all \( (i, j) \notin K \), we can define \( \Theta_i : \tau_i(\Omega_i) \to \mathbb{R}^d \) and \( \Theta_j : \tau_j(\Omega_j) \to \mathbb{R}^d \) by

\[ \Theta_i([x]) = \theta_i(x), \quad x \in \Omega_i \quad \Theta_j([y]) = \theta_j(y), \quad y \in \Omega_j. \]

For \( (i, j) \in K \), as the the \( \tau_i \)’s are injections we can define \( \Theta_i : \tau_i(\Omega_i - \Omega_{ij}) \to \mathbb{R}^d \) and \( \Theta_j : \tau_j(\Omega_j - \Omega_{ji}) \to \mathbb{R}^d \) by

\[ \Theta_i([x]) = \theta_i(x), \quad x \in \Omega_i - \Omega_{ij} \quad \Theta_j([y]) = \theta_j(y), \quad y \in \Omega_j - \Omega_{ji}. \]

It remains to define \( \Theta_i \) on \( \tau_i(\Omega_{ij}) \) and \( \Theta_j \) on \( \tau_j(\Omega_{ji}) \) in such a way that they agree on \( \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \). However, condition (C) in Definition 20.3 says that for all \( x \in \Omega_{ij} \),

\[ \theta_i(x) = \theta_j(\varphi_{ji}(x)). \]

Consequently, if we define \( \Theta_i \) on \( \tau_i(\Omega_{ij}) \) and \( \Theta_j \) on \( \tau_j(\Omega_{ji}) \) by

\[ \Theta_i([x]) = \theta_i(x), \quad x \in \Omega_{ij} \quad \Theta_j([y]) = \theta_j(y), \quad y \in \Omega_{ji}, \]

as \( x \sim \varphi_{ji}(x) \), we have

\[ \Theta_i([x]) = \theta_i(x) = \theta_j(\varphi_{ji}(x)) = \Theta_j([\varphi_{ji}(x)]) = \Theta_j([x]), \]

which means that \( \Theta_i \) and \( \Theta_j \) agree on \( \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \). But then, the functions \( \Theta_i \) agree whenever their domains overlap and so, they patch to yield a function \( \Theta \) with domain \( M_G \) and image \( M \). By construction, \( \theta_i = \Theta \circ \tau_i \), and as a manifold induced by \( M \) is a parametric pseudo-manifold, the last statement is obvious. \( \square \)

The function \( \Theta : M_G \to M \) given by Proposition 20.4 shows how the parametric pseudo-manifold \( M \) differs from the abstract manifold \( M_G \). As we said before, a practical method for approximating 3D meshes based on parametric pseudo surfaces is described in Siqueira, Xu and Gallier [163].
Chapter 21

Tensor Algebras, Symmetric Algebras and Exterior Algebras

We begin by defining tensor products of vector spaces over a field and then we investigate some basic properties of these tensors, in particular the existence of bases and duality. After this we investigate special kinds of tensors, namely symmetric tensors and skew-symmetric tensors. Tensor products of modules over a commutative ring with identity will be discussed very briefly. They show up naturally when we consider the space of sections of a tensor product of vector bundles.

Given a linear map \( f: E \to F \) (where \( E \) and \( F \) are two vector spaces over a field \( K \)), we know that if we have a basis \( (u_i)_{i \in I} \) for \( E \), then \( f \) is completely determined by its values \( f(u_i) \) on the basis vectors. For a multilinear map \( f: E^n \to F \), we don’t know if there is such a nice property but it would certainly be very useful.

In many respects tensor products allow us to define multilinear maps in terms of their action on a suitable basis. The crucial idea is to linearize, that is, to create a new vector space \( E^\otimes n \) such that the multilinear map \( f: E^n \to F \) is turned into a linear map \( f_\otimes: E^\otimes n \to F \) which is equivalent to \( f \) in a strong sense. If in addition, \( f \) is symmetric, then we can define a symmetric tensor power \( \text{Sym}^n(E) \), and every symmetric multilinear map \( f: E^n \to F \) is turned into a linear map \( f_\otimes: \text{Sym}^n(E) \to F \) which is equivalent to \( f \) in a strong sense. Similarly, if \( f \) is alternating, then we can define a skew-symmetric tensor power \( \wedge^n(E) \), and every alternating multilinear map is turned into a linear map \( f_\wedge: \wedge^n(E) \to F \) which is equivalent to \( f \) in a strong sense.

Tensor products can be defined in various ways, some more abstract than others. We tried to stay down to earth, without excess.

Before proceeding any further, we review some facts about dual spaces and pairings. Pairings will be used to deal with dual spaces of tensors.
21.1 Linear Algebra Preliminaries: Dual Spaces and Pairings

We assume that we are dealing with vector spaces over a field $K$. As usual the dual space $E^*$ of a vector space $E$ is defined by $E^* = \text{Hom}(E, K)$.

**Definition 21.1.** Given two vector spaces $E$ and $F$ over a field $K$, a map $\langle -, - \rangle : E \times F \to K$ is a nondegenerate pairing iff it is bilinear and iff $\langle u, v \rangle = 0$ for all $v \in F$ implies $u = 0$, and $\langle u, v \rangle = 0$ for all $u \in E$ implies $v = 0$. A nondegenerate pairing induces two linear maps $\varphi : E \to F^*$ and $\psi : F \to E^*$ defined such that for all $u \in E$ and all $v \in F$, $\varphi(u)$ is the linear form in $F^*$ and $\psi(v)$ is the linear form in $E^*$ given by

\[
\varphi(u)(y) = \langle u, y \rangle \quad \text{for all } y \in F
\]

\[
\psi(v)(x) = \langle x, v \rangle \quad \text{for all } x \in E.
\]

Schematically $\varphi(u) = \langle u, - \rangle$ and $\phi(v) = \langle -, v \rangle$.

**Proposition 21.1.** For every nondegenerate pairing $\langle -, - \rangle : E \times F \to K$, the induced maps $\varphi : E \to F^*$ and $\psi : F \to E^*$ are linear and injective. Furthermore, if $E$ and $F$ are finite dimensional, then $\varphi : E \to F^*$ and $\psi : F \to E^*$ are bijective.

**Proof.** The maps $\varphi : E \to F^*$ and $\psi : F \to E^*$ are linear because $u, v \mapsto \langle u, v \rangle$ is bilinear. Assume that $\varphi(u) = 0$. This means that $\varphi(u)(y) = \langle u, y \rangle = 0$ for all $y \in F$, and as our pairing is nondegenerate, we must have $u = 0$. Similarly, $\psi$ is injective. If $E$ and $F$ are finite dimensional, then $\dim(E) = \dim(E^*)$ and $\dim(F) = \dim(F^*)$. However, the injectivity of $\varphi$ and $\psi$ implies that that $\dim(E) \leq \dim(F^*)$ and $\dim(F) \leq \dim(E^*)$. Consequently $\dim(E) \leq \dim(F)$ and $\dim(F) \leq \dim(E)$, so $\dim(E) = \dim(F)$. Therefore, $\dim(E) = \dim(F^*)$ and $\varphi$ is bijective (and similarly $\dim(F) = \dim(E^*)$ and $\psi$ is bijective).

Proposition 21.1 shows that when $E$ and $F$ are finite dimensional, a nondegenerate pairing induces canonical isomorphisms $\varphi : E \to F^*$ and $\psi : F \to E^*$; that is, isomorphisms that do not depend on the choice of bases. An important special case is the case where $E = F$ and we have an inner product (a symmetric, positive definite bilinear form) on $E$.

**Remark:** When we use the term “canonical isomorphism,” we mean that such an isomorphism is defined independently of any choice of bases. For example, if $E$ is a finite dimensional vector space and $(e_1, \ldots, e_n)$ is any basis of $E$, we have the dual basis $(e_1^*, \ldots, e_n^*)$ of $E^*$ (where, $e_i^*(e_j) = \delta_{ij}$), and thus the map $e_i \mapsto e_i^*$ is an isomorphism between $E$ and $E^*$. This isomorphism is not canonical.

On the other hand, if $\langle -, - \rangle$ is an inner product on $E$, then Proposition 21.1 shows that the nondegenerate pairing $\langle -, - \rangle$ on $E \times E$ induces a canonical isomorphism between $E$ and $E^*$. This isomorphism is often denoted $\flat : E \to E^*$, and we usually write $u^\flat$ for $\flat(u)$, with
Given any basis, \((e_1, \ldots, e_n)\) of \(E\) (not necessarily orthonormal), let \((g_{ij})\) be the \(n \times n\)-matrix given by \(g_{ij} = \langle e_i, e_j \rangle\) (the Gram matrix of the inner product). Recall that the dual basis \((e^*_1, \ldots, e^*_n)\) of \(E^*\) consists of the coordinate forms \(e^*_i \in E^*\), which are characterized by the following properties:

\[ e^*_i(e_j) = \delta_{ij}, \quad 1 \leq i, j \leq n. \]

The inverse of the Gram matrix \((g_{ij})\) is often denoted by \((g^{ij})\) (by raising the indices).

The tradition of raising and lowering indices is pervasive in the literature on tensors. It is indeed useful to have some notational convention to distinguish between vectors and linear forms (also called one-forms or covectors). The usual convention is that coordinates of vectors are written using superscripts, as in \(u = \sum_{i=1}^{n} u^i e_i\), and coordinates of one-forms are written using subscripts, as in \(\omega = \sum_{i=1}^{n} \omega_i e^*_i\). Actually, since vectors are indexed with subscripts, one-forms are indexed with superscripts, so \(e^*_i\) should be written as \(e^i\). The motivation is that summation signs can then be omitted, according to the Einstein summation convention. According to this convention, whenever a summation variable (such as \(i\)) appears both as a subscript and a superscript in an expression, it is assumed that it is involved in a summation. For example the sum \(\sum_{i=1}^{n} u^i e_i\) is abbreviated as

\[ u^i e_i, \]

and the sum \(\sum_{i=1}^{n} \omega_i e^i\) is abbreviated as

\[ \omega_i e^i. \]

In this text we will not use the Einstein summation convention, which we find somewhat confusing, and we will also write \(e^*_i\) instead of \(e^i\). The maps \(\flat\) and \(\sharp\) can be described explicitly in terms of the Gram matrix of the inner product and its inverse.

**Proposition 21.2.** For any vector space \(E\), given a basis \((e_1, \ldots, e_n)\) for \(E\) and its dual basis \((e^*_1, \ldots, e^*_n)\) for \(E^*\), for any inner product \(\langle -,- \rangle\) on \(E\), if \((g_{ij})\) is its Gram matrix, with \(g_{ij} = \langle e_i, e_j \rangle\), and \((g^{ij})\) is its inverse, then for every vector \(u = \sum_{j=1}^{n} w^j e_j \in E\) and every one-form \(\omega = \sum_{i=1}^{n} \omega_i e^*_i \in E^*\), we have

\[ u^\flat = \sum_{i=1}^{n} \omega_i e^*_i, \quad \text{with} \quad \omega_i = \sum_{j=1}^{n} g_{ij} w^j, \]

and

\[ \omega^\sharp = \sum_{j=1}^{n} (\omega^\sharp)^j e_j, \quad \text{with} \quad (\omega^\sharp)^j = \sum_{j=1}^{n} g^{ij} \omega_j. \]
Proof. For every \( u = \sum_{j=1}^{n} u^j e_j \), since \( u^b(v) = \langle u, v \rangle \) for all \( v \in E \), we have

\[
u^b(e_i) = \langle u, e_i \rangle = \left\langle \sum_{j=1}^{n} u^j e_j, e_i \right\rangle = \sum_{j=1}^{n} u^j \langle e_j, e_i \rangle = \sum_{j=1}^{n} g_{ij} u^j,
\]

so we get

\[
u^b = \sum_{i=1}^{n} \omega_i e_i^*, \quad \text{with} \quad \omega_i = \sum_{j=1}^{n} g_{ij} u^j.
\]

If we write \( \omega \in E^* \) as \( \omega = \sum_{i=1}^{n} \omega_i e_i^* \) and \( \omega^\sharp \in E \) as \( \omega^\sharp = \sum_{j=1}^{n} (\omega^\sharp)^j e_j \), since

\[
\omega_i = \omega(e_i) = \langle \omega^\sharp, e_i \rangle = \sum_{j=1}^{n} (\omega^\sharp)^j g_{ij}, \quad 1 \leq i \leq n,
\]

we get

\[
(\omega^\sharp)^j = \sum_{j=1}^{n} g^{ij} \omega_j,
\]

where \( (g^{ij}) \) is the inverse of the matrix \( (g_{ij}) \).

The map \( b \) has the effect of lowering (flattening!) indices, and the map \( \sharp \) has the effect of raising (sharpening!) indices.

Here is an explicit example of Proposition 21.2. Let \( (e_1, e_2) \) be a basis of \( E \) such that

\[
\langle e_1, e_1 \rangle = 1, \quad \langle e_1, e_2 \rangle = 2, \quad \langle e_2, e_2 \rangle = 5.
\]

Then

\[
g = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}, \quad g^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}.
\]

Set \( u = u^1 e_1 + u^2 e_2 \) and observe that

\[
u^b(e_1) = \langle u^1 e_1 + u^2 e_2, e_1 \rangle = \langle e_1, e_1 \rangle u^1 + \langle e_2, e_1 \rangle u^2 = g_{11} u^1 + g_{12} u^2 = u^1 + 2u^2
\]

\[
u^b(e_2) = \langle u^1 e_1 + u^2 e_2, e_2 \rangle = \langle e_1, e_2 \rangle u^1 + \langle e_2, e_2 \rangle u^2 = g_{21} u^1 + g_{22} u^2 = 2u^1 + 5u^2,
\]

which in turn implies that

\[
u^b = \omega_1 e_1^* + \omega_2 e_2^* = u^b(e_1) e_1^* + u^b(e_2) e_2^* = (u^1 + 2u^2) e_1^* + (2u^1 + 5u^2) e_2^*.
\]

Given \( \omega = \omega_1 e_1^* + \omega_2 e_2^* \), we calculate \( \omega^\sharp = (\omega^\sharp)^1 e_1 + (\omega^\sharp)^2 e_2 \) from the following two linear equalities:

\[
\omega_1 = \omega(e_1) = \langle \omega^\sharp, e_1 \rangle = \langle (\omega^\sharp)^1 e_1 + (\omega^\sharp)^2 e_2, e_1 \rangle
\]

\[
= \langle e_1, e_1 \rangle (\omega^\sharp)^1 + \langle e_2, e_1 \rangle (\omega^\sharp)^2 = (\omega^\sharp)^1 + 2(\omega^\sharp)^2 = g_{11}(\omega^\sharp)^1 + g_{12}(\omega^\sharp)^2
\]

\[
\omega_2 = \omega(e_2) = \langle \omega^\sharp, e_2 \rangle = \langle (\omega^\sharp)^1 e_1 + (\omega^\sharp)^2 e_2, e_2 \rangle
\]

\[
= \langle e_1, e_2 \rangle (\omega^\sharp)^1 + \langle e_2, e_2 \rangle (\omega^\sharp)^2 = 2(\omega^\sharp)^1 + 5(\omega^\sharp)^2 = g_{21}(\omega^\sharp)^1 + g_{22}(\omega^\sharp)^2.
\]
These equalities are concisely written as
\[
\begin{bmatrix}
\omega_1 \\
\omega_2
\end{bmatrix} = \begin{bmatrix} 1 & 2 \\
2 & 5
\end{bmatrix} \begin{bmatrix}
(\omega^\sharp)^1 \\
(\omega^\sharp)^2
\end{bmatrix} = g \begin{bmatrix}
(\omega^\sharp)^1 \\
(\omega^\sharp)^2
\end{bmatrix}.
\]
Then
\[
\begin{bmatrix}
(\omega^\sharp)^1 \\
(\omega^\sharp)^2
\end{bmatrix} = g^{-1} \begin{bmatrix}
\omega_1 \\
\omega_2
\end{bmatrix} = \begin{bmatrix} 5 & -2 \\
-2 & 1
\end{bmatrix} \begin{bmatrix}
\omega_1 \\
\omega_2
\end{bmatrix},
\]
which in turn implies
\[
(\omega^\sharp)^1 = 5\omega_1 - 2\omega_2, \quad (\omega^\sharp)^2 = -2\omega_1 + \omega_2,
\]
i.e.
\[
\omega^\sharp = (5\omega_1 - 2\omega_2)e_1 + (-2\omega_1 + \omega_2)e_2.
\]

The inner product \(\langle -,- \rangle\) on \(E\) induces an inner product on \(E^*\) denoted \(\langle -,- \rangle_{E^*}\), and given by
\[
\langle \omega_1, \omega_2 \rangle_{E^*} = \langle \omega^\sharp_1, \omega^\sharp_2 \rangle, \quad \text{for all } \omega_1, \omega_2 \in E^*.
\]
Then, we have
\[
\langle u^b, v^b \rangle_{E^*} = \langle (u^\sharp)^b, (v^\sharp)^b \rangle = \langle u, v \rangle \quad \text{for all } u, v \in E.
\]

If \((e_1, \ldots, e_n)\) is a basis of \(E\) and \(g_{ij} = \langle e_i, e_j \rangle\), as
\[
(e_i^\sharp)^b = \sum_{k=1}^{n} g^{ik}e_k,
\]
an easy computation shows that
\[
\langle e_i^*, e_j^* \rangle_{E^*} = \langle (e_i^\sharp)^b, (e_j^\sharp)^b \rangle = g^{ij};
\]
that is, in the basis \((e_1^*, \ldots, e_n^*)\), the inner product on \(E^*\) is represented by the matrix \((g^{ij})\), the inverse of the matrix \((g_{ij})\).

The inner product on a finite vector space also yields a natural isomorphism between the space \(\text{Hom}(E,E;K)\) of bilinear forms on \(E\), and the space \(\text{Hom}(E,E)\) of linear maps from \(E\) to itself. Using this isomorphism, we can define the trace of a bilinear form in an intrinsic manner. This technique is used in differential geometry, for example, to define the divergence of a differential one-form.

**Proposition 21.3.** If \(\langle -,- \rangle\) is an inner product on a finite vector space \(E\) (over a field, \(K\)), then for every bilinear form \(f : E \times E \rightarrow K\), there is a unique linear map \(f^\sharp : E \rightarrow E\) such that
\[
f(u,v) = \langle f^\sharp(u), v \rangle, \quad \text{for all } u, v \in E.
\]
The map \(f \mapsto f^\sharp\) is a linear isomorphism between \(\text{Hom}(E,E;K)\) and \(\text{Hom}(E,E)\).
Proof. For every \( g \in \text{Hom}(E, E) \), the map given by
\[
f(u, v) = \langle g(u), v \rangle, \quad u, v \in E,
\]
is clearly bilinear. It is also clear that the above defines a linear map from \( \text{Hom}(E, E) \) to \( \text{Hom}(E, E; K) \). This map is injective, because if \( f(u, v) = 0 \) for all \( u, v \in E \), as \( \langle -, - \rangle \) is an inner product, we get \( g(u) = 0 \) for all \( u \in E \). Furthermore, both spaces \( \text{Hom}(E, E) \) and \( \text{Hom}(E, E; K) \) have the same dimension, so our linear map is an isomorphism.

If \((e_1, \ldots, e_n)\) is an orthonormal basis of \( E \), then we check immediately that the trace of a linear map \( g \) (which is independent of the choice of a basis) is given by
\[
\text{tr}(g) = \sum_{i=1}^{n} \langle g(e_i), e_i \rangle,
\]
where \( n = \dim(E) \).

**Definition 21.2.** We define the trace of the bilinear form \( f \) by
\[
\text{tr}(f) = \text{tr}(f^\sharp).
\]

From Proposition 21.3, \( \text{tr}(f) \) is given by
\[
\text{tr}(f) = \sum_{i=1}^{n} f(e_i, e_i),
\]
for any orthonormal basis \((e_1, \ldots, e_n)\) of \( E \). We can also check directly that the above expression is independent of the choice of an orthonormal basis.

We demonstrate how to calculate \( \text{tr}(f) \) where \( f: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R} \) with \( f((x_1, y_1), (x_2, y_2)) = x_1x_2 + 2x_2y_1 + 3x_1y_2 - y_1y_2 \). Under the standard basis for \( \mathbb{R}^2 \), the bilinear form \( f \) is represented as
\[
\begin{pmatrix} x_1 & y_1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}.
\]
This matrix representation shows that
\[
f^\sharp = \begin{pmatrix} 1 & 3 \\ 2 & -1 \end{pmatrix}^\top = \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix},
\]
and hence
\[
\text{tr}(f) = \text{tr}(f^\sharp) = \text{tr}\begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} = 0.
\]

We will also need the following proposition to show that various families are linearly independent.
Proposition 21.4. Let $E$ and $F$ be two nontrivial vector spaces and let $(u_i)_{i \in I}$ be any family of vectors $u_i \in E$. The family $(u_i)_{i \in I}$ is linearly independent iff for every family $(v_i)_{i \in I}$ of vectors $v_i \in F$, there is some linear map $f: E \to F$ so that $f(u_i) = v_i$ for all $i \in I$.

Proof. Left as an exercise. \qed

21.2 Tensors Products

First we define tensor products, and then we prove their existence and uniqueness up to isomorphism.

Definition 21.3. Let $K$ be a given field, and let $E_1, \ldots, E_n$ be $n \geq 2$ given vector spaces. For any vector space $F$, a map $f: E_1 \times \cdots \times E_n \to F$ is multilinear iff it is linear in each of its argument; that is,

\[
\begin{align*}
    f(u_1, \ldots, u_i, v + w, u_{i+1}, \ldots, u_n) &= f(u_1, \ldots, u_i, v, u_{i+1}, \ldots, u_n) \\
    &\quad + f(u_1, \ldots, u_i, w, u_{i+1}, \ldots, u_n) \\
    f(u_1, \ldots, u_i, \lambda v, u_{i+1}, \ldots, u_n) &= \lambda f(u_1, \ldots, u_i, v, u_{i+1}, \ldots, u_n),
\end{align*}
\]

for all $u_j \in E_j$ ($j \neq i$), all $v, w \in E_i$ and all $\lambda \in K$, for $i = 1 \ldots, n$.

The set of multilinear maps as above forms a vector space denoted $L(E_1, \ldots, E_n; F)$ or $\text{Hom}(E_1, \ldots, E_n; F)$. When $n = 1$, we have the vector space of linear maps $L(E, F)$ (also denoted $\text{Hom}(E, F)$). (To be very precise, we write $\text{Hom}_K(E_1, \ldots, E_n; F)$ and $\text{Hom}_K(E, F)$.)

Definition 21.4. A tensor product of $n \geq 2$ vector spaces $E_1, \ldots, E_n$ is a vector space $T$ together with a multilinear map $\varphi: E_1 \times \cdots \times E_n \to T$, such that for every vector space $F$ and for every multilinear map $f: E_1 \times \cdots \times E_n \to F$, there is a unique linear map $f_\otimes: T \to F$ with

\[
f(u_1, \ldots, u_n) = f_\otimes(\varphi(u_1, \ldots, u_n)),
\]

for all $u_1 \in E_1, \ldots, u_n \in E_n$, or for short

\[
f = f_\otimes \circ \varphi.
\]

Equivalently, there is a unique linear map $f_\otimes$ such that the following diagram commutes:

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\varphi} & T \\
\downarrow f & & \downarrow f_\otimes \\
F & & \\
\end{array}
\]

The above property is called the universal mapping property of the tensor product $(T, \varphi)$.
We show that any two tensor products \((T_1, \varphi_1)\) and \((T_2, \varphi_2)\) for \(E_1, \ldots, E_n\), are isomorphic.

**Proposition 21.5.** Given any two tensor products \((T_1, \varphi_1)\) and \((T_2, \varphi_2)\) for \(E_1, \ldots, E_n\), there is an isomorphism \(h: T_1 \to T_2\) such that

\[
\varphi_2 = h \circ \varphi_1.
\]

**Proof.** Focusing on \((T_1, \varphi_1)\), we have a multilinear map \(\varphi_2: E_1 \times \cdots \times E_n \to T_2\), and thus there is a unique linear map \((\varphi_2)_{\otimes}: T_1 \to T_2\) with

\[
\varphi_2 = (\varphi_2)_{\otimes} \circ \varphi_1
\]

as illustrated by the following commutative diagram:

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\varphi_1} & T_1 \\
\varphi_2 \downarrow & & \downarrow (\varphi_2)_{\otimes} \\
& T_2, & \\
\end{array}
\]

Similarly, focusing now on \((T_2, \varphi_2)\), we have a multilinear map \(\varphi_1: E_1 \times \cdots \times E_n \to T_1\), and thus there is a unique linear map \((\varphi_1)_{\otimes}: T_2 \to T_1\) with

\[
\varphi_1 = (\varphi_1)_{\otimes} \circ \varphi_2
\]

as illustrated by the following commutative diagram:

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\varphi_2} & T_2 \\
\varphi_1 \downarrow & & \downarrow (\varphi_1)_{\otimes} \\
& T_1, & \\
\end{array}
\]

Putting these diagrams together, we obtain the commutative diagrams

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\varphi_1} & T_1 \\
\varphi_2 \downarrow & & \downarrow (\varphi_2)_{\otimes} \\
& T_2, & \\
\end{array}
\]

and

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xleftarrow{\varphi_2} & T_2 \\
\varphi_1 \downarrow & & \downarrow (\varphi_1)_{\otimes} \\
& T_1, & \\
\end{array}
\]
which means that
\[ \varphi_1 = (\varphi_1) \circ (\varphi_2) \circ \varphi_1 \quad \text{and} \quad \varphi_2 = (\varphi_2) \circ (\varphi_1) \circ \varphi_2. \]

On the other hand, focusing on \((T_1, \varphi_1)\), we have a multilinear map \(\varphi_1: E_1 \times \cdots \times E_n \to T_1\), but the unique linear map \(h: T_1 \to T_1\) with
\[ \varphi_1 = h \circ \varphi_1 \]
is \(h = \text{id}\), as illustrated by the following commutative diagram
\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\varphi_1} & T_1 \\
\downarrow{\varphi_1} & & \downarrow{id} \\
& T_1 &
\end{array}
\]

and since \((\varphi_1) \circ (\varphi_2)\) is linear as a composition of linear maps, we must have
\[ (\varphi_1) \circ (\varphi_2) = \text{id}. \]

Similarly, we have the commutative diagram
\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\varphi_2} & T_2 \\
\downarrow{\varphi_2} & & \downarrow{id} \\
& T_2 &
\end{array}
\]

and we must have
\[ (\varphi_2) \circ (\varphi_1) = \text{id}. \]

This shows that \((\varphi_1)\) and \((\varphi_2)\) are inverse linear maps, and thus, \((\varphi_2): T_1 \to T_2\) is an isomorphism between \(T_1\) and \(T_2\).

Now that we have shown that tensor products are unique up to isomorphism, we give a construction that produces them. Tensor products are obtained from free vector spaces by a quotient process, so let us begin by describing the construction of the free vector space generated by a set.

For simplicity assume that our set \(I\) is finite, say
\[ I = \{\heartsuit, \diamondsuit, \spadesuit, \clubsuit\}. \]

The construction works for any field \(K\) (and in fact for any commutative ring \(A\), in which case we obtain the free \(A\)-module generated \(I\)). Assume that \(K = \mathbb{R}\). The free vector space generated by \(I\) is the set of all formal linear combinations of the form
\[ a\heartsuit + b\diamondsuit + c\spadesuit + d\clubsuit, \]
with \( a, b, c, d \in \mathbb{R} \). It is assumed that the order of the terms does not matter. For example,
\[
2\bigcirc - 5\bigdiamond + 3\blacklozenge = -5\bigdiamond + 2\bigcirc + 3\blacklozenge.
\]

Addition and multiplication by a scalar are defined as follows:
\[
(a_1\bigcirc + b_1\bigdiamond + c_1\blacklozenge + d_1\blacklozenge) + (a_2\bigcirc + b_2\bigdiamond + c_2\blacklozenge + d_2\blacklozenge)
= (a_1 + a_2)\bigcirc + (b_1 + b_2)\bigdiamond + (c_1 + c_2)\blacklozenge + (d_1 + d_2)\blacklozenge,
\]
and
\[
\alpha \cdot (a\bigcirc + b\bigdiamond + c\blacklozenge + d\blacklozenge) = \alpha a\bigcirc + \alpha b\bigdiamond + \alpha c\blacklozenge + \alpha d\blacklozenge,
\]
for all \( a, b, c, d, \alpha \in \mathbb{R} \). With these operations, it is immediately verified that we obtain a vector space denoted \( \mathbb{R}^{(I)} \). The set \( I \) can be viewed as embedded in \( \mathbb{R}^{(I)} \) by the injection \( \iota \) given by
\[
\iota(\bigcirc) = 1\bigcirc, \quad \iota(\bigdiamond) = 1\bigdiamond, \quad \iota(\blacklozenge) = 1\blacklozenge, \quad \iota(\blacklozenge) = 1\blacklozenge.
\]
Thus, \( \mathbb{R}^{(I)} \) can be viewed as the vector space with the special basis \( I = \{\bigcirc, \bigdiamond, \blacklozenge, \blacklozenge\} \). In our case, \( \mathbb{R}^{(I)} \) is isomorphic to \( \mathbb{R}^4 \).

The exact same construction works for any field \( K \), and we obtain a vector space denoted by \( K^{(I)} \) and an injection \( \iota: I \rightarrow K^{(I)} \).

The main reason why the free vector space \( K^{(I)} \) over a set \( I \) is interesting is that it satisfies a universal mapping property. This means that for every vector space \( F \) (over the field \( K \)), any function \( h: I \rightarrow F \), where \( F \) is considered just a set, has a unique linear extension \( \overline{h}: K^{(I)} \rightarrow F \). By extension, we mean that \( \overline{h}(i) = h(i) \) for all \( i \in I \), or more rigorously that \( h = \overline{h} \circ \iota \).

For example, if \( I = \{\bigcirc, \bigdiamond, \blacklozenge, \blacklozenge\} \), \( K = \mathbb{R} \), and \( F = \mathbb{R}^3 \), the function \( h \) given by
\[
h(\bigcirc) = (1, 1, 1), \quad h(\bigdiamond) = (1, 1, 0), \quad h(\blacklozenge) = (1, 0, 0), \quad h(\blacklozenge) = (0, 0, 1)
\]
has a unique linear extension \( \overline{h}: \mathbb{R}^{(I)} \rightarrow \mathbb{R}^3 \) to the free vector space \( \mathbb{R}^{(I)} \), given by
\[
\overline{h}(a\bigcirc + b\bigdiamond + c\blacklozenge + d\blacklozenge) = a\overline{h}(\bigcirc) + b\overline{h}(\bigdiamond) + c\overline{h}(\blacklozenge) + d\overline{h}(\blacklozenge)
= ah(\bigcirc) + bh(\bigdiamond) + ch(\blacklozenge) + dh(\blacklozenge)
= a(1, 1, 1) + b(1, 1, 0) + c(1, 0, 0) + d(0, 0, 1)
= (a + b + c, b + c, a - d).
\]

To generalize the construction of a free vector space to infinite sets \( I \), we observe that the formal linear combination \( a\bigcirc + b\bigdiamond + c\blacklozenge + d\blacklozenge \) can be viewed as the function \( f: I \rightarrow \mathbb{R} \) given by
\[
f(\bigcirc) = a, \quad f(\bigdiamond) = b, \quad f(\blacklozenge) = c, \quad f(\blacklozenge) = d,
\]
where \( a, b, c, d \in \mathbb{R} \). More generally, we can replace \( \mathbb{R} \) by any field \( K \). If \( I \) is finite, then the set of all such functions is a vector space under pointwise addition and pointwise scalar
multiplication. If $I$ is infinite, since addition and scalar multiplication only makes sense for finite vectors, we require that our functions $f : I \to K$ take the value 0 except for possibly finitely many arguments. We can think of such functions as an infinite sequences $(f_i)_{i \in I}$ of elements $f_i$ of $K$ indexed by $I$, with only finitely many nonzero $f_i$. The formalization of this construction goes as follows.

Given any set $I$ viewed as an index set, let $K^{(I)}$ be the set of all functions $f : I \to K$ such that $f(i) \neq 0$ only for finitely many $i \in I$. As usual, denote such a function by $(f_i)_{i \in I}$; it is a family of finite support. We make $K^{(I)}$ into a vector space by defining addition and scalar multiplication by

$$(f_i) + (g_i) = (f_i + g_i)$$

$$(\lambda f_i) = (\lambda f_i).$$

The family $(e_i)_{i \in I}$ is defined such that $(e_i)_j = 0$ if $j \neq i$ and $(e_i)_i = 1$. It is a basis of the vector space $K^{(I)}$, so that every $w \in K^{(I)}$ can be uniquely written as a finite linear combination of the $e_i$. There is also an injection $\iota : I \to K^{(I)}$ such that $\iota(i) = e_i$ for every $i \in I$. Furthermore, it is easy to show that for any vector space $F$, and for any function $h : I \to F$, there is a unique linear map $\overline{h} : K^{(I)} \to F$ such that $h = \overline{h} \circ \iota$, as in the following diagram:

$$
I \xrightarrow{\iota} K^{(I)} \\
\downarrow h \quad \downarrow \pi \\
F.
$$

We call $(K^{(I)}, \iota)$ the free vector space generated by $I$ (or over $I$). The commutativity of the above diagram is called the universal mapping property of the free vector space $(K^{(I)}, \iota)$ over $I$. Using the proof technique of Proposition 21.5, it is not hard to prove that any two vector spaces satisfying the above universal mapping property are isomorphic.

We can now return to the construction of tensor products. For simplicity consider two vector spaces $E_1$ and $E_2$. Whatever $E_1 \otimes E_2$ and $\varphi : E_1 \times E_2 \to E_1 \otimes E_2$ are, since $\varphi$ is supposed to be bilinear, we must have

$$
\varphi(u_1 + u_2, v_1) = \varphi(u_1, v_1) + \varphi(u_2, v_1)
$$

$$
\varphi(u_1, v_1 + v_2) = \varphi(u_1, v_1) + \varphi(u_1, v_2)
$$

$$
\varphi(\lambda u_1, v_1) = \lambda \varphi(u_1, v_1)
$$

$$
\varphi(u_1, \mu v_1) = \mu \varphi(u_1, v_1)
$$

for all $u_1, u_2 \in E_1$, all $v_1, v_2 \in E_2$, and all $\lambda, \mu \in K$. Since $E_1 \otimes E_2$ must satisfy the universal mapping property of Definition 21.4, we may want to define $E_1 \otimes E_2$ as the free vector space $K^{(E_1 \times E_2)}$ generated by $I = E_1 \times E_2$ and let $\varphi$ be the injection of $E_1 \times E_2$ into $K^{(E_1 \times E_2)}$. The problem is that in $K^{(E_1 \times E_2)}$, vectors such that

$$(u_1 + u_2, v_1) \quad \text{and} \quad (u_1, v_1) + (u_2, v_2)$$
are different, when they should really be the same, since $\varphi$ is bilinear. Since $K^{(E_1 \times E_2)}$ is free, there are no relations among the generators and this vector space is too big for our purpose.

The remedy is simple: take the quotient of the free vector space $K^{(E_1 \times E_2)}$ by the subspace $N$ generated by the vectors of the form

$$
(u_1 + u_2, v_1) - (u_1, v_1) - (u_2, v_1)
$$

$$
(u_1, v_1 + v_2) - (u_1, v_1) - (u_1, v_2)
$$

$$
(\lambda u_1, v_1) - \lambda \varphi(u_1, v_1)
$$

$$
(u_1, \mu v_1) - \mu \varphi(u_1, v_1).
$$

Then, if we let $E_1 \otimes E_2$ be the quotient space $K^{(E_1 \times E_2)}/N$ and let $\varphi$ to be the quotient map, this forces $\varphi$ to be bilinear. Checking that $(K^{(E_1 \times E_2)}/N, \varphi)$ satisfies the universal mapping property is straightforward. Here is the detailed construction.

**Theorem 21.6.** Given $n \geq 2$ vector spaces $E_1, \ldots, E_n$, a tensor product $(E_1 \otimes \cdots \otimes E_n, \varphi)$ for $E_1, \ldots, E_n$ can be constructed. Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \otimes \cdots \otimes u_n$, the tensor product $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors $u_1 \otimes \cdots \otimes u_n$, where $u_i \in E_i, i = 1, \ldots, n$, and for every multilinear map $f : E_1 \times \cdots \times E_n \to F$, the unique linear map $f_\otimes : E_1 \otimes \cdots \otimes E_n \to F$ such that $f = f_\otimes \circ \varphi$ is defined by

$$
f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n)
$$

on the generators $u_1 \otimes \cdots \otimes u_n$ of $E_1 \otimes \cdots \otimes E_n$.

**Proof.** First we apply the construction of a free vector space to the cartesian product $I = E_1 \times \cdots \times E_n$, obtaining the free vector space $M = K^{(I)}$ on $I = E_1 \times \cdots \times E_n$. Since every $e_i \in M$ is uniquely associated with some $n$-tuple $i = (u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$, we denote $e_i$ by $(u_1, \ldots, u_n)$.

Next let $N$ be the subspace of $M$ generated by the vectors of the following type:

$$
(u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n),
$$

$$
(u_1, \ldots, \lambda u_i, \ldots, u_n) - \lambda(u_1, \ldots, u_i, \ldots, u_n).
$$

We let $E_1 \otimes \cdots \otimes E_n$ be the quotient $M/N$ of the free vector space $M$ by $N$, $\pi : M \to M/N$ be the quotient map, and set

$$
\varphi = \pi \circ \iota.
$$

By construction, $\varphi$ is multilinear, and since $\pi$ is surjective and the $\iota(i) = e_i$ generate $M$, the fact that each $i$ is of the form $i = (u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$ implies that $\varphi(u_1, \ldots, u_n)$ generate $M/N$. Thus, if we denote $\varphi(u_1, \ldots, u_n)$ as $u_1 \otimes \cdots \otimes u_n$, the space $E_1 \otimes \cdots \otimes E_n$ is generated by the vectors $u_1 \otimes \cdots \otimes u_n$, with $u_i \in E_i$. 
21.2. TENSORS PRODUCTS

It remains to show that \((E_1 \otimes \cdots \otimes E_n, \varphi)\) satisfies the universal mapping property. To this end, we begin by proving that there is a map \(h\) such that \(f = h \circ \varphi\). Since \(M = K^{(E_1 \times \cdots \times E_n)}\) is free on \(I = E_1 \times \cdots \times E_n\), there is a unique linear map \(\overline{f} : K^{(E_1 \times \cdots \times E_n)} \to F\), such that

\[ f = \overline{f} \circ \iota, \]

as in the diagram below:

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \overset{\iota}{\longrightarrow} & K^{(E_1 \times \cdots \times E_n)} = M \\
\downarrow{f} & & \downarrow{\overline{f}} \\
F & & \\
\end{array}
\]

Because \(f\) is multilinear, note that we must have \(\overline{f}(w) = 0\) for every \(w \in N\); for example, on the generator

\[(u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n)\]

we have

\[
\begin{align*}
\overline{f}((u_1, \ldots, u_i + v_i, \ldots, u_n) - (u_1, \ldots, u_i, \ldots, u_n) - (u_1, \ldots, v_i, \ldots, u_n)) \\
= f(u_1, \ldots, u_i + v_i, \ldots, u_n) - f(u_1, \ldots, u_i, \ldots, u_n) - f(u_1, \ldots, v_i, \ldots, u_n) \\
= f(u_1, \ldots, u_i, \ldots, u_n) + f(u_1, \ldots, v_i, \ldots, u_n) - f(u_1, \ldots, u_i, \ldots, u_n) \\
& \quad - f(u_1, \ldots, v_i, \ldots, u_n) \\
= 0.
\end{align*}
\]

But then, \(\overline{f} : M \to F\) factors through \(M/N\), which means that there is a unique linear map \(h : M/N \to F\) such that \(\overline{f} = h \circ \pi\) making the following diagram commute

\[
\begin{array}{ccc}
M & \overset{\pi}{\longrightarrow} & M/N \\
\downarrow{\overline{f}} & & \downarrow{h} \\
F & & \\
\end{array}
\]

by defining \(h([z]) = \overline{f}(z)\) for every \(z \in M\), where \([z]\) denotes the equivalence class in \(M/N\) of \(z \in M\). Indeed, the fact that \(\overline{f}\) vanishes on \(N\) insures that \(h\) is well defined on \(M/N\), and it is clearly linear by definition. Since \(f = \overline{f} \circ \iota\), from the equation \(\overline{f} = h \circ \pi\), by composing on the right with \(\iota\), we obtain

\[ f = \overline{f} \circ \iota = h \circ \pi \circ \iota = h \circ \varphi, \]

as in the following commutative diagram:
We now prove the uniqueness of \( h \). For any linear map \( f_\otimes \colon E_1 \otimes \cdots \otimes E_n \to F \) such that \( f = f_\otimes \circ \varphi \), since the vectors \( u_1 \otimes \cdots \otimes u_n \) generate \( E_1 \otimes \cdots \otimes E_n \) and since \( \varphi(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n \), the map \( f_\otimes \) is uniquely defined by
\[
f_\otimes(u_1 \otimes \cdots \otimes u_n) = f(u_1, \ldots, u_n).
\]
Since \( f = h \circ \varphi \), the map \( h \) is unique, and we let \( f_\otimes = h \).

The map \( \varphi \) from \( E_1 \times \cdots \times E_n \) to \( E_1 \otimes \cdots \otimes E_n \) is often denoted by \( \iota_\otimes \), so that
\[
\iota_\otimes(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n.
\]

What is important about Theorem 21.6 is not so much the construction itself but the fact that it produces a tensor product with the universal mapping property with respect to multilinear maps. Indeed, Theorem 21.6 yields a canonical isomorphism
\[
L(E_1 \otimes \cdots \otimes E_n, F) \cong L(E_1, \ldots, E_n; F)
\]
between the vector space of linear maps \( L(E_1 \otimes \cdots \otimes E_n, F) \), and the vector space of multilinear maps \( L(E_1, \ldots, E_n; F) \), via the linear map \( - \circ \varphi \) defined by
\[
h \mapsto h \circ \varphi,
\]
where \( h \in L(E_1 \otimes \cdots \otimes E_n, F) \). Indeed, \( h \circ \varphi \) is clearly multilinear, and since by Theorem 21.6, for every multilinear map \( f \in L(E_1, \ldots, E_n; F) \), there is a unique linear map \( f_\otimes \in L(E_1 \otimes \cdots \otimes E_n, F) \) such that \( f = f_\otimes \circ \varphi \), the map \( - \circ \varphi \) is bijective. As a matter of fact, its inverse is the map
\[
f \mapsto f_\otimes.
\]
We record this fact as the following proposition.

**Proposition 21.7.** Given a tensor product \( (E_1 \otimes \cdots \otimes E_n, \varphi) \), the linear map \( h \mapsto h \circ \varphi \) is a canonical isomorphism
\[
L(E_1 \otimes \cdots \otimes E_n, F) \cong L(E_1, \ldots, E_n; F)
\]
between the vector space of linear maps \( L(E_1 \otimes \cdots \otimes E_n, F) \), and the vector space of multilinear maps \( L(E_1, \ldots, E_n; F) \).

Using the “Hom” notation, the above canonical isomorphism is written
\[
\text{Hom}(E_1 \otimes \cdots \otimes E_n, F) \cong \text{Hom}(E_1, \ldots, E_n; F).
\]

**Remarks:**
(1) To be very precise, since the tensor product depends on the field $K$, we should subscript the symbol $\otimes$ with $K$ and write

$$E_1 \otimes_K \cdots \otimes_K E_n.$$  

However, we often omit the subscript $K$ unless confusion may arise.

(2) For $F = K$, the base field, Proposition 21.7 yields a canonical isomorphism between the vector space $L(E_1 \otimes \cdots \otimes E_n, K)$, and the vector space of multilinear forms $L(E_1, \ldots, E_n; K)$. However, $L(E_1 \otimes \cdots \otimes E_n, K)$ is the dual space $(E_1 \otimes \cdots \otimes E_n)^*$, and thus the vector space of multilinear forms $L(E_1, \ldots, E_n; K)$ is canonically isomorphic to $(E_1 \otimes \cdots \otimes E_n)^*$.

Since this isomorphism is used often, we record it as the following proposition.

**Proposition 21.8.** Given a tensor product $E_1 \otimes \cdots \otimes E_n$, there is a canonical isomorphism

$$L(E_1, \ldots, E_n; K) \cong (E_1 \otimes \cdots \otimes E_n)^*$$

between the vector space of multilinear maps $L(E_1, \ldots, E_n; K)$ and the dual $(E_1 \otimes \cdots \otimes E_n)^*$ of the tensor product $E_1 \otimes \cdots \otimes E_n$.

The fact that the map $\varphi: E_1 \times \cdots \times E_n \to E_1 \otimes \cdots \otimes E_n$ is multilinear, can also be expressed as follows:

$$u_1 \otimes \cdots \otimes (v_i + w_i) \otimes \cdots \otimes u_n = (u_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes u_n) + (u_1 \otimes \cdots \otimes w_i \otimes \cdots \otimes u_n),$$

$$u_1 \otimes \cdots \otimes (\lambda u_i) \otimes \cdots \otimes u_n = \lambda(u_1 \otimes \cdots \otimes u_i \otimes \cdots \otimes u_n).$$

Of course, this is just what we wanted! Tensors in $E_1 \otimes \cdots \otimes E_n$ are also called $n$-tensors, and tensors of the form $u_1 \otimes \cdots \otimes u_n$, where $u_i \in E_i$ are called simple (or decomposable) $n$-tensors. Those $n$-tensors that are not simple are often called compound $n$-tensors.

Not only do tensor products act on spaces, but they also act on linear maps (they are functors).

**Proposition 21.9.** Given two linear maps $f: E \to E'$ and $g: F \to F'$, there is a unique linear map

$$f \otimes g: E \otimes F \to E' \otimes F'$$

such that

$$(f \otimes g)(u \otimes v) = f(u) \otimes g(v),$$

for all $u \in E$ and all $v \in F$.  

Proof. We can define \( h : E \times F \to E' \otimes F' \) by
\[
h(u, v) = f(u) \otimes g(v).
\]
It is immediately verified that \( h \) is bilinear, and thus it induces a unique linear map
\[
f \otimes g : E \otimes F \to E' \otimes F'
\]
making the following diagram commute:

\[
\begin{array}{ccc}
E \times F & \overset{\iota \otimes}{\longrightarrow} & E \otimes F \\
\downarrow h & & \downarrow f \otimes g \\
E' \otimes F' & & \\
\end{array}
\]

such that \((f \otimes g)(u \otimes v) = f(u) \otimes g(v)\), for all \( u \in E \) and all \( v \in F \).

The linear map \( f \otimes g : E \otimes F \to E' \otimes F' \) is called the tensor product of \( f \) and \( g \).

Another way to define \( f \otimes g \) proceeds as follows. Given two linear maps \( f : E \to E' \) and \( g : F \to F' \), the map \( f \times g \) is the linear map from \( E \times F \) to \( E' \times F' \) given by
\[
(f \times g)(u, v) = (f(u), g(v)), \quad \text{for all } u \in E \text{ and all } v \in F.
\]

Then, the map \( h \) in the proof of Proposition 21.9 is given by \( h = \iota'_\otimes \circ (f \times g) \), and \( f \otimes g \) is the unique linear map making the following diagram commute:

\[
\begin{array}{ccc}
E \times F & \overset{\iota \otimes}{\longrightarrow} & E \otimes F \\
\downarrow f \times g & & \downarrow f \otimes g \\
E' \times F' & \overset{\iota'_\otimes}{\longrightarrow} & E' \otimes F'.
\end{array}
\]

Remark: The notation \( f \otimes g \) is potentially ambiguous, because \( \text{Hom}(E, F) \) and \( \text{Hom}(E', F') \) are vector spaces, so we can form the tensor product \( \text{Hom}(E, F) \otimes \text{Hom}(E', F') \) which contains elements also denoted \( f \otimes g \). To avoid confusion, the first kind of tensor product of linear maps (which yields a linear map in \( \text{Hom}(E \otimes F, E' \otimes F') \)) can be denoted by \( T(f, g) \). If we denote the tensor product \( E \otimes F \) by \( T(E, F) \), this notation makes it clearer that \( T \) is a bifunctor. If \( E, E' \) and \( F, F' \) are finite dimensional, by picking bases it is not hard to show that the map induced by \( f \otimes g \mapsto T(f, g) \) is an isomorphism
\[
\text{Hom}(E, F) \otimes \text{Hom}(E', F') \cong \text{Hom}(E \otimes F, E' \otimes F').
\]
If we also have linear maps \(f': E' \to E''\) and \(g': F' \to F''\), then we have the commutative diagram

\[
\begin{array}{ccc}
E \times F & \xrightarrow{\kappa_{\otimes}} & E \otimes F \\
\downarrow f \times g & & \downarrow f \otimes g \\
E' \times F' & \xrightarrow{\kappa_{\otimes}'} & E' \otimes F' \\
\downarrow f' \times g' & & \downarrow f' \otimes g' \\
E'' \times F'' & \xrightarrow{\kappa_{\otimes}''} & E'' \otimes F'',
\end{array}
\]

and thus the commutative diagram

\[
\begin{array}{ccc}
E \times F & \xrightarrow{\kappa_{\otimes}} & E \otimes F \\
\downarrow (f' \times g') \circ (f \times g) & & \downarrow (f' \otimes g') \circ (f \otimes g) \\
E'' \times F'' & \xrightarrow{\kappa_{\otimes}''} & E'' \otimes F''.
\end{array}
\]

We also have the commutative diagram

\[
\begin{array}{ccc}
E \times F & \xrightarrow{\kappa_{\otimes}} & E \otimes F \\
\downarrow (f' \circ f) \times (g' \circ g) & & \downarrow (f' \circ f) \otimes (g' \circ g) \\
E'' \times F'' & \xrightarrow{\kappa_{\otimes}''} & E'' \otimes F''.
\end{array}
\]

Since we immediately verify that

\[(f' \circ f) \times (g' \circ g) = (f' \times g') \circ (f \times g),\]

by uniqueness of the map between \(E \otimes F\) and \(E'' \otimes F''\) in the above diagram, we conclude that

\[(f' \circ f) \otimes (g' \circ g) = (f' \otimes g') \circ (f \otimes g).\]  \(\ast\)

The above formula yields the following useful fact.

**Proposition 21.10.** If \(f: E \to E'\) and \(g: F \to F'\) are isomorphims, then \(f \otimes g: E \otimes F \to E' \otimes F'\) is also an isomorphism.

**Proof.** If \(f^{-1}: E' \to E\) is the inverse of \(f: E \to E'\) and \(g^{-1}: F' \to F\) is the inverse of \(g: F \to F'\), then \(f^{-1} \otimes g^{-1}: E' \otimes F' \to E \otimes F\) is the inverse of \(f \otimes g: E \otimes F \to E' \otimes F'\), which is shown as follows:

\[
(f \otimes g) \circ (f^{-1} \otimes g^{-1}) = (f \circ f^{-1}) \otimes (g \circ g^{-1})
\]

\[
= \text{id}_{E'} \otimes \text{id}_{F'}
\]

\[
= \text{id}_{E' \otimes F'},
\]
and
\[(f^{-1} \otimes g^{-1}) \circ (f \otimes g) = (f^{-1} \circ f) \otimes (g^{-1} \circ g) = \text{id}_E \otimes \text{id}_F = \text{id}_{E \otimes F}.
\]
Therefore, \(f \otimes g : E \otimes F \to E' \otimes F'\) is an isomorphism. \(\square\)

The generalization to the tensor product \(f_1 \otimes \cdots \otimes f_n\) of \(n \geq 3\) linear maps \(f_i : E_i \to F_i\) is immediate, and left to the reader.

### 21.3 Bases of Tensor Products

We showed that \(E_1 \otimes \cdots \otimes E_n\) is generated by the vectors of the form \(u_1 \otimes \cdots \otimes u_n\). However, these vectors are not linearly independent. This situation can be fixed when considering bases.

To explain the idea of the proof, consider the case when we have two spaces \(E\) and \(F\) both of dimension 3. Given a basis \((e_1, e_2, e_3)\) of \(E\) and a basis \((f_1, f_2, f_3)\) of \(F\), we would like to prove that
\[e_1 \otimes f_1, \quad e_1 \otimes f_2, \quad e_1 \otimes f_3, \quad e_2 \otimes f_1, \quad e_2 \otimes f_2, \quad e_2 \otimes f_3, \quad e_3 \otimes f_1, \quad e_3 \otimes f_2, \quad e_3 \otimes f_3\]
are linearly independent. To prove this, it suffices to show that for any vector space \(G\), if \(w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33}\) are any vectors in \(G\), then there is a bilinear map \(h : E \times F \to G\) such that
\[h(e_i, e_j) = w_{ij}, \quad 1 \leq i, j \leq 3,
\]
because \(h\) yields a unique linear map \(h_\otimes : E \otimes F \to G\) such that
\[h_\otimes(e_i \otimes e_j) = w_{ij}, \quad 1 \leq i, j \leq 3,
\]
and by Proposition 21.4, the vectors
\[e_1 \otimes f_1, \quad e_1 \otimes f_2, \quad e_1 \otimes f_3, \quad e_2 \otimes f_1, \quad e_2 \otimes f_2, \quad e_2 \otimes f_3, \quad e_3 \otimes f_1, \quad e_3 \otimes f_2, \quad e_3 \otimes f_3\]
are linearly independent. This suggests understanding how a bilinear function \(f : E \times F \to G\) is expressed in terms of its values \(f(e_i, f_j)\) on the basis vectors \((e_1, e_2, e_3)\) and \((f_1, f_2, f_3)\), and this can be done easily. Using bilinearity we obtain
\[f(u_1e_1 + u_2e_2 + u_3e_3, v_1f_1 + v_2f_2 + v_3f_3) = u_1v_1f(e_1, f_1) + u_1v_2f(e_1, f_2) + u_1v_3f(e_1, f_3)
+ u_2v_1f(e_2, f_1) + u_2v_2f(e_2, f_2) + u_2v_3f(e_2, f_3)
+ u_3v_1f(e_3, f_1) + u_3v_2f(e_3, f_2) + u_3v_3f(e_3, f_3).
\]
Therefore, given \( w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, w_{31}, w_{32}, w_{33} \in G \) the function \( h \) given by

\[
    h(u_1e_1 + u_2e_2 + u_3e_3, v_1f_1 + v_2f_2 + v_3f_3) = u_1v_1w_{11} + u_1v_2w_{12} + u_1v_3w_{13} \\
    + u_2v_1w_{21} + u_2v_2w_{22} + u_2v_3w_{23} \\
    + u_3v_1w_{31} + u_3v_2w_{32} + u_3v_3w_{33}
\]

is clearly bilinear, and by construction \( h(e_i, f_j) = w_{ij} \), so it does the job.

The generalization of this argument to any number of vector spaces of any dimension (even infinite) is straightforward.

**Proposition 21.11.** Given \( n \geq 2 \) vector spaces \( E_1, \ldots, E_n \), if \( (u_i^k)_{i \in I_k} \) is a basis for \( E_k \), \( 1 \leq k \leq n \), then the family of vectors

\[
    (u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n)_{(i_1, \ldots, i_n) \in I_1 \times \cdots \times I_n}
\]

is a basis of the tensor product \( E_1 \otimes \cdots \otimes E_n \).

**Proof.** For each \( k, 1 \leq k \leq n \), every \( v^k \in E_k \) can be written uniquely as

\[
    v^k = \sum_{j \in I_k} v_j^k u_j^k,
\]

for some family of scalars \( (v_j^k)_{j \in I_k} \). Let \( F \) be any nontrivial vector space. We show that for every family

\[
    (w_{i_1, \ldots, i_n})_{(i_1, \ldots, i_n) \in I_1 \times \cdots \times I_n},
\]

of vectors in \( F \), there is some linear map \( h: E_1 \otimes \cdots \otimes E_n \to F \) such that

\[
    h(u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n) = w_{i_1, \ldots, i_n}.
\]

Then, by Proposition 21.4, it follows that

\[
    (u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n)_{(i_1, \ldots, i_n) \in I_1 \times \cdots \times I_n}
\]

is linearly independent. However, since \( (u_i^k)_{i \in I_k} \) is a basis for \( E_k \), the \( u_{i_1}^1 \otimes \cdots \otimes u_{i_n}^n \) also generate \( E_1 \otimes \cdots \otimes E_n \), and thus, they form a basis of \( E_1 \otimes \cdots \otimes E_n \).

We define the function \( f: E_1 \times \cdots \times E_n \to F \) as follows: For any \( n \) nonempty finite subsets \( J_1, \ldots, J_n \) such that \( J_k \subseteq I_k \) for \( k = 1, \ldots, n \),

\[
    f(\sum_{j_1 \in J_1} v_{j_1}^1 u_{j_1}^1, \ldots, \sum_{j_n \in J_n} v_{j_n}^n u_{j_n}^n) = \sum_{j_1 \in J_1, j_n \in J_n} v_{j_1}^1 \cdots v_{j_n}^n w_{j_1, \ldots, j_n}.
\]

It is immediately verified that \( f \) is multilinear. By the universal mapping property of the tensor product, the linear map \( f_{\otimes}: E_1 \otimes \cdots \otimes E_n \to F \) such that \( f = f_{\otimes} \circ \varphi \), is the desired map \( h \). \( \square \)
In particular, when each $I_k$ is finite and of size $m_k = \dim(E_k)$, we see that the dimension of the tensor product $E_1 \otimes \cdots \otimes E_n$ is $m_1 \cdots m_n$. As a corollary of Proposition 21.11, if $(u^k_i)_{i \in I_k}$ is a basis for $E_k$, $1 \leq k \leq n$, then every tensor $z \in E_1 \otimes \cdots \otimes E_n$ can be written in a unique way as

$$z = \sum_{(i_1, \ldots, i_n) \in I_1 \times \cdots \times I_n} \lambda_{i_1, \ldots, i_n} u^1_{i_1} \otimes \cdots \otimes u^n_{i_n},$$

for some unique family of scalars $\lambda_{i_1, \ldots, i_n} \in K$, all zero except for a finite number.

### 21.4 Some Useful Isomorphisms for Tensor Products

**Proposition 21.12.** Given 3 vector spaces $E, F, G$, there exists unique canonical isomorphisms

1. $E \otimes F \cong F \otimes E$
2. $(E \otimes F) \otimes G \cong E \otimes (F \otimes G) \cong E \otimes F \otimes G$
3. $(E \oplus F) \otimes G \cong (E \otimes G) \oplus (F \otimes G)$
4. $K \otimes E \cong E$

such that respectively

(a) $u \otimes v \mapsto v \otimes u$
(b) $(u \otimes v) \otimes w \mapsto u \otimes (v \otimes w) \mapsto u \otimes v \otimes w$
(c) $(u, v) \otimes w \mapsto (u \otimes w, v \otimes w)$
(d) $\lambda \otimes u \mapsto \lambda u$.

**Proof.** Except for (3), these isomorphisms are proved using the universal mapping property of tensor products.

(1) The map from $E \times F$ to $F \otimes E$ given by $(u, v) \mapsto v \otimes u$ is clearly bilinear, thus it induces a unique linear $\alpha : E \otimes F \to F \otimes E$ making the following diagram commute

$$
\begin{array}{ccc}
E \times F & \xrightarrow{\iota \otimes} & E \otimes F \\
\downarrow & & \downarrow \alpha \\
F \otimes E,
\end{array}
$$

such that

$$\alpha(u \otimes v) = v \otimes u, \quad \text{for all } u \in E \text{ and all } v \in F,$$
Similarly, The map from $F \times E$ to $E \otimes F$ given by $(v, u) \mapsto u \otimes v$ is clearly bilinear, thus it induces a unique linear $\beta: F \otimes E \to E \otimes F$ making the following diagram commute

\[
\begin{array}{c}
F \times E \xrightarrow{i \otimes} F \otimes E \\
\downarrow \beta \\
E \otimes F,
\end{array}
\]

such that

\[
\beta(v \otimes u) = u \otimes v, \quad \text{for all } u \in E \text{ and all } v \in F,
\]

It is immediately verified that

\[
(\beta \circ \alpha)(u \otimes v) = u \otimes v \quad \text{and} \quad (\alpha \circ \beta)(v \otimes u) = v \otimes u
\]

for all $u \in E$ and all $v \in F$. Since the tensors of the form $u \otimes v$ span $E \otimes F$ and similarly the tensors of the form $v \otimes u$ span $F \otimes E$, the map $\beta \circ \alpha$ is actually the identity on $E \otimes F$, and similarly $\alpha \circ \beta$ is the identity on $F \otimes E$, so $\alpha$ and $\beta$ are isomorphisms.

(2) Fix some $w \in G$. The map

\[
(u, v) \mapsto u \otimes v \otimes w
\]

from $E \times F$ to $E \otimes F \otimes G$ is bilinear, and thus there is a linear map $f_w: E \otimes F \to E \otimes F \otimes G$ making the following diagram commute

\[
\begin{array}{c}
E \times F \xrightarrow{i \otimes} E \otimes F \\
\downarrow f_w \\
E \otimes F \otimes G,
\end{array}
\]

such that $f_w(u \otimes v) = u \otimes v \otimes w$.

Next, consider the map

\[
(z, w) \mapsto f_w(z),
\]

from $(E \otimes F) \times G$ into $E \otimes F \otimes G$. It is easily seen to be bilinear, and thus it induces a linear map $f: (E \otimes F) \otimes G \to E \otimes F \otimes G$ making the following diagram commute

\[
\begin{array}{c}
(E \otimes F) \times G \xrightarrow{i \otimes} (E \otimes F) \otimes G \\
\downarrow f \\
E \otimes F \otimes G,
\end{array}
\]

such that $f((u \otimes v) \otimes w) = u \otimes v \otimes w$.

Also consider the map

\[
(u, v, w) \mapsto (u \otimes v) \otimes w
\]
from $E \times F \times G$ to $(E \otimes F) \otimes G$. It is trilinear, and thus there is a linear map $g$: $E \otimes F \otimes G \to (E \otimes F) \otimes G$ making the following diagram commute

$$\begin{array}{ccc}
E \times F \times G & \xrightarrow{\imath \otimes} & E \otimes F \otimes G \\
\downarrow{g} & & \downarrow{g} \\
(E \otimes F) \otimes G, & & \\
\end{array}$$

such that $g(u \otimes v \otimes w) = (u \otimes v) \otimes w$. Clearly, $f \circ g$ and $g \circ f$ are identity maps, and thus $f$ and $g$ are isomorphisms. The other case is similar.

(3) Given a fixed vector space $G$, for any two vector spaces $M$ and $N$ and every linear map $f$: $M \to N$, let $\tau_G(f) = f \otimes \text{id}_G$ be the unique linear map making the following diagram commute:

$$\begin{array}{ccc}
M \times G & \xrightarrow{\imath_M \otimes} & M \otimes G \\
\downarrow{f \times \text{id}_G} & & \downarrow{f \otimes \text{id}_G} \\
N \times G & \xrightarrow{\imath_N \otimes} & N \otimes G. \\
\end{array}$$

The identity (*) proved in Section 21.2 shows that if $g$: $N \to P$ is another linear map, then

$$\tau_G(g) \circ \tau_G(f) = (g \otimes \text{id}_G) \circ (f \otimes \text{id}_G) = (g \circ f) \otimes (\text{id}_G \circ \text{id}_G) = (g \circ f) \otimes \text{id}_G = \tau_G(g \circ f).$$

Clearly, $\tau_G(0) = 0$, and a direct computation on generators also shows that

$$\tau_G(\text{id}_M) = (\text{id}_M \otimes \text{id}_G) = \text{id}_{M \otimes G}$$

and that if $f'$: $M \to N$ is another linear map, then

$$\tau_G(f + f') = \tau_G(f) + \tau_G(f').$$

In fancy terms, $\tau_G$ is a functor. Now, if $E \oplus F$ is a direct sum, it is a standard fact of linear algebra that if $\pi_E$: $E \oplus F \to E$ and $\pi_F$: $E \oplus F \to F$ are the projection maps, then

$$\pi_E \circ \pi_E = \pi_E \quad \pi_F \circ \pi_F = \pi_F \quad \pi_E \circ \pi_F = 0 \quad \pi_F \circ \pi_E = 0 \quad \pi_E + \pi_F = \text{id}_{E \oplus F}.$$ 

If we apply $\tau_G$ to these identites, we get

$$\tau_G(\pi_E) \circ \tau_G(\pi_E) = \tau_G(\pi_E) \quad \tau_G(\pi_F) \circ \tau_G(\pi_F) = \tau_G(\pi_F)$$

$$\tau_G(\pi_E) \circ \tau_G(\pi_F) = 0 \quad \tau_G(\pi_F) \circ \tau_G(\pi_E) = 0 \quad \tau_G(\pi_E) + \tau_G(\pi_F) = \text{id}_{(E \oplus F) \otimes G}.$$ 

Observe that $\tau_G(\pi_E) = \pi_E \otimes \text{id}_G$ is a map from $(E \oplus F) \otimes G$ onto $E \otimes G$ and that $\tau_G(\pi_F) = \pi_F \otimes \text{id}_G$ is a map from $(E \oplus F) \otimes G$ onto $F \otimes G$, and by linear algebra, the above equations mean that we have a direct sum

$$(E \otimes G) \oplus (F \otimes G) \cong (E \oplus F) \otimes G.$$
(4) We have the linear map $\epsilon : E \to K \otimes E$ given by

$$\epsilon(u) = 1 \otimes u, \quad \text{for all } u \in E.$$ 

The map $(\lambda, u) \mapsto \lambda u$ from $K \times E$ to $E$ is bilinear, so it induces a unique linear map $\eta : K \otimes E \to E$ making the following diagram commute

$$
\begin{array}{ccc}
K \times E & \xrightarrow{\iota \otimes} & K \otimes E \\
& \searrow & \downarrow \eta \\
& & E,
\end{array}
$$

such that $\eta(\lambda \otimes u) = \lambda u$, for all $\lambda \in K$ and all $u \in E$. We have

$$(\eta \circ \epsilon)(u) = \eta(1 \otimes u) = 1u = u,$$

and

$$(\epsilon \circ \eta)(\lambda \otimes u) = \epsilon(\lambda u) = 1 \otimes (\lambda u) = \lambda(1 \otimes u) = \lambda \otimes u,$$

which shows that both $\epsilon \circ \eta$ and $\eta \circ \epsilon$ are the identity, so $\epsilon$ and $\eta$ are isomorphisms. \hfill \Box

**Remark:** The isomorphism (3) can be generalized to finite and even arbitrary direct sums $\bigoplus_{i \in I} E_i$ of vector spaces (where $I$ is an arbitrary nonempty index set). We have an isomorphism

$$(\bigoplus_{i \in I} E_i) \otimes G \cong \bigoplus_{i \in I} (E_i \otimes G).$$

This isomorphism (with isomorphism (1)) can be used to give another proof of Proposition 21.11 (see Bertin [21], Chapter 4, Section 1) or Lang [111], Chapter XVI, Section 2).

**Proposition 21.13.** Given any three vector spaces, $E, F, G$, we have the canonical isomorphism

$$\text{Hom}(E, F; G) \cong \text{Hom}(E, \text{Hom}(F, G)).$$

**Proof.** Any bilinear map $f : E \times F \to G$ gives the linear map $\varphi(f) \in \text{Hom}(E, \text{Hom}(F, G))$, where $\varphi(f)(u)$ is the linear map in $\text{Hom}(F, G)$ given by

$$\varphi(f)(u)(v) = f(u, v).$$

Conversely, given a linear map $g \in \text{Hom}(E, \text{Hom}(F, G))$, we get the bilinear map $\psi(g)$ given by

$$\psi(g)(u, v) = g(u)(v),$$

and it is clear that $\varphi$ and $\psi$ and mutual inverses. \hfill \Box
Since by Proposition 21.7 there is a canonical isomorphism

\[ \text{Hom}(E \otimes F, G) \cong \text{Hom}(E, F; G), \]

together with the isomorphism

\[ \text{Hom}(E, F; G) \cong \text{Hom}(E, \text{Hom}(F, G)) \]
given by Proposition 21.13, we obtain the important corollary:

**Proposition 21.14.** For any three vector spaces, \( E, F, G \), we have the canonical isomorphism

\[ \text{Hom}(E \otimes F, G) \cong \text{Hom}(E, \text{Hom}(F, G)). \]

### 21.5 Duality for Tensor Products

In this section, all vector spaces are assumed to have finite dimension, unless specified otherwise. Let us now see how tensor products behave under duality. For this, we define a pairing between \( E_1^* \otimes \cdots \otimes E_n^* \) and \( E_1 \otimes \cdots \otimes E_n \) as follows: For any fixed \( (v_1^*, \ldots, v_n^*) \in E_1^* \times \cdots \times E_n^* \), we have the multilinear map

\[ l_{v_1^*, \ldots, v_n^*} : (u_1, \ldots, u_n) \mapsto v_1^*(u_1) \cdots v_n^*(u_n) \]

from \( E_1 \times \cdots \times E_n \) to \( K \). The map \( l_{v_1^*, \ldots, v_n^*} \) extends uniquely to a linear map \( L_{v_1^*, \ldots, v_n^*} : E_1 \otimes \cdots \otimes E_n \to K \) making the following diagram commute

\[
\begin{array}{ccc}
E_1 \times \cdots \times E_n & \xrightarrow{\iota \otimes} & E_1 \otimes \cdots \otimes E_n \\
\downarrow l_{v_1^*, \ldots, v_n^*} & & \downarrow L_{v_1^*, \ldots, v_n^*} \\
 & K & \\
\end{array}
\]

We also have the multilinear map

\[ (v_1^*, \ldots, v_n^*) \mapsto L_{v_1^*, \ldots, v_n^*} \]

from \( E_1^* \times \cdots \times E_n^* \) to \( \text{Hom}(E_1 \otimes \cdots \otimes E_n, K) \), which extends to a unique linear map \( L \) from \( E_1^* \otimes \cdots \otimes E_n^* \) to \( \text{Hom}(E_1 \otimes \cdots \otimes E_n, K) \) making the following diagram commute

\[
\begin{array}{ccc}
E_1^* \times \cdots \times E_n^* & \xrightarrow{\iota \otimes} & E_1^* \otimes \cdots \otimes E_n^* \\
\downarrow L_{v_1^*, \ldots, v_n^*} & & \downarrow L \\
\text{Hom}(E_1 \otimes \cdots \otimes E_n; K) & & \\
\end{array}
\]

However, in view of the isomorphism

\[ \text{Hom}(U \otimes V, W) \cong \text{Hom}(U, \text{Hom}(V, W)) \]
given by Proposition 21.14, with $U = E_1^* \otimes \cdots \otimes E_n^*$, $V = E_1 \otimes \cdots \otimes E_n$ and $W = K$, we can view $L$ as a linear map

$$L: (E_1^* \otimes \cdots \otimes E_n^*) \otimes (E_1 \otimes \cdots \otimes E_n) \rightarrow K,$$

which corresponds to a bilinear map

$$(E_1^* \otimes \cdots \otimes E_n^*) \times (E_1 \otimes \cdots \otimes E_n) \rightarrow K,$$

via the isomorphism $(U \otimes V)^* \cong \text{Hom}(U, V; K)$ given by Proposition 21.8. It is easy to check that this bilinear map is nondegenerate, and thus by Proposition 21.1, we have a canonical isomorphism

$$(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*.$$

Here is our main proposition about duality of tensor products.

**Proposition 21.15.** We have canonical isomorphisms

$$(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*,$$

and

$$\mu: E_1^* \otimes \cdots \otimes E_n^* \cong \text{Hom}(E_1, \ldots, E_n; K).$$

**Proof.** The second isomorphism follows from the isomorphism $(E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*$ together with the isomorphism $\text{Hom}(E_1, \ldots, E_n; K) \cong (E_1 \otimes \cdots \otimes E_n)^*$ given by Proposition 21.8. \qed

**Remark:** The isomorphism $\mu: E_1^* \otimes \cdots \otimes E_n^* \cong \text{Hom}(E_1, \ldots, E_n; K)$ can be described explicitly as the linear extension to $E_1^* \otimes \cdots \otimes E_n^*$ of the map given by

$$\mu(v_1^* \otimes \cdots \otimes v_n^*)(u_1, \ldots, u_n) = v_1^*(u_1) \cdots v_n^*(u_n).$$

We prove another useful canonical isomorphism that allows us to treat linear maps as tensors.

Let $E$ and $F$ be two vector spaces and let $\alpha: E^* \times F \rightarrow \text{Hom}(E, F)$ be the map defined such that

$$\alpha(u^*, f)(x) = u^*(x)f,$$

for all $u^* \in E^*$, $f \in F$, and $x \in E$. This map is clearly bilinear, and thus it induces a linear map $\alpha_\otimes: E^* \otimes F \rightarrow \text{Hom}(E, F)$ making the following diagram commute

$$\begin{array}{ccc}
E^* \times F & \xrightarrow{\otimes} & E^* \otimes F \\
\alpha \downarrow & & \downarrow \alpha_\otimes \\
\text{Hom}(E, F), & & \\
\end{array}$$

such that

$$\alpha_\otimes(u^* \otimes f)(x) = u^*(x)f.$$
Proposition 21.16. If $E$ and $F$ are vector spaces, then the following properties hold:

1. The linear map $\alpha_\otimes : E^* \otimes F \to \text{Hom}(E, F)$ is injective.

2. If $E$ is finite-dimensional, then $\alpha_\otimes : E^* \otimes F \to \text{Hom}(E, F)$ is a canonical isomorphism.

3. If $F$ is finite-dimensional, then $\alpha_\otimes : E^* \otimes F \to \text{Hom}(E, F)$ is a canonical isomorphism.

Proof. (1) Let $(e_i^*)_{i \in I}$ be a basis of $E^*$ and let $(f_j)_{j \in J}$ be a basis of $F$. Then, we know that $(e_i^* \otimes f_j)_{i \in I, j \in J}$ is a basis of $E^* \otimes F$. To prove that $\alpha_\otimes$ is injective, let us show that its kernel is reduced to $(0)$. For any vector

$$\omega = \sum_{i \in I', j \in J'} \lambda_{ij} e_i^* \otimes f_j$$

in $E^* \otimes F$, with $I'$ and $J'$ some finite sets, assume that $\alpha_\otimes(\omega) = 0$. This means that for every $x \in E$, we have $\alpha_\otimes(\omega)(x) = 0$; that is,

$$\sum_{i \in I', j \in J'} \alpha_\otimes(\lambda_{ij} e_i^* \otimes f_j)(x) = \sum_{j \in J'} \left( \sum_{i \in I} \lambda_{ij} e_i^* (x) \right) f_j = 0.$$

Since $(f_j)_{j \in J}$ is a basis of $F$, for every $j \in J'$, we must have

$$\sum_{i \in I'} \lambda_{ij} e_i^* (x) = 0, \quad \text{for all } x \in E.$$

But, then $(e_i^*)_{i \in I'}$ would be linearly dependent, contradicting the fact that $(e_i^*)_{i \in I}$ is a basis of $E^*$, so we must have

$$\lambda_{ij} = 0, \quad \text{for all } i \in I' \text{ and all } j \in J',$$

which shows that $\omega = 0$. Therefore, $\alpha_\otimes$ is injective.

(2) Let $(e_j)_{1 \leq j \leq n}$ be a finite basis of $E$, and as usual, let $e_j^* \in E^*$ be the linear form defined by

$$e_j^*(e_k) = \delta_{j,k},$$

where $\delta_{j,k} = 1$ iff $j = k$ and 0 otherwise. We know that $(e_j^*)_{1 \leq j \leq n}$ is a basis of $E^*$ (this is where we use the finite dimension of $E$). Now, for any linear map $f \in \text{Hom}(E, F)$, for every $x = x_1 e_1 + \cdots + x_n e_n \in E$, we have

$$f(x) = f(x_1 e_1 + \cdots + x_n e_n) = x_1 f(e_1) + \cdots + x_n f(e_n) = e_1^*(x) f(e_1) + \cdots + e_n^*(x) f(e_n).$$

Consequently, every linear map $f \in \text{Hom}(E, F)$ can be expressed as

$$f(x) = e_1^*(x) f_1 + \cdots + e_n^*(x) f_n,$$
for some $f_i \in F$. Furthermore, if we apply $f$ to $e_i$, we get $f(e_i) = f_i$, so the $f_i$ are unique. Observe that

$$(\alpha_\otimes(e_1^* \otimes f_1 + \cdots + e_n^* \otimes f_n))(x) = \sum_{i=1}^{n} (\alpha_\otimes(e_i^* \otimes f_i))(x) = \sum_{i=1}^{n} e_i^*(x)f_i.$$  

Thus, $\alpha_\otimes$ is surjective, so $\alpha_\otimes$ is a bijection.

(3) Let $(f_1, \ldots, f_m)$ be a finite basis of $F$, and let $(f_1^*, \ldots, f_m^*)$ be its dual basis. Given any linear map $h: E \rightarrow F$, for all $u \in E$, since $f_i^*(f_j) = \delta_{ij}$, we have

$$h(u) = \sum_{i=1}^{m} f_i^*(h(u))f_i.$$  

If

$$h(u) = \sum_{j=1}^{m} v_j^*(u)f_j$$  

for some linear forms $(v_1^*, \ldots, v_m^*) \in (E^*)^m$, then

$$f_i^*(h(u)) = \sum_{j=1}^{m} v_j^*(u)f_i^*(f_j) = v_i^*(u)$$  

for all $u \in E$, which shows that $v_i^* = f_i^* \circ h$ for $i = 1, \ldots, m$. This means that $h$ has a unique expression in terms of linear forms as in $(\star)$. Define the map $\alpha$ from $(E^*)^m$ to $\text{Hom}(E, F)$ by

$$\alpha(v_1^*, \ldots, v_m^*)(u) = \sum_{j=1}^{m} v_j^*(u)f_j$$  

for all $u \in E$.

This map is linear. For any $h \in \text{Hom}(E, F)$, we showed earlier that the expression of $h$ in $(\star)$ is unique, thus $\alpha$ is an isomorphism. Similarly, $E^* \otimes F$ is isomorphic to $(E^*)^m$. Any tensor $\omega \in E^* \otimes F$ can be written as a linear combination

$$\sum_{k=1}^{p} u_k^* \otimes y_k$$  

for some $u_k^* \in E^*$ and some $y_k \in F$, and since $(f_1, \ldots, f_m)$ is a basis of $F$, each $y_k$ can be written as a linear combination of $(f_1, \ldots, f_m)$, so $\omega$ can be expressed as

$$\omega = \sum_{i=1}^{m} v_i^* \otimes f_i,$$  

for some linear forms $v_i^* \in E^*$ which are linear combinations of the $u_k^*$. If we pick a basis $(w_i^*)_{i \in I}$ for $E^*$, then we know that the family $(w_i^* \otimes f_j)_{i \in I, 1 \leq j \leq m}$ is a basis of $E^* \otimes F$, and this implies that the $v_i^*$ in $(\dagger)$ are unique. Define the linear map $\beta$ from $(E^*)^m$ to $E^* \otimes F$ by

$$\beta(v_1^*, \ldots, v_m^*) = \sum_{i=1}^{m} v_i^* \otimes f_i.$$
Since every tensor \( \omega \in E^* \otimes F \) can be written in a unique way as in (†), this map is an isomorphism.

Note that in Proposition 21.16, we have an isomorphism if either \( E \) or \( F \) has finite dimension. In view of the canonical isomorphism

\[
\text{Hom}(E_1, \ldots, E_n; F) \cong \text{Hom}(E_1 \otimes \cdots \otimes E_n, F)
\]
given by Proposition 21.7 and the canonical isomorphism \((E_1 \otimes \cdots \otimes E_n)^* \cong E_1^* \otimes \cdots \otimes E_n^*\) given by Proposition 21.15, if the \( E_i \)'s are finite-dimensional, then Proposition 21.16 yields the canonical isomorphism

\[
\text{Hom}(E_1, \ldots, E_n; F) \cong E_1^* \otimes \cdots \otimes E_n^* \otimes F.
\]

## 21.6 Tensor Algebras

The tensor product

\[
\underbrace{V \otimes \cdots \otimes V}_m
\]

is also denoted as

\[
\bigotimes^m V \quad \text{or} \quad V^\otimes m
\]

and is called the \( m \)-th tensor power of \( V \) (with \( V^\otimes 1 = V \), and \( V^\otimes 0 = K \)). We can pack all the tensor powers of \( V \) into the “big” vector space

\[
T(V) = \bigoplus_{m \geq 0} V^\otimes m,
\]

also denoted \( T^*(V) \) or \( \bigotimes V \) to avoid confusion with the tangent bundle. This is an interesting object because we can define a multiplication operation on it which makes it into an algebra called the tensor algebra of \( V \).

When \( V \) is of finite dimension \( n \), we can pick some basis \((e_1, \ldots, e_n)\) of \( V \), and then every tensor \( \omega \in T(V) \) can be expressed as a linear combination of terms of the form \( e_{i_1} \otimes \cdots \otimes e_{i_k} \), where \((i_1, \ldots, i_k)\) is any sequence of elements from the set \( \{1, \ldots, n\} \). We can think of the tensors \( e_{i_1} \otimes \cdots \otimes e_{i_k} \) as monomials in the noncommuting variables \( e_1, \ldots, e_n \). Thus the space \( T(V) \) corresponds to the algebra of polynomials with coefficients in \( K \) in \( n \) noncommuting variables.

Let us review the definition of an algebra over a field. Let \( K \) denote any (commutative) field, although for our purposes, we may assume that \( K = \mathbb{R} \) (and occasionally, \( K = \mathbb{C} \)). Since we will only be dealing with associative algebras with a multiplicative unit, we only define algebras of this kind.
21.6. TENSOR ALGEBRAS

Definition 21.5. Given a field $K$, a $K$-algebra is a $K$-vector space $A$ together with a bilinear operation $\cdot: A \times A \to A$, called multiplication, which makes $A$ into a ring with unity $1$ (or $1_A$, when we want to be very precise). This means that $\cdot$ is associative and that there is a multiplicative identity element $1$ so that $1 \cdot a = a \cdot 1 = a$, for all $a \in A$. Given two $K$-algebras $A$ and $B$, a $K$-algebra homomorphism $h: A \to B$ is a linear map that is also a ring homomorphism, with $h(1_A) = 1_B$; that is,

$$h(a_1 \cdot a_2) = h(a_1) \cdot h(a_2) \quad \text{for all } a_1, a_2 \in A$$

$$h(1_A) = 1_B.$$ 

The set of $K$-algebra homomorphisms between $A$ and $B$ is denoted $\text{Hom}_{\text{alg}}(A, B)$.

For example, the ring $\text{M}_n(K)$ of all $n \times n$ matrices over a field $K$ is a $K$-algebra.

There is an obvious notion of ideal of a $K$-algebra: An ideal $\mathfrak{a} \subseteq A$ is a linear subspace of $A$ that is also a two-sided ideal with respect to multiplication in $A$; this means that for all $a \in \mathfrak{a}$ and all $\alpha, \beta \in A$, we have $\alpha a \beta \in \mathfrak{a}$. If the field $K$ is understood, we usually simply say an algebra instead of a $K$-algebra.

We would like to define a multiplication operation on $T(V)$ which makes it into a $K$-algebra. As

$$T(V) = \bigoplus_{i \geq 0} V^\otimes i,$$

for every $i \geq 0$, there is a natural injection $\iota_n: V^\otimes n \to T(V)$, and in particular, an injection $\iota_0: K \to T(V)$. The multiplicative unit $1$ of $T(V)$ is the image $\iota_0(1)$ in $T(V)$ of the unit $1$ of the field $K$. Since every $v \in T(V)$ can be expressed as a finite sum

$$v = \iota_{n_1}(v_1) + \cdots + \iota_{n_k}(v_k),$$

where $v_i \in V^\otimes n_i$ and the $n_i$ are natural numbers with $n_i \neq n_j$ if $i \neq j$, to define multiplication in $T(V)$, using bilinearity, it is enough to define multiplication operations

$$\cdot: V^\otimes m \times V^\otimes n \to V^\otimes (m+n),$$

which, using the isomorphisms $V^\otimes n \cong \iota_n(V^\otimes n)$, yield multiplication operations $\cdot: \iota_m(V^\otimes m) \times \iota_n(V^\otimes n) \to \iota_{m+n}(V^\otimes (m+n))$. First, for $\omega_1 \in V^\otimes m$ and $\omega_2 \in V^\otimes n$, we let

$$\omega_1 \cdot \omega_2 = \omega_1 \otimes \omega_2.$$ 

This defines a bilinear map so it defines a multiplication $V^\otimes m \times V^\otimes n \to V^\otimes m \otimes V^\otimes n$. This is not quite what we want, but there is a canonical isomorphism

$$V^\otimes m \otimes V^\otimes n \cong V^\otimes (m+n)$$

which yields the desired multiplication $\cdot: V^\otimes m \times V^\otimes n \to V^\otimes (m+n)$.

The isomorphism $V^\otimes m \otimes V^\otimes n \cong V^\otimes (m+n)$ can be established by induction using the isomorphism $(E \otimes F) \otimes G \cong E \otimes (F \otimes G)$. First, we prove by induction on $m \geq 2$ that

$$V^\otimes (m-1) \otimes V \cong V^\otimes m,$$
and then by induction on \( n \geq 1 \) than
\[
V^\otimes m \otimes V^\otimes n \cong V^\otimes (m+n).
\]

In summary the multiplication \( V^\otimes m \times V^\otimes n \rightarrow V^\otimes (m+n) \) is defined so that
\[
(v_1 \otimes \cdots \otimes v_m) \cdot (w_1 \otimes \cdots \otimes w_n) = v_1 \otimes \cdots \otimes v_m \otimes w_1 \otimes \cdots \otimes w_n.
\]
(This has to be made rigorous by using isomorphisms involving the associativity of tensor products, for details, see Jacobson [98], Section 3.9, or Bertin [21], Chapter 4, Section 2.)

Remark: It is important to note that multiplication in \( T(V) \) is not commutative. Also, in all rigor, the unit \( 1 \) of \( T(V) \) is not equal to \( 1 \), the unit of the field \( K \). However, in view of the injection \( \iota_0: K \rightarrow T(V) \), for the sake of notational simplicity, we will denote \( 1 \) by \( 1 \). More generally, in view of the injections \( \iota_n: V^\otimes n \rightarrow T(V) \), we identify elements of \( V^\otimes n \) with their images in \( T(V) \).

The algebra \( T(V) \) satisfies a universal mapping property which shows that it is unique up to isomorphism. For simplicity of notation, let \( i: V \rightarrow T(V) \) be the natural injection of \( V \) into \( T(V) \).

**Proposition 21.17.** Given any \( K \)-algebra \( A \), for any linear map \( f: V \rightarrow A \), there is a unique \( K \)-algebra homomorphism \( \overline{f}: T(V) \rightarrow A \) so that
\[
f = \overline{f} \circ i,
\]
as in the diagram below:

\[
\begin{array}{ccc}
V & \xrightarrow{i} & T(V) \\
\downarrow{f} & & \downarrow{\overline{f}} \\
A
\end{array}
\]

Proof. Left an an exercise (use Theorem 21.6). A proof can be found in Knapp [106] (Appendix A, Proposition A.14) or Bertin [21] (Chapter 4, Theorem 2.4).

Proposition 21.17 implies that there is a natural isomorphism
\[
\text{Hom}_{\text{alg}}(T(V), A) \cong \text{Hom}(V, A),
\]
where the algebra \( A \) on the right-hand side is viewed as a vector space. Given a linear map \( h: V_1 \rightarrow V_2 \) between two vectors spaces \( V_1, V_2 \) over a field \( K \), there is a unique \( K \)-algebra homomorphism \( \otimes h: T(V_1) \rightarrow T(V_2) \) making the following diagram commute:

\[
\begin{array}{ccc}
V_1 & \xrightarrow{i_1} & T(V_1) \\
\downarrow{h} & & \downarrow{\otimes h} \\
V_2 & \xrightarrow{i_2} & T(V_2)
\end{array}
\]
Most algebras of interest arise as well-chosen quotients of the tensor algebra $T(V)$. This is true for the exterior algebra $\bigwedge(V)$ (also called Grassmann algebra), where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v$, where $v \in V$, and for the symmetric algebra $\text{Sym}(V)$, where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes w - w \otimes v$, where $v, w \in V$.

Algebras such as $T(V)$ are graded, in the sense that there is a sequence of subspaces $V \otimes^n \subseteq T(V)$ such that

$$T(V) = \bigoplus_{k \geq 0} V \otimes^k,$$

and the multiplication $\otimes$ behaves well w.r.t. the grading, i.e., $\otimes: V \otimes^m \times V \otimes^n \to V \otimes^{m+n}$. Generally, a $K$-algebra $E$ is said to be a graded algebra iff there is a sequence of subspaces $E^n \subseteq E$ such that

$$E = \bigoplus_{k \geq 0} E^k,$$

(with $E^0 = K$) and the multiplication $\cdot$ respects the grading; that is, $\cdot: E^m \times E^n \to E^{m+n}$. Elements in $E^n$ are called homogeneous elements of rank (or degree) $n$.

In differential geometry and in physics it is necessary to consider slightly more general tensors.

**Definition 21.6.** Given a vector space $V$, for any pair of nonnegative integers $(r, s)$, the tensor space $T^{r,s}(V)$ of type $(r, s)$ is the tensor product

$$T^{r,s}(V) = V \otimes^r \otimes (V^*) \otimes^s = \underbrace{V \otimes \cdots \otimes V \otimes V^* \otimes \cdots \otimes V^*},$$

with $T^{0,0}(V) = K$. We also define the tensor algebra $T^{*,*}(V)$ as the direct sum (coproduct)

$$T^{*,*}(V) = \bigoplus_{r,s \geq 0} T^{r,s}(V).$$

Tensors in $T^{r,s}(V)$ are called homogeneous of degree $(r, s)$.

Note that tensors in $T^{r,0}(V)$ are just our “old tensors” in $V \otimes^r$. We make $T^{*,*}(V)$ into an algebra by defining multiplication operations

$$T^{r_1,s_1}(V) \times T^{r_2,s_2}(V) \to T^{r_1+r_2,s_1+s_2}(V)$$

in the usual way, namely: For $u = u_1 \otimes \cdots \otimes u_{r_1} \otimes u_1^* \otimes \cdots \otimes u_{s_1}^*$ and $v = v_1 \otimes \cdots \otimes v_{r_2} \otimes v_1^* \otimes \cdots \otimes v_{s_2}^*$, let

$$u \otimes v = u_1 \otimes \cdots \otimes u_{r_1} \otimes v_1 \otimes \cdots \otimes v_{r_2} \otimes u_1^* \otimes \cdots \otimes u_{s_1}^* \otimes v_1^* \otimes \cdots \otimes v_{s_2}^*.$$
Denote by $\text{Hom}(V^r, (V^*)^s; W)$ the vector space of all multilinear maps from $V^r \times (V^*)^s$ to $W$. Then, we have the universal mapping property which asserts that there is a canonical isomorphism

$$\text{Hom}(T^{r,s}(V), W) \cong \text{Hom}(V^r, (V^*)^s; W).$$

In particular,

$$(T^{r,s}(V))^* \cong \text{Hom}(V^r, (V^*)^s; K).$$

For finite dimensional vector spaces, the duality of Section 21.5 is also easily extended to the tensor spaces $T^{r,s}(V)$. We define the pairing

$$T^{r,s}(V^*) \times T^{r,s}(V) \to K$$

as follows: if

$$v^* = v_1^* \otimes \cdots \otimes v_r^* \otimes u_{r+1}^* \otimes \cdots \otimes u_{r+s}^* \in T^{r,s}(V^*)$$

and

$$u = u_1 \otimes \cdots \otimes u_r \otimes v_{r+1}^* \otimes \cdots \otimes v_{r+s}^* \in T^{r,s}(V),$$

then

$$(v^*, u) = v_1^*(u_1) \cdots v_{r+s}^*(u_{r+s}).$$

This is a nondegenerate pairing, and thus we get a canonical isomorphism

$$(T^{r,s}(V))^* \cong T^{r,s}(V^*).$$

Consequently, we get a canonical isomorphism

$$T^{r,s}(V^*) \cong \text{Hom}(V^r, (V^*)^s; K).$$

We summarize these results in the following proposition.

**Proposition 21.18.** Let $V$ be a vector space and let

$$T^{r,s}(V) = V^\otimes r \otimes (V^*)^\otimes s = \underbrace{V \otimes \cdots \otimes V}_r \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_s.$$

We have the canonical isomorphisms

$$(T^{r,s}(V))^* \cong T^{r,s}(V^*),$$

and

$$T^{r,s}(V^*) \cong \text{Hom}(V^r, (V^*)^s; K).$$
Remark: The tensor spaces, $T^{r,s}(V)$ are also denoted $T^r_s(V)$. A tensor $\alpha \in T^{r,s}(V)$ is said to be contravariant in the first $r$ arguments and covariant in the last $s$ arguments. This terminology refers to the way tensors behave under coordinate changes. Given a basis $(e_1, \ldots, e_n)$ of $V$, if $(e_1^*, \ldots, e_n^*)$ denotes the dual basis, then every tensor $\alpha \in T^{r,s}(V)$ is given by an expression of the form

$$\alpha = \sum_{i_1, \ldots, i_r}^{j_1, \ldots, j_s} a_{i_1, \ldots, i_r}^{j_1, \ldots, j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e_{j_1}^* \otimes \cdots \otimes e_{j_s}^*.$$ 

The tradition in classical tensor notation is to use lower indices on vectors and upper indices on linear forms and in accordance to Einstein summation convention (or Einstein notation) the position of the indices on the coefficients is reversed. Einstein summation convention (already encountered in Section 21.1) is to assume that a summation is performed for all values of every index that appears simultaneously once as an upper index and once as a lower index. According to this convention, the tensor $\alpha$ above is written

$$\alpha = a_{j_1, \ldots, j_s}^{i_1, \ldots, i_r} e_{j_1} \otimes \cdots \otimes e_{i_r} \otimes e_{j_1}^* \otimes \cdots \otimes e_{j_s}^*.$$ 

An older view of tensors is that they are multidimensional arrays of coefficients, $(a_{j_1, \ldots, j_s}^{i_1, \ldots, i_r})$, subject to the rules for changes of bases.

Another operation on general tensors, contraction, is useful in differential geometry.

Definition 21.7. For all $r, s \geq 1$, the contraction $c_{i,j}: T^{r,s}(V) \rightarrow T^{r-1,s-1}(V)$, with $1 \leq i \leq r$ and $1 \leq j \leq s$, is the linear map defined on generators by

$$c_{i,j}(u_1 \otimes \cdots \otimes u_r \otimes v_1^* \otimes \cdots \otimes v_s^*) = v_j^*(u_i) u_1 \otimes \cdots \otimes \widehat{u_i} \otimes \cdots \otimes u_r \otimes v_1^* \otimes \cdots \otimes \widehat{v_j^*} \otimes \cdots \otimes v_s^*,$$

where the hat over an argument means that it should be omitted.

Let us figure out what is $c_{1,1}: T^{1,1}(V) \rightarrow \mathbb{R}$, that is $c_{1,1}: V \otimes V^* \rightarrow \mathbb{R}$. If $(e_1, \ldots, e_n)$ is a basis of $V$ and $(e_1^*, \ldots, e_n^*)$ is the dual basis, by Proposition 21.16 every $h \in V \otimes V^* \cong \text{Hom}(V,V)$ can be expressed as

$$h = \sum_{i,j=1}^{n} a_{ij} e_i \otimes e_j^*.$$ 

As

$$c_{1,1}(e_i \otimes e_j^*) = \delta_{i,j},$$

we get

$$c_{1,1}(h) = \sum_{i=1}^{n} a_{ii} = \text{tr}(h),$$

where $\text{tr}(h)$ denotes the trace of $h$. 


where \( \text{tr}(h) \) is the trace of \( h \), where \( h \) is viewed as the linear map given by the matrix, \((a_{ij})\). Actually, since \( c_{1,1} \) is defined independently of any basis, \( c_{1,1} \) provides an intrinsic definition of the trace of a linear map \( h \in \text{Hom}(V,V) \).

\textbf{Remark:} Using the Einstein summation convention, if

\[
\alpha = a_{j_1,\ldots, j_r}^{i_1,\ldots, i_r} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s},
\]

then

\[
c_{k,l}(\alpha) = a_{j_1,\ldots, j_{k-1}, j_{k+1},\ldots, j_r}^{i_1,\ldots, i_k, i_{k+1},\ldots, i_r} e_{i_1} \otimes \cdots \otimes \hat{e}_{i_k} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes \hat{e}_{j_k} \otimes \cdots \otimes e^{j_s}.
\]

If \( E \) and \( F \) are two \( K \)-algebras, we know that their tensor product \( E \otimes F \) exists as a vector space. We can make \( E \otimes F \) into an algebra as well. Indeed, we have the multilinear map

\[
E \times F \times E \times F \rightarrow E \otimes F
\]

given by \((a, b, c, d) \mapsto (ac) \otimes (bd)\), where \( ac \) is the product of \( a \) and \( c \) in \( E \) and \( bd \) is the product of \( b \) and \( d \) in \( F \). By the universal mapping property, we get a linear map,

\[
E \otimes F \otimes E \otimes F \rightarrow E \otimes F.
\]

Using the isomorphism

\[
E \otimes F \otimes E \otimes F \cong (E \otimes F) \otimes (E \otimes F),
\]

we get a linear map

\[
(E \otimes F) \otimes (E \otimes F) \rightarrow E \otimes F,
\]

and thus a bilinear map,

\[
(E \otimes F) \times (E \otimes F) \rightarrow E \otimes F
\]

which is our multiplication operation in \( E \otimes F \). This multiplication is determined by

\[
(a \otimes b) \cdot (c \otimes d) = (ac) \otimes (bd).
\]

One immediately checks that \( E \otimes F \) with this multiplication is a \( K \)-algebra.

We now turn to symmetric tensors.

\section{Symmetric Tensor Powers}

Our goal is to come up with a notion of tensor product that will allow us to treat symmetric multilinear maps as linear maps. Note that we have to restrict ourselves to a single vector space \( E \), rather then \( n \) vector spaces \( E_1, \ldots, E_n \), so that symmetry makes sense.
21.7. SYMMETRIC TENSOR POWERS

Definition 21.8. A multilinear map \( f : E^n \to F \) is symmetric iff
\[
f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = f(u_1, \ldots, u_n),
\]
for all \( u_i \in E \) and all permutations, \( \sigma : \{1, \ldots, n\} \to \{1, \ldots, n\} \). The group of permutations on \( \{1, \ldots, n\} \) (the symmetric group) is denoted \( S_n \). The vector space of all symmetric multilinear maps \( f : E^n \to F \) is denoted by \( \text{Sym}^n(E; F) \) or \( \text{Hom}_{\text{symlin}}(E^n, F) \). Note that \( \text{Sym}^1(E; F) = \text{Hom}(E, F) \).

We could proceed directly as in Theorem 21.6 and construct symmetric tensor products from scratch. However, since we already have the notion of a tensor product, there is a more economical method. First we define symmetric tensor powers.

Definition 21.9. An \( n \)-th symmetric tensor power of a vector space \( E \), where \( n \geq 1 \), is a vector space \( S \) together with a symmetric multilinear map \( \varphi : E^n \to S \) such that, for every vector space \( F \) and for every symmetric multilinear map \( f : E^n \to F \), there is a unique linear map \( f \circ \varphi : S \to F \), with
\[
f(u_1, \ldots, u_n) = f \circ \varphi(u_1, \ldots, u_n),
\]
for all \( u_1, \ldots, u_n \in E \), or for short
\[
f = f \circ \varphi.
\]
Equivalently, there is a unique linear map \( f \circ \varphi \) such that the following diagram commutes:

\[
\begin{array}{ccc}
E^n & \xrightarrow{\varphi} & S \\
\downarrow{f} & & \downarrow{f \circ \varphi} \\
F & & F
\end{array}
\]

The above property is called the universal mapping property of the symmetric tensor power \( (S, \varphi) \).

We next show that any two symmetric \( n \)-th tensor powers \( (S_1, \varphi_1) \) and \( (S_2, \varphi_2) \) for \( E \) are isomorphic.

Proposition 21.19. Given any two symmetric \( n \)-th tensor powers \( (S_1, \varphi_1) \) and \( (S_2, \varphi_2) \) for \( E \), there is an isomorphism \( h : S_1 \to S_2 \) such that
\[
\varphi_2 = h \circ \varphi_1.
\]

Proof. Replace tensor product by \( n \)-th symmetric tensor power in the proof of Proposition 21.5. \( \square \)

We now give a construction that produces a symmetric \( n \)-th tensor power of a vector space \( E \).
Theorem 21.20. Given a vector space $E$, a symmetric $n$-th tensor power $(S^n(E), \varphi)$ for $E$ can be constructed ($n \geq 1$). Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \odot \cdots \odot u_n$, the symmetric tensor power $S^n(E)$ is generated by the vectors $u_1 \odot \cdots \odot u_n$, where $u_1, \ldots, u_n \in E$, and for every symmetric multilinear map $f: E^n \to F$, the unique linear map $f_\otimes: S^n(E) \to F$ such that $f = f_\otimes \circ \varphi$ is defined by

$$f_\otimes(u_1 \odot \cdots \odot u_n) = f(u_1, \ldots, u_n)$$
on the generators $u_1 \odot \cdots \odot u_n$ of $S^n(E)$.

Proof. The tensor power $E^\otimes n$ is too big, and thus we define an appropriate quotient. Let $C$ be the subspace of $E^\otimes n$ generated by the vectors of the form

$$u_1 \otimes \cdots \otimes u_n - u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)},$$

for all $u_i \in E$, and all permutations $\sigma: \{1, \ldots, n\} \to \{1, \ldots, n\}$. We claim that the quotient space $(E^\otimes n)/C$ does the job.

Let $p: E^\otimes n \to (E^\otimes n)/C$ be the quotient map, and let $\varphi: E^n \to (E^\otimes n)/C$ be the map given by

$$\varphi = p \circ \varphi_0,$$

where $\varphi_0: E^n \to E^\otimes n$ is the injection given by $\varphi_0(u_1, \ldots, u_n) = u_1 \otimes \cdots \otimes u_n$.

Let us denote $\varphi(u_1, \ldots, u_n)$ as $u_1 \odot \cdots \odot u_n$. It is clear that $\varphi$ is symmetric. Since the vectors $u_1 \otimes \cdots \otimes u_n$ generate $E^\otimes n$, and $p$ is surjective, the vectors $u_1 \odot \cdots \odot u_n$ generate $(E^\otimes n)/C$.

It remains to show that $((E^\otimes n)/C, \varphi)$ satisfies the universal mapping property. To this end we begin by proving that there is a map $h$ such that $f = h \circ \varphi$. Given any symmetric multilinear map $f: E^n \to F$, by Theorem 21.6 there is a linear map $f_\otimes: E^\otimes n \to F$ such that $f = f_\otimes \circ \varphi_0$, as in the diagram below:

$$\begin{CD}
E^n @>\varphi_0>> E^\otimes n \\
@VfVV @Vf_\otimes VV \\
F @VVhV \\
\end{CD}$$

However, since $f$ is symmetric, we have $f_\otimes(z) = 0$ for every $z \in C$. Thus, we get an induced linear map $h: (E^\otimes n)/C \to F$ making the following diagram commute:

$$\begin{CD}
E^n @>\varphi_0>> (E^\otimes n)/C \\
@VfVV @VVpV \\
F @VhV \\
\end{CD}$$
if we define $h([z]) = f \odot (z)$ for every $z \in E^\otimes n$, where $[z]$ is the equivalence class in $(E^\otimes n)/C$ of $z \in E^\otimes n$: The above diagram shows that $f = h \circ p \circ \varphi_0 = h \circ \varphi$. We now prove the uniqueness of $h$. For any linear map $f \odot : (E^\otimes n)/C \to F$ such that $f = f \odot \circ \varphi$, since $\varphi(u_1, \ldots, u_n) = u_1 \odot \cdots \odot u_n$ and the vectors $u_1 \odot \cdots \odot u_n$ generate $(E^\otimes n)/C$, the map $f \odot$ is uniquely defined by

$$f \odot (u_1 \odot \cdots \odot u_n) = f(u_1, \ldots, u_n).$$

Since $f = h \circ \varphi$, the map $h$ is unique, and we let $f \odot = h$. Thus, $S^n(E) = (E^\otimes n)/C$ and $\varphi$ constitute a symmetric $n$-th tensor power of $E$.

The map $\varphi$ from $E^n$ to $S^n(E)$ is often denoted $\iota \odot$, so that

$$\iota \odot (u_1, \ldots, u_n) = u_1 \odot \cdots \odot u_n.$$

Again, the actual construction is not important. What is important is that the symmetric $n$-th power has the universal mapping property with respect to symmetric multilinear maps.

**Remark:** The notation $\odot$ for the commutative multiplication of symmetric tensor powers is not standard. Another notation commonly used is $\cdot$. We often abbreviate “symmetric tensor power” as “symmetric power.” The symmetric power $S^n(E)$ is also denoted Sym$^n E$ but we prefer to use the notation Sym to denote spaces of symmetric multilinear maps. To be consistent with the use of $\odot$, we could have used the notation $\bigodot^n E$. Clearly, $S^1(E) \cong E$ and it is convenient to set $S^0(E) = K$.

The fact that the map $\varphi : E^n \to S^n(E)$ is symmetric and multilinear can also be expressed as follows:

$$u_1 \odot \cdots \odot (v_i + w_i) \odot \cdots \odot u_n = (u_1 \odot \cdots \odot v_i \odot \cdots \odot u_n) + (u_1 \odot \cdots \odot w_i \odot \cdots \odot u_n),$$

$$u_1 \odot \cdots \odot (\lambda u_i) \odot \cdots \odot u_n = \lambda(u_1 \odot \cdots \odot u_i \odot \cdots \odot u_n),$$

$$u_{\sigma(1)} \odot \cdots \odot u_{\sigma(n)} = u_1 \odot \cdots \odot u_n,$$

for all permutations $\sigma \in S_n$.

The last identity shows that the “operation” $\odot$ is commutative. This allows us to view the symmetric tensor $u_1 \odot \cdots \odot u_n$ as an object called a multiset.

Given a set $A$, a multiset with elements from $A$ is a generalization of the concept of a set that allows multiple instances of elements from $A$ to occur. For example, if $A = \{a, b, c\}$, the following are multisets:

$$M_1 = \{a, a, b\}, \quad M_2 = \{a, a, b, b, c\}, \quad M_3 = \{a, a, b, b, c, d, d, d\}.$$  

Here is another way to represent multisets as tables showing the multiplicities of the elements in the multiset:

$$M_1 = \begin{pmatrix} a & b & c & d \\ 2 & 1 & 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} a & b & c & d \\ 2 & 2 & 1 & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} a & b & c & d \\ 2 & 2 & 1 & 3 \end{pmatrix}.$$
The above are just graphs of functions from the set \( A = \{a, b, c, d\} \) to \( \mathbb{N} \). This suggests that formally, a finite multiset \( M \) over a set \( A \) is a function \( M: A \to \mathbb{N} \) such that \( M(a) \neq 0 \) for finitely many \( a \in A \). The multiplicity of an element \( a \in A \) in \( M \) is \( M(a) \). The set of all multisets over \( A \) is denoted by \( \mathbb{N}^{(A)} \), and we let \( \text{dom}(M) = \{a \in A \mid M(a) \neq 0\} \), which is a finite set. The set \( \text{dom}(M) \) is the set of elements in \( A \) that actually occur in \( M \). For any multiset \( M \in \mathbb{N}^{(A)} \), note that \( \sum_{a \in A} M(a) \) makes sense, since \( \sum_{a \in \text{dom}(a)} M(a) \), and \( \text{dom}(M) \) is finite; this sum is the total number of elements in the multiset \( A \) and is called the size of \( M \). Let \( |M| = \sum_{a \in A} M(a) \).

Going back to our symmetric tensors, we can view the tensors of the form \( u_1 \odot \cdots \odot u_n \) as multisets of size \( n \) over the set \( E \).

Theorem 21.20 implies the following proposition.

**Proposition 21.21.** There is a canonical isomorphism

\[
\text{Hom}(S^n(E), F) \cong \text{Sym}^n(E; F),
\]

between the vector space of linear maps \( \text{Hom}(S^n(E), F) \) and the vector space of symmetric multilinear maps \( \text{Sym}^n(E; F) \) given by the linear map \( - \circ \varphi \) defined by \( h \mapsto h \circ \varphi \), with \( h \in \text{Hom}(S^n(E), F) \).

**Proof.** The map \( h \circ \varphi \) is clearly symmetric multilinear. By Theorem 21.20, for every symmetric multilinear map \( f \in \text{Sym}^n(E; F) \) there is a unique linear map \( f_\circ \in \text{Hom}(S^n(E), F) \) such that \( f = f_\circ \circ \varphi \), so the map \( - \circ \varphi \) is bijective. Its inverse is the map \( f \mapsto f_\circ \). \( \square \)

In particular, when \( F = K \), we get the following important fact:

**Proposition 21.22.** There is a canonical isomorphism

\[
(S^n(E))^* \cong \text{Sym}^n(E; K).
\]

Symmetric tensors in \( S^n(E) \) are also called symmetric \( n \)-tensors, and tensors of the form \( u_1 \odot \cdots \odot u_n \), where \( u_i \in E \), are called simple (or decomposable) symmetric \( n \)-tensors. Those symmetric \( n \)-tensors that are not simple are often called compound symmetric \( n \)-tensors.

Given two linear maps \( f: E \to E' \) and \( g: E \to E' \), since the map \( \iota_\circ \circ (f \times g) \) is bilinear and symmetric, there is a unique linear map \( f \circ g: S^2(E) \to S^2(E') \) making the following diagram commute:

\[
\begin{array}{ccc}
E^2 & \xrightarrow{\iota_\circ \circ (f \times g)} & S^2(E) \\
\downarrow{f \times g} & & \downarrow{f \circ g} \\
(E')^2 & \xrightarrow{\iota_\circ} & S^2(E').
\end{array}
\]

Observe that \( f \circ g \) is determined by

\[(f \circ g)(u \odot v) = f(u) \odot g(u).\]
If we also have linear maps \( f': E' \to E'' \) and \( g': E' \to E'' \), we can easily verify that
\[
(f' \circ f) \circ (g' \circ g) = (f' \circ g') \circ (f \circ g).
\]

The generalization to the symmetric tensor product \( f_1 \circ \cdots \circ f_n \) of \( n \geq 3 \) linear maps \( f_i: E \to E' \) is immediate, and left to the reader.

## 21.8 Bases of Symmetric Powers

The vectors \( u_1 \circ \cdots \circ u_m \) where \( u_1, \ldots, u_m \in E \) generate \( S^m(E) \), but they are not linearly independent. We will prove a version of Proposition 21.11 for symmetric tensor powers using multisets.

Recall that a (finite) multiset over a set \( I \) is a function \( M: I \to \mathbb{N} \), such that \( M(i) \neq 0 \) for finitely many \( i \in I \). The set of all multisets over \( I \) is denoted as \( \mathbb{N}^{(I)} \) and we let \( \text{dom}(M) = \{i \in I \mid M(i) \neq 0\} \), the finite set of elements in \( I \) that actually occur in \( M \). The size of the multiset \( M \) is \( |M| = \sum_{a \in I} M(a) \).

To explain the idea of the proof, consider the case when \( m = 2 \) and \( E \) has dimension 3. Given a basis \((e_1, e_2, e_3)\) of \( E \), we would like to prove that
\[
e_1 \circ e_1, \quad e_1 \circ e_2, \quad e_1 \circ e_3, \quad e_2 \circ e_2, \quad e_2 \circ e_3, \quad e_3 \circ e_3
\]
are linearly independent. To prove this, it suffices to show that for any vector space \( F \), if \( w_{11}, w_{12}, w_{13}, w_{22}, w_{23}, w_{33} \) are any vectors in \( F \), then there is a symmetric bilinear map \( h: E^2 \to F \) such that
\[
h(e_i, e_j) = w_{ij}, \quad 1 \leq i \leq j \leq 3.
\]
Because \( h \) yields a unique linear map \( h_\circ: S^2E \to F \) such that
\[
h_\circ(e_i \circ e_j) = w_{ij}, \quad 1 \leq i \leq j \leq 3,
\]
by Proposition 21.4, the vectors
\[
e_1 \circ e_1, \quad e_1 \circ e_2, \quad e_1 \circ e_3, \quad e_2 \circ e_2, \quad e_2 \circ e_3, \quad e_3 \circ e_3
\]
are linearly independent. This suggests understanding how a symmetric bilinear function \( f: E^2 \to F \) is expressed in terms of its values \( f(e_i, e_j) \) on the basis vectors \((e_1, e_2, e_3)\), and this can be done easily. Using bilinearity and symmetry, we obtain
\[
f(u_1e_1 + u_2e_2 + u_3e_3, v_1e_1 + v_2e_2 + v_3e_3) = u_1v_1f(e_1, e_1) + (u_1v_2 + u_2v_1)f(e_1, e_2)
+ (u_1v_3 + u_3v_1)f(e_1, e_3) + u_2v_2f(e_2, e_2)
+ (u_2v_3 + u_3v_2)f(e_2, e_3) + u_3v_3f(e_3, e_3).
\]
Therefore, given \( w_{11}, w_{12}, w_{13}, w_{22}, w_{23}, w_{33} \in F \), the function \( h \) given by

\[
h(u_1e_1 + u_2e_2 + u_3e_3, v_1e_1 + v_2e_2 + v_3e_3) = u_1v_1w_{11} + (u_1v_2 + u_2v_1)w_{12} \\
+ (u_1v_3 + u_3v_1)w_{13} + u_2v_2w_{22} \\
+ (u_2v_3 + u_3v_2)w_{23} + u_3v_3w_{33}
\]

is clearly bilinear symmetric, and by construction \( h(e_i, e_j) = w_{ij} \), so it does the job.

The generalization of this argument to any \( m \geq 2 \) and to a space \( E \) of any dimension (even infinite) is conceptually clear, but notationally messy. If \( \dim(E) = n \) and if \((e_1, \ldots, e_n)\) is a basis of \( E \), for any \( m \) vectors \( v_j = \sum_{i=1}^{n} u_{i,j}e_i \) in \( F \), for any symmetric multilinear map \( f: E^m \to F \), we have

\[
f(v_1, \ldots, v_m) = \sum_{k_1 + \cdots + k_n = m} \left( \sum_{I_1 \cup \cdots \cup I_n = \{1, \ldots, m\}} \left( \prod_{i_1 \in I_1} u_{1,i_1} \right) \cdots \left( \prod_{i_n \in I_n} u_{n,i_n} \right) \right) f(e_{i_1}, \ldots, e_{i_{k_1}}, \ldots, e_{i_{k_2}}, \ldots, e_{i_{k_n}}).
\]

Given any set \( J \) of \( n \geq 1 \) elements, say \( J = \{j_1, \ldots, j_n\} \), and given any \( m \geq 2 \), for any sequence \((k_1, \ldots, k_n)\) of natural numbers \( k_i \in \mathbb{N} \) such that \( k_1 + \cdots + k_n = m \), the multiset \( M \) of size \( m \)

\[
M = \{j_{1k_1}, \ldots, j_{1k_{k_1}}, j_{2k_2}, \ldots, j_{2k_{k_2}}, \ldots, j_{nk_n}, \ldots, j_{nk_{k_n}}\}
\]

is denoted by \( M(m, J, k_1, \ldots, k_n) \). Note that \( M(j_i) = k_i \), for \( i = 1, \ldots, n \). Given any \( k \geq 1 \), and any \( u \in E \), we denote \( u \odot \cdots \odot u \) as \( u^{\odot k} \).

We can now prove the following proposition.

**Proposition 21.23.** Given a vector space \( E \), if \( (e_i)_{i \in I} \) is a basis for \( E \), then the family of vectors

\[
\left( e_{i_1}^{\odot M(i_1)} \odot \cdots \odot e_{i_k}^{\odot M(i_k)} \right)_{M \in \mathbb{N}^{(I)}, |M| = m, \{i_1, \ldots, i_k\} = \text{dom}(M)}
\]

is a basis of the symmetric \( m \)-th tensor power \( S^m(E) \).

**Proof.** The proof is very similar to that of Proposition 21.11. First assume that \( E \) has finite dimension \( n \). In this case \( I = \{1, \ldots, n\} \), and any multiset \( M \in \mathbb{N}^{(I)} \) of size \(|M| = m\) is of the form \( M(m, \{1, \ldots, n\}, k_1, \ldots, k_n) \), with \( k_i = M(i) \) and \( k_1 + \cdots + k_n = m \).

For any nontrivial vector space \( F \), for any family of vectors

\[
(w_M)_{M \in \mathbb{N}^{(I)}, |M| = m},
\]

we show the existence of a symmetric multilinear map \( h: S^m(E) \to F \), such that for every \( M \in \mathbb{N}^{(I)} \) with \(|M| = m\), we have

\[
h(e^{\otimes M(i_1)}_{i_1} \otimes \cdots \otimes e^{\otimes M(i_k)}_{i_k}) = w_M,
\]

where \( \{i_1, \ldots, i_k\} = \text{dom}(M) \). We define the map \( f: E^m \to F \) as follows: for any \( m \) vectors \( v_1, \ldots, v_m \in F \) we can write \( v_k = \sum_{i=1}^n u_{i,k} e_i \) for \( k = 1, \ldots, m \) and we set

\[
f(v_1, \ldots, v_m)
= \sum_{k_1+\cdots+k_n=m} \left( \sum_{I_1 \cup \cdots \cup I_n = \{1, \ldots, m\}} \left( \prod_{i_1 \in I_1} u_{1,i_1} \right) \cdots \left( \prod_{i_n \in I_n} u_{n,i_n} \right) \right) w_M(m,\{1,\ldots,n\},k_1,\ldots,k_n).
\]

It is not difficult to verify that \( f \) is symmetric and multilinear. By the universal mapping property of the symmetric tensor product, the linear map \( f: S^m(E) \to F \) such that \( f = f \circ \varphi \), is the desired map \( h \). Then, by Proposition 21.4, it follows that the family

\[
\left( e^{\otimes M(i_1)}_{i_1} \otimes \cdots \otimes e^{\otimes M(i_k)}_{i_k} \right)_{M \in \mathbb{N}^{(I)}, |M| = m, \{i_1, \ldots, i_k\} = \text{dom}(M)}
\]

is linearly independent. Using the commutativity of \( \otimes \), we can also show that these vectors generate \( S^m(E) \), and thus, they form a basis for \( S^m(E) \).

If \( I \) is infinite dimensional, then for any \( m \) vectors \( v_1, \ldots, v_m \in F \) there is a finite subset \( J \) of \( I \) such that \( v_k = \sum_{j \in J} u_{j,k} e_j \) for \( k = 1, \ldots, m \), and if we write \( n = |J| \), then the formula for \( f(v_1, \ldots, v_m) \) is obtained by replacing the set \( \{1, \ldots, n\} \) by \( J \). The details are left as an exercise.

As a consequence, when \( I \) is finite, say of size \( p = \dim(E) \), the dimension of \( S^m(E) \) is the number of finite multisets \( (j_1, \ldots, j_p) \), such that \( j_1 + \cdots + j_p = m \), \( j_k \geq 0 \). We leave as an exercise to show that this number is \( \binom{p+m-1}{m} \). Thus, if \( \dim(E) = p \), then the dimension of \( S^m(E) \) is \( \binom{p+m-1}{m} \). Compare with the dimension of \( E^\otimes m \), which is \( p^m \). In particular, when \( p = 2 \), the dimension of \( S^m(E) \) is \( m+1 \). This can also be seen directly.

**Remark:** The number \( \binom{p+m-1}{m} \) is also the number of homogeneous monomials

\[
X_1^{j_1} \cdots X_p^{j_p}
\]

of total degree \( m \) in \( p \) variables (we have \( j_1 + \cdots + j_p = m \)). This is not a coincidence! Given a vector space \( E \) and a basis \( (e_i)_{i \in I} \) for \( E \), Proposition 21.23 shows that every symmetric tensor \( z \in S^m(E) \) can be written in a unique way as

\[
z = \sum_{M \in \mathbb{N}^{(I)}, \sum_{i \in I} M(i) = m, \{i_1, \ldots, i_k\} = \text{dom}(M)} \lambda_M e^{\otimes M(i_1)}_{i_1} \otimes \cdots \otimes e^{\otimes M(i_k)}_{i_k},
\]
for some unique family of scalars \( \lambda_M \in K \), all zero except for a finite number.

This looks like a homogeneous polynomial of total degree \( m \), where the monomials of total degree \( m \) are the symmetric tensors \( e_i^{\otimes M(i_1)} \otimes \cdots \otimes e_i^{\otimes M(i_k)} \) in the “indeterminates” \( e_i \), where \( i \in I \) (recall that \( M(i_1) + \cdots + M(i_k) = m \)) and implies that polynomials can be defined in terms of symmetric tensors.

## 21.9 Some Useful Isomorphisms for Symmetric Powers

We can show the following property of the symmetric tensor product, using the proof technique of Proposition 21.12:

\[
S^n(E \oplus F) \cong \bigoplus_{k=0}^{n} S^k(E) \otimes S^{n-k}(F).
\]

## 21.10 Duality for Symmetric Powers

In this section all vector spaces are assumed to have finite dimension. We define a nondegenerate pairing \( S^n(E^*) \times S^n(E) \rightarrow K \) as follows: Consider the multilinear map

\[
(E^*)^n \times E^n \rightarrow K
\]

given by

\[
(v_1^*, \ldots, v_n^*, u_1, \ldots, u_n) \mapsto \sum_{\sigma \in S_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).
\]

Note that the expression on the right-hand side is “almost” the determinant \( \det(v_j^*(u_i)) \), except that the sign \( \text{sgn}(\sigma) \) is missing (where \( \text{sgn}(\sigma) \) is the signature of the permutation \( \sigma \); that is, the parity of the number of transpositions into which \( \sigma \) can be factored). Such an expression is called a *permanent*.

It can be verified that this expression is symmetric w.r.t. the \( u_i \)'s and also w.r.t. the \( v_j^* \). For any fixed \( (v_1^*, \ldots, v_n^*) \in (E^*)^n \), we get a symmetric multilinear map

\[
l_{v_1^*, \ldots, v_n^*} : (u_1, \ldots, u_n) \mapsto \sum_{\sigma \in S_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n)
\]

from \( E^n \) to \( K \). The map \( l_{v_1^*, \ldots, v_n^*} \) extends uniquely to a linear map \( L_{v_1^*, \ldots, v_n^*} : S^n(E) \rightarrow K \) making the following diagram commute:

\[
\begin{array}{ccc}
E^n & \xrightarrow{l_{v_1^*, \ldots, v_n^*}} & S^n(E) \\
\downarrow & & \downarrow \\
K & & K
\end{array}
\]
We also have the symmetric multilinear map
\[(v_1^*, \ldots, v_n^*) \mapsto L_{v_1^* \ldots v_n^*}\]
from \((E^*)^n\) to \(\text{Hom}(S^n(E), K)\), which extends to a linear map \(L\) from \(S^n(E^*)\) to \(\text{Hom}(S^n(E), K)\) making the following diagram commute:

\[
\begin{array}{ccc}
(E^*)^n & \xrightarrow{\iota \otimes^*} & S^n(E^*) \\
& \downarrow & \downarrow L \\
& & \text{Hom}(S^n(E), K).
\end{array}
\]

However, in view of the isomorphism
\[
\text{Hom}(U \otimes V, W) \cong \text{Hom}(U, \text{Hom}(V, W)),
\]
with \(U = S^n(E^*)\), \(V = S^n(E)\) and \(W = K\), we can view \(L\) as a linear map
\[
L: S^n(E^*) \otimes S^n(E) \rightarrow K,
\]
which by Proposition 21.8 corresponds to a bilinear map
\[
S^n(E^*) \times S^n(E) \rightarrow K.
\]

Now this pairing is nondegenerate. This can be shown using bases and we leave it as an exercise to the reader (see Knapp [106], Appendix A). Therefore we get a canonical isomorphism
\[
(S^n(E))^* \cong S^n(E^*).
\]
The following proposition summarizes the duality properties of symmetric powers.

**Proposition 21.24.** We have the canonical isomorphisms
\[
(S^n(E))^* \cong S^n(E^*)
\]
and
\[
S^n(E^*) \cong \text{Sym}^n(E; K) = \text{Hom}_{\text{symlin}}(E^n, K),
\]
which allows us to interpret symmetric tensors over \(E^*\) as symmetric multilinear maps.

**Proof.** The isomorphism
\[
\mu: S^n(E^*) \cong \text{Sym}^n(E; K)
\]
follows from the isomorphisms \((S^n(E))^* \cong S^n(E^*)\) and \((S^n(E))^* \cong \text{Sym}^n(E; K)\) given by Proposition 21.22. \(\square\)
Remark: The isomorphism \( \mu: S^n(E^*) \cong \text{Sym}^n(E;K) \) discussed above can be described explicitly as the linear extension of the map given by

\[
\mu(v_1^* \odot \cdots \odot v_n^*)(u_1, \ldots, u_n) = \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n).
\]

The map from \( E^n \) to \( S^n(E) \) given by \( (u_1, \ldots, u_n) \mapsto u_1 \odot \cdots \odot u_n \) yields a surjection \( \pi: E \otimes^n \to S^n(E) \). Because we are dealing with vector spaces, this map has some section; that is, there is some injection \( \eta: S^n(E) \to E \otimes^n \) with \( \pi \circ \eta = \text{id} \). If our field \( K \) has characteristic 0, then there is a special section having a natural definition involving a symmetrization process defined as follows: For every permutation \( \sigma \), we have the map \( r_\sigma: E^n \to E \otimes^n \) given by

\[
r_\sigma(u_1, \ldots, u_n) = u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.
\]

As \( r_\sigma \) is clearly multilinear, \( r_\sigma \) extends to a linear map \( (r_\sigma)_\otimes: E \otimes^n \to E \otimes^n \) making the following diagram commute

\[
\begin{array}{ccc}
E^n & \xrightarrow{\iota_\otimes} & E \otimes^n \\
\downarrow{r_\sigma} & & \downarrow{(r_\sigma)_\otimes} \\
E \otimes^n & \xrightarrow{(r_\sigma)_\otimes} & E \otimes^n,
\end{array}
\]

and we get a map \( \mathfrak{S}_n \times E \otimes^n \to E \otimes^n \), namely

\[
\sigma \cdot z = (r_\sigma)_\otimes(z).
\]

It is immediately checked that this is a left action of the symmetric group \( \mathfrak{S}_n \) on \( E \otimes^n \), and the tensors \( z \in E \otimes^n \) such that

\[
\sigma \cdot z = z, \quad \text{for all } \sigma \in \mathfrak{S}_n
\]

are called symmetrized tensors.

We define the map \( \eta: E^n \to E \otimes^n \) by

\[
\eta(u_1, \ldots, u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} \sigma \cdot (u_1 \otimes \cdots \otimes u_n) = \frac{1}{n!} \sum_{\sigma \in \mathfrak{S}_n} u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.
\]

As the right hand side is clearly symmetric, we get a linear map \( \eta_\otimes: S^n(E) \to E \otimes^n \) making the following diagram commute:

\[
\begin{array}{ccc}
E^n & \xrightarrow{\iota_\otimes} & S^n(E) \\
\downarrow{\eta} & & \downarrow{\eta_\otimes} \\
E \otimes^n & \xrightarrow{\eta_\otimes} & E \otimes^n,
\end{array}
\]
Clearly, \( \eta \circ (S^n(E)) \) is the set of symmetrized tensors in \( E^\otimes n \). If we consider the map \( S = \eta \circ \pi: E^\otimes n \to E^\otimes n \) where \( \pi \) is the surjection \( E^\otimes n \to S^n(E) \), it is easy to check that \( S \circ S = S \). Therefore, \( S \) is a projection, and by linear algebra, we know that

\[
E^\otimes n = S(E^\otimes n) \oplus \ker S = \eta \circ (S^n(E)) \oplus \ker S.
\]

It turns out that \( \ker S = E^\otimes n \cap \mathfrak{I} = \ker \pi \), where \( \mathfrak{I} \) is the two-sided ideal of \( T(E) \) generated by all tensors of the form \( u \otimes v - v \otimes u \in E^\otimes 2 \) (for example, see Knapp [106], Appendix A). Therefore, \( \eta \circ \) is injective,

\[
E^\otimes n = \eta \circ (S^n(E)) \oplus (E^\otimes n \cap \mathfrak{I}) = \eta \circ (S^n(E)) \oplus \ker \pi,
\]

and the symmetric tensor power \( S^n(E) \) is naturally embedded into \( E^\otimes n \).

### 21.11 Symmetric Algebras

As in the case of tensors, we can pack together all the symmetric powers \( S^n(V) \) into an algebra

\[
S(V) = \bigoplus_{m \geq 0} S^m(V),
\]

called the symmetric tensor algebra of \( V \). We could adapt what we did in Section 21.6 for general tensor powers to symmetric tensors but since we already have the algebra \( T(V) \), we can proceed faster. If \( \mathfrak{I} \) is the two-sided ideal generated by all tensors of the form \( u \otimes v - v \otimes u \in V^\otimes 2 \), we set

\[
S^*(V) = T(V)/\mathfrak{I}.
\]

Observe that since the ideal \( \mathfrak{I} \) is generated by elements in \( V^\otimes 2 \), every tensor in \( \mathfrak{I} \) is a linear combination of tensors of the form \( \omega_1 \otimes (u \otimes v - v \otimes u) \otimes \omega_2 \), with \( \omega_1 \in V^\otimes n_1 \) and \( \omega_2 \in V^\otimes n_2 \) for some \( n_1, n_2 \in \mathbb{N} \), which implies that

\[
\mathfrak{I} = \bigoplus_{m \geq 0} (\mathfrak{I} \cap V^\otimes m).
\]

Then, \( S^*(V) \) automatically inherits a multiplication operation which is commutative, and since \( T(V) \) is graded, that is

\[
T(V) = \bigoplus_{m \geq 0} V^\otimes m,
\]

we have

\[
S^*(V) = \bigoplus_{m \geq 0} V^\otimes m / (\mathfrak{I} \cap V^\otimes m).
\]

However, it is easy to check that

\[
S^m(V) \cong V^\otimes m / (\mathfrak{I} \cap V^\otimes m),
\]
so

\[ S^\bullet(V) \cong S(V). \]

When \( V \) is of finite dimension \( n \), \( S(V) \) corresponds to the algebra of polynomials with coefficients in \( K \) in \( n \) variables (this can be seen from Proposition 21.23). When \( V \) is of infinite dimension and \((u_i)_{i \in I}\) is a basis of \( V \), the algebra \( S(V) \) corresponds to the algebra of polynomials in infinitely many variables in \( I \). What’s nice about the symmetric tensor algebra \( S(V) \) is that it provides an intrinsic definition of a polynomial algebra in any set of \( I \) variables.

It is also easy to see that \( S(V) \) satisfies the following universal mapping property:

**Proposition 21.25.** Given any commutative \( K \)-algebra \( A \), for any linear map \( f : V \to A \), there is a unique \( K \)-algebra homomorphism \( \overline{f} : S(V) \to A \) so that

\[ f = \overline{f} \circ i, \]

as in the diagram below:

\[
\begin{array}{ccc}
V & \longrightarrow & S(V) \\
\downarrow f & & \downarrow \overline{f} \\
& \searrow & \\
& A & \\
\end{array}
\]

**Remark:** If \( E \) is finite-dimensional, recall the isomorphism \( \mu : S^n(E^*) \to \text{Sym}^n(E; K) \) defined as the linear extension of the map given by

\[ \mu(v_1^* \circ \cdots \circ v_n^*)(u_1, \ldots, u_n) = \sum_{\sigma \in \mathfrak{S}_n} v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n). \]

Now we have also a multiplication operation \( S^m(E^*) \times S^n(E^*) \to S^{m+n}(E^*) \). The following question then arises:

Can we define a multiplication \( \text{Sym}^m(E; K) \times \text{Sym}^n(E; K) \to \text{Sym}^{m+n}(E; K) \) directly on symmetric multilinear forms, so that the following diagram commutes:

\[
\begin{array}{ccc}
S^m(E^*) \times S^n(E^*) & \xrightarrow{\circ} & S^{m+n}(E^*) \\
\downarrow \mu_m \times \mu_n & & \downarrow \mu_{m+n} \\
\text{Sym}^m(E; K) \times \text{Sym}^n(E; K) & \longrightarrow & \text{Sym}^{m+n}(E; K).
\end{array}
\]

The answer is yes! The solution is to define this multiplication such that for \( f \in \text{Sym}^m(E; K) \) and \( g \in \text{Sym}^n(E; K) \),

\[ (f \cdot g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \text{shuffle}(m,n)} f(u_{\sigma(1)}, \ldots, u_{\sigma(m)})g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}), \quad (*) \]
where shuffle\((m, n)\) consists of all \((m, n)\)-"shuffles;" that is, permutations \(\sigma\) of \(\{1, \ldots, m+n\}\) such that \(\sigma(1) < \cdots < \sigma(m)\) and \(\sigma(m+1) < \cdots < \sigma(m+n)\). Observe that a \((m, n)\)-shuffle is completely determined by the sequence \(\sigma(1) < \cdots < \sigma(m)\).

For example, suppose \(m = 2\) and \(n = 1\). Given \(v_1^*, v_2^*, v_3^* \in E^*\), the multiplication structure on \(S(E^*)\) implies that \((v_1^* \circ v_2^*) \cdot v_3^* = v_1^* \circ v_2^* \circ v_3^* \in S^3(E^*)\). Furthermore, for \(u_1, u_2, u_3, \in E\),

\[
\mu_3(v_1^* \circ v_2^* \circ v_3^*)(u_1, u_2, u_3) = \sum_{\sigma \in \text{shuffle}(2,1)} v_{\sigma(1)}^*(u_1) v_{\sigma(2)}^*(u_2) v_{\sigma(3)}^*(u_3)
\]

\[
= v_1^*(u_1) v_2^*(u_2) v_3^*(u_3) + v_1^*(u_1) v_3^*(u_2) v_2^*(u_3) + v_2^*(u_1) v_1^*(u_2) v_3^*(u_3) + v_2^*(u_1) v_3^*(u_2) v_1^*(u_3) + v_3^*(u_1) v_1^*(u_2) v_2^*(u_3) + v_3^*(u_1) v_2^*(u_2) v_1^*(u_3).
\]

Now the \((2, 1)\)-shuffles of \(\{1, 2, 3\}\) are the following three permutations, namely

\[
\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.
\]

If \(f \cong \mu_2(v_1^* \circ v_2^*)\) and \(g \cong \mu_1(v_3^*)\), then (*) implies that

\[
(f \cdot g)(u_1, u_2, u_3) = \sum_{\sigma \in \text{shuffle}(2,1)} f(u_{\sigma(1)}, u_{\sigma(2)})g(u_{\sigma(3)})
\]

\[
= f(u_1, u_2)g(u_3) + f(u_1, u_3)g(u_2) + f(u_2, u_3)g(u_1)
\]

\[
= \mu_2(v_1^* \circ v_2^*)(u_1, u_2)\mu_1(v_3^*)(u_3) + \mu_2(v_1^* \circ v_2^*)(u_1, u_3)\mu_1(v_3^*)(u_2) + \mu_2(v_1^* \circ v_2^*)(u_2, u_3)\mu_1(v_3^*)(u_1)
\]

\[
= (v_1^*(u_1)v_2^*(u_2) + v_2^*(u_1)v_1^*(u_2))v_3^*(u_3) + (v_1^*(u_1)v_3^*(u_2) + v_3^*(u_1)v_1^*(u_2))v_2^*(u_3) + (v_2^*(u_1)v_3^*(u_2) + v_3^*(u_1)v_2^*(u_2))v_1^*(u_3)
\]

\[
= \mu_3(v_1^* \circ v_2^* \circ v_3^*)(u_1, u_2, u_3).
\]

We leave it as an exercise for the reader to prove Equation (*).

Another useful canonical isomorphism (of \(K\)-algebras) is

\[
S(E \oplus F) \cong S(E) \otimes S(F).
\]

### 21.12 Tensor Products of Modules over a Commutative Ring

This section provides some background on modules which is needed for Section 28.4 about metrics on vector bundles and for Chapter 29 on connections and curvature on vector bundles.
What happens is that given a manifold \( M \), the space \( \mathfrak{X}(M) \) of vector fields on \( M \) and the space \( \mathcal{A}^p(M) \) of differential \( p \)-forms on \( M \) are vector spaces, but vector fields and \( p \)-forms can also be multiplied by smooth functions in \( C^\infty(M) \). This operation is a left action of \( C^\infty(M) \) which satisfies all the axioms of the scalar multiplication in a vector space, but since \( C^\infty(M) \) is not a field, the resulting structure is not a vector space. Instead it is a module, a more general notion.

**Definition 21.10.** If \( R \) is a commutative ring with identity (say 1), a *module over \( R \) (or \( R \)-module)* is an abelian group \( M \) with a scalar multiplication \( \cdot : R \times M \to M \) such that all the axioms of a vector space are satisfied.

At first glance, a module does not seem any different from a vector space, but the lack of multiplicative inverses in \( R \) has drastic consequences, one being that unlike vector spaces, modules are generally not free; that is, have no bases. Furthermore, a module may have *torsion elements*, that is, elements \( m \in M \) such that \( \lambda \cdot m = 0 \), even though \( m \neq 0 \) and \( \lambda \neq 0 \). For example, for any nonzero integer \( n \in \mathbb{Z} \), the \( \mathbb{Z} \)-module \( \mathbb{Z}/n\mathbb{Z} \) has no basis and \( n \cdot \overline{m} = 0 \) for all \( \overline{m} \in \mathbb{Z}/n\mathbb{Z} \). Similarly, \( \mathbb{Q} \) as a \( \mathbb{Z} \)-module has no basis. In fact, any two distinct nonzero elements \( p_1/q_1 \) and \( p_2/q_2 \) are linearly dependent, since

\[
(p_2q_1)
\begin{pmatrix}
p_1 \\
q_1
\end{pmatrix}
-(p_1q_2)
\begin{pmatrix}
p_2 \\
q_2
\end{pmatrix}
= 0.
\]

Nevertheless, it is possible to define tensor products of modules over a ring, just as in Section 21.2, and the results of this section continue to hold. The results of Section 21.4 also continue to hold since they are based on the universal mapping property. However, the results of Section 21.3 on bases generally fail, except for free modules. Similarly, the results of Section 21.5 on duality generally fail. Tensor algebras can be defined for modules, as in Section 21.6. Symmetric tensor and alternating tensors can be defined for modules, but again, results involving bases generally fail.

Tensor products of modules have some unexpected properties. For example, if \( p \) and \( q \) are relatively prime integers, then

\[
\mathbb{Z}/p\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/q\mathbb{Z} = (0).
\]

This is because, by Bezout’s identity, there are \( a, b \in \mathbb{Z} \) such that

\[ ap + bq = 1, \]

so, for all \( x \in \mathbb{Z}/p\mathbb{Z} \) and all \( y \in \mathbb{Z}/q\mathbb{Z} \), we have

\[
x \otimes y = ap(x \otimes y) + bq(x \otimes y) \\
= a(px \otimes y) + b(x \otimes qy) \\
= a(0 \otimes y) + b(x \otimes 0) \\
= 0.
\]
It is possible to salvage certain properties of tensor products holding for vector spaces by restricting the class of modules under consideration. For example, projective modules have a pretty good behavior w.r.t. tensor products.

A free $R$-module $F$ is a module that has a basis (i.e., there is a family $(e_i)_{i \in I}$ of linearly independent vectors in $F$ that span $F$). Projective modules have many equivalent characterizations. Here is one that is best suited for our needs:

**Definition 21.11.** An $R$-module $P$ is projective if it is a summand of a free module; that is, if there is a free $R$-module $F$, and some $R$-module $Q$, so that

$$F = P \oplus Q.$$  

For example, we show in Section 28.4 that the space $\Gamma(\xi)$ of global sections of a vector bundle $\xi$ over a base manifold $B$ is a finitely generated $C^\infty(B)$-projective module.

Given any $R$-module $M$, we let $M^* = \text{Hom}_R(M, R)$ be its dual. We have the following proposition:

**Proposition 21.26.** For any finitely-generated projective $R$-modules $P$ and any $R$-module $Q$, we have the isomorphisms:

$$P^{**} \cong P,$$

$$\text{Hom}_R(P, Q) \cong P^* \otimes_R Q.$$

**Proof sketch.** We only consider the second isomorphism. Since $P$ is projective, we have some $R$-modules $P_1, F$ with

$$P \oplus P_1 = F,$$

where $F$ is some free module. Now, we know that for any $R$-modules $U, V, W$, we have

$$\text{Hom}_R(U \oplus V, W) \cong \text{Hom}_R(U, W) \prod \text{Hom}_R(V, W) \cong \text{Hom}_R(U, W) \oplus \text{Hom}_R(V, W),$$

so

$$P^* \oplus P_1^* \cong F^*,$$

$$\text{Hom}_R(P, Q) \oplus \text{Hom}_R(P_1, Q) \cong \text{Hom}_R(F, Q).$$

By tensoring with $Q$ and using the fact that tensor distributes w.r.t. coproducts, we get

$$(P^* \otimes_R Q) \oplus (P_1^* \otimes Q) \cong (P^* \oplus P_1^*) \otimes_R Q \cong F^* \otimes_R Q.$$  

Now, the proof of Proposition 21.16 goes through because $F$ is free and finitely generated, so

$$\alpha_\otimes: (P^* \otimes_R Q) \oplus (P_1^* \otimes Q) \cong F^* \otimes_R Q \longrightarrow \text{Hom}_R(F, Q) \cong \text{Hom}_R(P, Q) \oplus \text{Hom}_R(P_1, Q)$$

is an isomorphism, and as $\alpha_\otimes$ maps $P^* \otimes_R Q$ to $\text{Hom}_R(P, Q)$, it yields an isomorphism between these two spaces.  

\[ \square \]
The isomorphism \( \alpha : P^* \otimes_R Q \cong \text{Hom}_R(P, Q) \) of Proposition 21.26 is still given by

\[
\alpha(u^* \otimes f)(x) = u^*(x)f, \quad u^* \in P^*, \; f \in Q, \; x \in P.
\]

It is convenient to introduce the **evaluation map** \( \text{Ev}_x : P^* \otimes_R Q \rightarrow Q \) defined for every \( x \in P \) by

\[
\text{Ev}_x(u^* \otimes f) = u^*(x)f, \quad u^* \in P^*, \; f \in Q.
\]

In Section 29.2 we will need to consider a slightly weaker version of the universal mapping property of tensor products. The situation is this: We have a commutative \( R \)-algebra \( S \), where \( R \) is a field (or even a commutative ring), we have two \( R \)-modules \( U \) and \( V \), and moreover, \( U \) is a right \( S \)-module and \( V \) is a left \( S \)-module. In Section 29.2, this corresponds to \( R = \mathbb{R}, \ S = C^\infty(B), \ U = \mathcal{A}(\xi) \) and \( V = \Gamma(\xi) \), where \( \xi \) is a vector bundle. Then, we can form the tensor product \( U \otimes_R V \), and we let \( U \otimes_S V \) be the quotient module \( (U \otimes_R V)/W \), where \( W \) is the submodule of \( U \otimes_R V \) generated by the elements of the form

\[
us \otimes_R v - u \otimes_R sv.
\]

As \( S \) is commutative, we can make \( U \otimes_S V \) into an \( S \)-module by defining the action of \( S \) via

\[
s(u \otimes_S v) = us \otimes_S v.
\]

It is immediately verified that this \( S \)-module is isomorphic to the tensor product of \( U \) and \( V \) as \( S \)-modules, and the following universal mapping property holds:

**Proposition 21.27.** For every \( R \)-bilinear map \( f : U \times V \rightarrow Z \), if \( f \) satisfies the property

\[
f(us, v) = f(u, sv), \quad \text{for all } u \in U, \; v \in V, \; s \in S,
\]

then \( f \) induces a unique \( R \)-linear map \( \hat{f} : U \otimes_S V \rightarrow Z \) such that

\[
f(u, v) = \hat{f}(u \otimes_S v), \quad \text{for all } u \in U, \; v \in V.
\]

Note that the linear map \( \hat{f} : U \otimes_S V \rightarrow Z \) is only \( R \)-linear; it is not \( S \)-linear in general.
Chapter 22

Exterior Tensor Powers and Exterior Algebras

22.1 Exterior Tensor Powers

In this chapter, we consider alternating (also called skew-symmetric) multilinear maps and exterior tensor powers (also called alternating tensor powers), denoted $\wedge^n(E)$. In many respects alternating multilinear maps and exterior tensor powers can be treated much like symmetric tensor powers, except that $\text{sgn}(\sigma)$ needs to be inserted in front of the formulae valid for symmetric powers.

Roughly speaking, we are now in the world of determinants rather than in the world of permanents. However, there are also some fundamental differences, one of which being that the exterior tensor power $\wedge^n(E)$ is the trivial vector space (0) when $E$ is finite-dimensional and when $n > \dim(E)$. As in the case of symmetric tensor powers, since we already have the tensor algebra $T(V)$, we can proceed rather quickly. But first let us review some basic definitions and facts.

**Definition 22.1.** Let $f : E^n \rightarrow F$ be a multilinear map. We say that $f$ alternating iff for all $u_i \in E$, $f(u_1, \ldots, u_n) = 0$ whenever $u_i = u_{i+1}$, for some $i$ with $1 \leq i \leq n - 1$; that is, $f(u_1, \ldots, u_n) = 0$ whenever two adjacent arguments are identical. We say that $f$ is skew-symmetric (or anti-symmetric) iff

$$f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = \text{sgn}(\sigma)f(u_1, \ldots, u_n),$$

for every permutation $\sigma \in \mathfrak{S}_n$, and all $u_i \in E$.

For $n = 1$, we agree that every linear map $f : E \rightarrow F$ is alternating. The vector space of all multilinear alternating maps $f : E^n \rightarrow F$ is denoted $\text{Alt}^n(E; F)$. Note that $\text{Alt}^1(E; F) = \text{Hom}(E, F)$. The following basic proposition shows the relationship between alternation and skew-symmetry.
Proposition 22.1. Let \( f : E^n \to F \) be a multilinear map. If \( f \) is alternating, then the following properties hold:

1. For all \( i \), with \( 1 \leq i \leq n - 1 \),
   \[ f(\ldots, u_i, u_{i+1}, \ldots) = -f(\ldots, u_{i+1}, u_i, \ldots). \]

2. For every permutation \( \sigma \in S_n \),
   \[ f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = \text{sgn}(\sigma) f(u_1, \ldots, u_n). \]

3. For all \( i, j \), with \( 1 \leq i < j \leq n \),
   \[ f(\ldots, u_i, \ldots u_j, \ldots) = 0 \quad \text{whenever } u_i = u_j. \]

Moreover, if our field \( K \) has characteristic different from 2, then every skew-symmetric multilinear map is alternating.

Proof. (i) By multilinearity applied twice, we have
\[
f(\ldots, u_i + u_{i+1}, u_i + u_{i+1}, \ldots) = f(\ldots, u_i, u_i, \ldots) + f(\ldots, u_i, u_{i+1}, \ldots) + f(\ldots, u_{i+1}, u_i, \ldots) + f(\ldots, u_{i+1}, u_{i+1}, \ldots).
\]
Since \( f \) is alternating, we get
\[
0 = f(\ldots, u_i, u_{i+1}, \ldots) + f(\ldots, u_{i+1}, u_i, \ldots);
\]
that is, \( f(\ldots, u_i, u_{i+1}, \ldots) = -f(\ldots, u_{i+1}, u_i, \ldots). \)

(ii) Clearly, the symmetric group, \( S_n \), acts on \( \text{Alt}^n(E; F) \) on the left, via
\[
\sigma \cdot f(u_1, \ldots, u_n) = f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}).
\]
Consequently, as \( S_n \) is generated by the transpositions (permutations that swap exactly two elements), since for a transposition, (ii) is simply (i), we deduce (ii) by induction on the number of transpositions in \( \sigma \).

(iii) There is a permutation \( \sigma \) that sends \( u_i \) and \( u_j \) respectively to \( u_1 \) and \( u_2 \). By hypothesis \( u_i = u_j \), so we have \( u_{\sigma(1)} = u_{\sigma(2)} \), and as \( f \) is alternating we have
\[
f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = 0.
\]
However, by (ii),
\[
f(u_1, \ldots, u_n) = \text{sgn}(\sigma) f(u_{\sigma(1)}, \ldots, u_{\sigma(n)}) = 0.
\]
Now, when \( f \) is skew-symmetric, if \( \sigma \) is the transposition swapping \( u_i \) and \( u_{i+1} = u_i \), as \( \text{sgn}(\sigma) = -1 \), we get
\[
f(\ldots, u_i, u_i, \ldots) = -f(\ldots, u_i, u_i, \ldots),
\]
so that
\[
2f(\ldots, u_i, u_i, \ldots) = 0,
\]
and in every characteristic except 2, we conclude that \( f(\ldots, u_i, u_i, \ldots) = 0 \), namely \( f \) is alternating. \( \Box \)
Proposition 22.1 shows that in every characteristic except 2, alternating and skew-symmetric multilinear maps are identical. Using Proposition 22.1 we easily deduce the following crucial fact:

**Proposition 22.2.** Let \( f : E^n \to F \) be an alternating multilinear map. For any families of vectors, \((u_1, \ldots, u_n)\) and \((v_1, \ldots, v_n)\), with \(u_i, v_i \in E\), if

\[
v_j = \sum_{i=1}^{n} a_{ij} u_i, \quad 1 \leq j \leq n,
\]

then

\[
f(v_1, \ldots, v_n) = \left( \sum_{\sigma \in S_n} \text{sgn}(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n} \right) f(u_1, \ldots, u_n) = \det(A) f(u_1, \ldots, u_n),
\]

where \( A \) is the \( n \times n \) matrix, \( A = (a_{ij}) \).

**Proof.** Use property (ii) of Proposition 22.1. \( \square \)

We are now ready to define and construct exterior tensor powers.

**Definition 22.2.** An \( n \)-th exterior tensor power of a vector space \( E \), where \( n \geq 1 \), is a vector space \( A \) together with an alternating multilinear map \( \varphi : E^n \to A \), such that for every vector space \( F \) and for every alternating multilinear map \( f : E^n \to F \), there is a unique linear map \( f_\wedge : A \to F \) with

\[
f(u_1, \ldots, u_n) = f_\wedge(\varphi(u_1, \ldots, u_n)),
\]

for all \( u_1, \ldots, u_n \in E \), or for short

\[
f = f_\wedge \circ \varphi.
\]

Equivalently, there is a unique linear map \( f_\wedge \) such that the following diagram commutes:

\[
\begin{array}{ccc}
E^n & \xrightarrow{\varphi} & A \\
\downarrow f & & \downarrow f_\wedge \\
F & & F
\end{array}
\]

The above property is called the universal mapping property of the exterior tensor power \((A, \varphi)\).

We now show that any two \( n \)-th exterior tensor powers \((A_1, \varphi_1)\) and \((A_2, \varphi_2)\) for \( E \) are isomorphic.

**Proposition 22.3.** Given any two \( n \)-th exterior tensor powers \((A_1, \varphi_1)\) and \((A_2, \varphi_2)\) for \( E \), there is an isomorphism \( h : A_1 \to A_2 \) such that

\[
\varphi_2 = h \circ \varphi_1.
\]
Proof. Replace tensor product by $n$ exterior tensor power in the proof of Proposition 21.5. □

We next give a construction that produces an $n$-th exterior tensor power of a vector space $E$.

**Theorem 22.4.** Given a vector space $E$, an $n$-th exterior tensor power $(\Lambda^n(E), \varphi)$ for $E$ can be constructed $(n \geq 1)$. Furthermore, denoting $\varphi(u_1, \ldots, u_n)$ as $u_1 \wedge \cdots \wedge u_n$, the exterior tensor power $\Lambda^n(E)$ is generated by the vectors $u_1 \wedge \cdots \wedge u_n$, where $u_1, \ldots, u_n \in E$, and for every alternating multilinear map $f : E^n \to F$, the unique linear map $f_\wedge : \Lambda^n(E) \to F$ such that $f = f_\wedge \circ \varphi$ is defined by

$$f_\wedge(u_1 \wedge \cdots \wedge u_n) = f(u_1, \ldots, u_n)$$

on the generators $u_1 \wedge \cdots \wedge u_n$ of $\Lambda^n(E)$.

**Proof sketch.** We can give a quick proof using the tensor algebra $T(E)$. Let $\mathcal{I}_a$ be the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes u \in E \otimes E$. Then, let

$$\Lambda^n(E) = E^{\otimes n}/(\mathcal{I}_a \cap E^{\otimes n})$$

and let $\pi$ be the projection $\pi : E^{\otimes n} \to \Lambda^n(E)$. If we let $u_1 \wedge \cdots \wedge u_n = \pi(u_1 \otimes \cdots \otimes u_n)$, it is easy to check that $(\Lambda^n(E), \wedge)$ satisfies the conditions of Theorem 22.4. □

**Remark:** We can also define

$$\Lambda(E) = T(E)/\mathcal{I}_a = \bigoplus_{n \geq 0} \Lambda^n(E),$$

the *exterior algebra* of $E$. This is the skew-symmetric counterpart of $S(E)$, and we will study it a little later.

For simplicity of notation, we may write $\Lambda^n E$ for $\Lambda^n(E)$. We also abbreviate “exterior tensor power” as “exterior power.” Clearly, $\Lambda^1(E) \cong E$, and it is convenient to set $\Lambda^0(E) = K$.

The fact that the map $\varphi : E^n \to \Lambda^n(E)$ is alternating and multilinear can also be expressed as follows:

$$u_1 \wedge \cdots \wedge (u_i + v_i) \wedge \cdots \wedge u_n = (u_1 \wedge \cdots \wedge u_i \wedge \cdots \wedge u_n) + (u_1 \wedge \cdots \wedge v_i \wedge \cdots \wedge u_n),$$

$$u_1 \wedge \cdots \wedge (\lambda u_i) \wedge \cdots \wedge u_n = \lambda (u_1 \wedge \cdots \wedge u_i \wedge \cdots \wedge u_n),$$

$$u_{\sigma(1)} \wedge \cdots \wedge u_{\sigma(n)} = \text{sgn}(\sigma) u_1 \wedge \cdots \wedge u_n,$$

for all $\sigma \in S_n$. 

The map $\varphi$ from $E^n$ to $\bigwedge^n(E)$ is often denoted $\iota_n$, so that
$$\iota_n(u_1, \ldots, u_n) = u_1 \wedge \cdots \wedge u_n.$$ 

Theorem 22.4 implies the following result.

**Proposition 22.5.** There is a canonical isomorphism
$$\text{Hom}(\bigwedge^n(E), F) \cong \text{Alt}^n(E; F)$$
between the vector space of linear maps $\text{Hom}(\bigwedge^n(E), F)$ and the vector space of alternating multilinear maps $\text{Alt}^n(E; F)$, given by the linear map $- \circ \varphi$ defined by $h \circ \varphi$, with $h \in \text{Hom}(\bigwedge^n(E), F)$. In particular, when $F = K$, we get a canonical isomorphism
$$\left(\bigwedge^n(E)^*\right)^* \cong \text{Alt}^n(E; K).$$

Tensors $\alpha \in \bigwedge^n(E)$ are called alternating $n$-tensors or alternating tensors of degree $n$ and we write $\deg(\alpha) = n$. Tensors of the form $u_1 \wedge \cdots \wedge u_n$, where $u_i \in E$, are called simple (or decomposable) alternating $n$-tensors. Those alternating $n$-tensors that are not simple are often called compound alternating $n$-tensors. Simple tensors $u_1 \wedge \cdots \wedge u_n \in \bigwedge^n(E)$ are also called $n$-vectors and tensors in $\bigwedge^n(E^*)$ are often called (alternating) $n$-forms.

Given two linear maps $f : E \to E'$ and $g : E \to E'$, since the map $\iota_n' \circ (f \times g)$ is bilinear and alternating, there is a unique linear map $f \wedge g : \bigwedge^2(E) \to \bigwedge^2(E')$ making the following diagram commute:

$$\begin{array}{ccc}
E^2 & \xrightarrow{\iota_n} & \bigwedge^2(E) \\
\downarrow{f \times g} & & \downarrow{f \wedge g} \\
(E')^2 & \xrightarrow{\iota_n'} & \bigwedge^2(E').
\end{array}$$

The map $f \wedge g : \bigwedge^2(E) \to \bigwedge^2(E')$ is determined by
$$(f \wedge g)(u \wedge v) = f(u) \wedge g(u).$$

If we also have linear maps $f' : E' \to E''$ and $g' : E' \to E''$, we can easily verify that
$$(f' \circ f) \wedge (g' \circ g) = (f' \wedge g') \circ (f \wedge g).$$

The generalization to the alternating product $f_1 \wedge \cdots \wedge f_n$ of $n \geq 3$ linear maps $f_i : E \to E'$ is immediate, and left to the reader.
22.2 Bases of Exterior Powers

Let $E$ be any vector space. For any basis $(u_i)_{i \in \Sigma}$ for $E$, we assume that some total ordering $\leq$ on the index set $\Sigma$ has been chosen. Call the pair $((u_i)_{i \in \Sigma}, \leq)$ an ordered basis. Then, for any nonempty finite subset $I \subseteq \Sigma$, let

$$u_I = u_{i_1} \wedge \cdots \wedge u_{i_m},$$

where $I = \{i_1, \ldots, i_m\}$, with $i_1 < \cdots < i_m$.

Since $\bigwedge^n(E)$ is generated by the tensors of the form $v_1 \wedge \cdots \wedge v_n$, with $v_i \in E$, in view of skew-symmetry, it is clear that the tensors $u_I$ with $|I| = n$ generate $\bigwedge^n(E)$. Actually they form a basis. To gain an intuitive understanding of this statement, let $m = 2$ and $E$ be a vector space of dimension 3 with lexicographically ordered basis $\{e_1, e_2, e_3\}$. We claim that

$$e_1 \wedge e_2, \quad e_1 \wedge e_3, \quad e_2 \wedge e_3$$

form a basis for $\bigwedge^2(E)$ since they not only generate $\bigwedge^2(E)$ but are linearly independent. The linear independence is argued as follows: given any vector space $F$, if $w_{12}, w_{13}, w_{23}$ are any vectors in $F$, then there is an alternating bilinear map $h: E^2 \rightarrow F$ such that

$$h(e_1, e_2) = w_{12}, \quad h(e_1, e_3) = w_{13}, \quad h(e_2, e_3) = w_{23}.$$

Because $h$ yields a unique linear map $h_\wedge: \bigwedge^2 E \rightarrow F$ such that

$$h_\wedge(e_i \wedge e_j) = w_{ij}, \quad 1 \leq i < j \leq 3,$$

by Proposition 21.4, the vectors

$$e_1 \wedge e_2, \quad e_1 \wedge e_3, \quad e_2 \wedge e_3$$

are linearly independent. This suggests understanding how an alternating bilinear function $f: E^2 \rightarrow F$ is expressed in terms of its values $f(e_i, e_j)$ on the basis vectors $(e_1, e_2, e_3)$. Using bilinearity and alternation, we obtain

$$f(u_1 e_1 + u_2 e_2 + u_3 e_3, v_1 e_1 + v_2 e_2 + v_3 e_3) = (u_1 v_2 - u_2 v_1) f(e_1, e_2) + (u_1 v_3 - u_3 v_1) f(e_1, e_3) + (u_2 v_3 - u_3 v_2) f(e_2, e_3).$$

Therefore, given $w_{12}, w_{13}, w_{23} \in F$, the function $h$ given by

$$h(u_1 e_1 + u_2 e_2 + u_3 e_3, v_1 e_1 + v_2 e_2 + v_3 e_3) = (u_1 v_2 - u_2 v_1) w_{12} + (u_1 v_3 - u_3 v_1) w_{13} + (u_2 v_3 - u_3 v_2) w_{23}$$

is clearly bilinear and alternating, and by construction $h(e_i, e_j) = w_{ij}$, with $1 \leq i < j \leq 3$ does the job.

We now prove the assertion that tensors $u_I$ with $|I| = n$ generate $\bigwedge^n(E)$ for arbitrary $n$. 
Proposition 22.6. Given any vector space $E$, if $E$ has finite dimension $d = \dim(E)$, then for all $n > d$, the exterior power $\wedge^n(E)$ is trivial; that is $\wedge^n(E) = (0)$. If $n \leq d$ or if $E$ is infinite dimensional, then for every ordered basis $((u_i)_{i \in \Sigma}, \leq)$, the family $(u_I)$ is basis of $\wedge^n(E)$, where $I$ ranges over finite nonempty subsets of $\Sigma$ of size $|I| = n$.

Proof. First assume that $E$ has finite dimension $d = \dim(E)$ and that $n > d$. We know that $\wedge^n(E)$ is generated by the tensors of the form $v_1 \wedge \cdots \wedge v_n$, with $v_i \in E$. If $u_1, \ldots, u_d$ is a basis of $E$, as every $v_i$ is a linear combination of the $u_j$, when we expand $v_1 \wedge \cdots \wedge v_n$ using multilinearity, we get a linear combination of the form

$$v_1 \wedge \cdots \wedge v_n = \sum_{(j_1, \ldots, j_n)} \lambda_{(j_1, \ldots, j_n)} u_{j_1} \wedge \cdots \wedge u_{j_n},$$

where each $(j_1, \ldots, j_n)$ is some sequence of integers $j_k \in \{1, \ldots, d\}$. As $n > d$, each sequence $(j_1, \ldots, j_n)$ must contain two identical elements. By alternation, $u_{j_1} \wedge \cdots \wedge u_{j_n} = 0$, and so $v_1 \wedge \cdots \wedge v_n = 0$. It follows that $\wedge^n(E) = (0)$.

Now, assume that either $\dim(E) = d$ and $n \leq d$, or that $E$ is infinite dimensional. The argument below shows that the $u_I$ are nonzero and linearly independent. As usual, let $u_i^* \in E^*$ be the linear form given by

$$u_i^*(u_j) = \delta_{ij}.$$

For any nonempty subset $I = \{i_1, \ldots, i_n\} \subseteq \Sigma$ with $i_1 < \cdots < i_n$, for any $n$ vectors $v_1, \ldots, v_n \in E$, let

$$l_I(v_1, \ldots, v_n) = \det(u_{i_j}^*(v_k)) = \begin{vmatrix} u_{i_1}^*(v_1) & \cdots & u_{i_1}^*(v_n) \\ \vdots & \ddots & \vdots \\ u_{i_n}^*(v_1) & \cdots & u_{i_n}^*(v_n) \end{vmatrix}.$$

If we let the $n$-tuple $(v_1, \ldots, v_n)$ vary we obtain a map $l_I$ from $E^n$ to $K$, and it is easy to check that this map is alternating multilinear. Thus $l_I$ induces a unique linear map $L_I: \wedge^n(E) \rightarrow K$ making the following diagram commute:

$$E^n \xrightarrow{\wedge} \wedge^n(E) \xrightarrow{l_I} K.$$

Observe that for any nonempty finite subset $J \subseteq \Sigma$ with $|J| = n$, we have

$$L_I(u_J) = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{if } I \neq J. \end{cases}$$
Note that when \( \dim(E) = d \) and \( n \leq d \), or when \( E \) is infinite-dimensional, the forms \( u^*_i, \ldots, u^*_n \) are all distinct, so the above does hold. Since \( L_I(u_I) = 1 \), we conclude that \( u_I \neq 0 \). Now, if we have a linear combination
\[
\sum_I \lambda_I u_I = 0,
\]
where the above sum is finite and involves nonempty finite subset \( I \subseteq \Sigma \) with \( |I| = n \), for every such \( I \), when we apply \( L_I \) we get \( \lambda_I = 0 \), proving linear independence. \( \square \)

As a corollary, if \( E \) is finite dimensional, say \( \dim(E) = d \), and if \( 1 \leq n \leq d \), then we have
\[
\dim(\bigwedge^n(E)) = \binom{n}{d},
\]
and if \( n > d \), then \( \dim(\bigwedge^n(E)) = 0 \).

**Remark:** When \( n = 0 \), if we set \( u_\emptyset = 1 \), then \( (u_\emptyset) = (1) \) is a basis of \( \bigwedge^0(V) = K \).

It follows from Proposition 22.6 that the family \( (u_I) \) where \( I \subseteq \Sigma \) ranges over finite subsets of \( \Sigma \) is a basis of \( \bigwedge(V) = \bigoplus_{n \geq 0} \bigwedge^n(V) \).

As a corollary of Proposition 22.6 we obtain the following useful criterion for linear independence:

**Proposition 22.7.** For any vector space \( E \), the vectors \( u_1, \ldots, u_n \in E \) are linearly independent iff \( u_1 \wedge \cdots \wedge u_n \neq 0 \).

**Proof.** If \( u_1 \wedge \cdots \wedge u_n \neq 0 \), then \( u_1, \ldots, u_n \) must be linearly independent. Otherwise, some \( u_i \) would be a linear combination of the other \( u_j \)'s (with \( j \neq i \)), and then, as in the proof of Proposition 22.6, \( u_1 \wedge \cdots \wedge u_n \) would be a linear combination of wedges in which two vectors are identical, and thus zero.

Conversely, assume that \( u_1, \ldots, u_n \) are linearly independent. Then we have the linear forms \( u^*_i \in E^* \) such that
\[
u^*_i(u_j) = \delta_{i,j} \quad 1 \leq i, j \leq n.
\]
As in the proof of Proposition 22.6, we have a linear map \( L_{u_1, \ldots, u_n} : \bigwedge^n(E) \to K \) given by
\[
L_{u_1, \ldots, u_n}(v_1 \wedge \cdots \wedge v_n) = \det(u^*_j(v_i)) = \begin{vmatrix} u^*_1(v_1) & \cdots & u^*_1(v_n) \\ \vdots & \ddots & \vdots \\ u^*_n(v_1) & \cdots & u^*_n(v_n) \end{vmatrix},
\]
for all \( v_1 \wedge \cdots \wedge v_n \in \bigwedge^n(E) \). As \( L_{u_1, \ldots, u_n}(u_1 \wedge \cdots \wedge u_n) = 1 \), we conclude that \( u_1 \wedge \cdots \wedge u_n \neq 0 \). \( \square \)

Proposition 22.7 shows that geometrically every nonzero wedge \( u_1 \wedge \cdots \wedge u_n \) corresponds to some oriented version of an \( n \)-dimensional subspace of \( E \).
22.3 Some Useful Isomorphisms for Exterior Powers

We can show the following property of the exterior tensor product, using the proof technique of Proposition 21.12:

\[ \bigwedge^n (E \oplus F) \cong \bigoplus_{k=0}^{n} \bigwedge^k (E) \otimes \bigwedge^{n-k} (F). \]

22.4 Duality for Exterior Powers

In this section all vector spaces are assumed to have finite dimension. We define a nondegenerate pairing \( \bigwedge^n (E^*) \times \bigwedge^n (E) \to K \) as follows: Consider the multilinear map

\[ (E^*)^n \times E^n \to K \]

given by

\[
(v_1^*, \ldots, v_n^*, u_1, \ldots, u_n) \mapsto \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) v_{\sigma(1)}^*(u_1) \cdots v_{\sigma(n)}^*(u_n) = \det(v_j^*(u_i))
\]

\[
= \begin{vmatrix}
 v_1^*(u_1) & \cdots & v_n^*(u_n) \\
 \vdots & \ddots & \vdots \\
 v_n^*(u_1) & \cdots & v_n^*(u_n)
\end{vmatrix}.
\]

It is easily checked that this expression is alternating w.r.t. the \( u_i \)'s and also w.r.t. the \( v_j^* \). For any fixed \((v_1^*, \ldots, v_n^*) \in (E^*)^n\), we get an alternating multilinear map

\[ l_{v_1^*, \ldots, v_n^*} : (u_1, \ldots, u_n) \mapsto \det(v_j^*(u_i)) \]

from \( E^n \) to \( K \). By the argument used in the symmetric case, we get a bilinear map

\[ \bigwedge^n (E^*) \times \bigwedge^n (E) \to K. \]

Now, this pairing in nondegenerate. This can be shown using bases and we leave it as an exercise to the reader. As a consequence we get a following canonical isomorphisms.

**Proposition 22.8.** There is a canonical isomorphism

\[ (\bigwedge^n (E))^* \cong \bigwedge^n (E^*). \]

There is also a canonical isomorphism

\[ \mu : \bigwedge^n (E^*) \cong \operatorname{Alt}^n (E; K) \]

which allows us to interpret alternating tensors over \( E^* \) as alternating multilinear maps.
Proof. The second isomorphism follows from the canonical isomorphism \((\wedge^n(E))^* \cong \wedge^n(E^*)\) and the canonical isomorphism \(\wedge^n(E))^* \cong \text{Alt}^n(E; K)\) given by Proposition 22.5. 

The isomorphism \(\mu: \wedge^n(E^*) \cong \text{Alt}^n(E; K)\) discussed above can be described explicitly as the linear extension of the map given by

\[\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(v_j^*(u_i)).\]

Remark: Variants of our isomorphism \(\mu\) are found in the literature. For example, there is a version \(\mu'\), where \(\mu' = \frac{1}{n!}\mu\), with the factor \(\frac{1}{n!}\) added in front of the determinant. Each version has its own merits and inconveniences. Morita [133] uses \(\mu'\) because it is more convenient than \(\mu\) when dealing with characteristic classes. On the other hand, when using \(\mu'\), some extra factor is needed in defining the wedge operation of alternating multilinear forms (see Section 22.5) and for exterior differentiation. The version \(\mu\) is the one adopted by Warner [175], Knapp [106], Fulton and Harris [70], and Cartan [36, 37].

If \(f: E \to F\) is any linear map, by transposition we get a linear map \(f^T: F^* \to E^*\) given by

\[f^T(v^*)(u) = v^*(f(u)), \quad v^* \in F^*.
\]

Consequently, we have

\[f^T(v^*)(u) = v^*(f(u)), \quad \text{for all } u \in E \text{ and all } v^* \in F^*.
\]

For any \(p \geq 1\), the map

\[(u_1, \ldots, u_p) \mapsto f(u_1) \wedge \cdots \wedge f(u_p)\]

from \(E^p\) to \(\wedge^p F\) is multilinear alternating, so it induces a unique linear map \(\wedge^p f: \wedge^p E \to \wedge^p F\) making the following diagram commute

\[
\begin{array}{ccc}
E^p & \xrightarrow{i^\wedge} & \wedge^p E \\
\downarrow & & \downarrow \wedge^p f \\
& \wedge^p F, & \\
\end{array}
\]

and defined on generators by

\[\left(\wedge^p f\right)(u_1 \wedge \cdots \wedge u_p) = f(u_1) \wedge \cdots \wedge f(u_p).
\]

Combining \(\wedge^p\) and duality, we get a linear map \(\wedge^p f^T: \wedge^p F^* \to \wedge^p E^*\) defined on generators by

\[\left(\wedge^p f^T\right)(v_1^* \wedge \cdots \wedge v_p^*) = f^T(v_1^*) \wedge \cdots \wedge f^T(v_p^*).\]
Proposition 22.9. If \( f: E \to F \) is any linear map between two finite-dimensional vector spaces \( E \) and \( F \), then

\[
\mu\left( \left( \bigwedge^p f^\top \right)(\omega) \right)(u_1, \ldots, u_p) = \mu(\omega)(f(u_1), \ldots, f(u_p)), \quad \omega \in \bigwedge^p F^*, \ u_1, \ldots, u_p \in E.
\]

Proof. It is enough to prove the formula on generators. By definition of \( \mu \), we have

\[
\mu\left( \left( \bigwedge^p f^\top \right) (v_1^* \wedge \cdots \wedge v_p^*) \right)(u_1, \ldots, u_p) = \mu(f^\top(v_1^*) \wedge \cdots \wedge f^\top(v_p^*))(u_1, \ldots, u_p)
= \det(f^\top(v_i^*)(u_i))
= \det(v_j^*(f(u_i)))
= \mu(v_1^* \wedge \cdots \wedge v_p^*)(f(u_1), \ldots, f(u_p)),
\]

as claimed. \( \square \)

The map \( \bigwedge^p f^\top \) is often denoted \( f^* \), although this is an ambiguous notation since \( p \) is dropped. Proposition 22.9 gives us the behavior of \( f^* \) under the identification of \( \bigwedge^p E^* \) and \( \text{Alt}^p(E; K) \) via the isomorphism \( \mu \).

As in the case of symmetric powers, the map from \( E^n \) to \( \bigwedge^n (E) \) given by \( (u_1, \ldots, u_n) \mapsto u_1 \wedge \cdots \wedge u_n \) yields a surjection \( \pi: E^\otimes n \to \bigwedge^n (E) \). Now, this map has some section, so there is some injection \( \eta: \bigwedge^n (E) \to E^\otimes n \) with \( \pi \circ \eta = \text{id} \). If our field \( K \) has characteristic 0, then there is a special section having a natural definition involving an antisymmetrization process.

Recall that we have a left action of the symmetric group \( S_n \) on \( E^\otimes n \). The tensors \( z \in E^\otimes n \) such that

\[
\sigma \cdot z = \text{sgn}(\sigma) z, \quad \text{for all } \sigma \in S_n
\]

are called antisymmetrized tensors. We define the map \( \eta: E^n \to E^\otimes n \) by

\[
\eta(u_1, \ldots, u_n) = \frac{1}{n!} \sum_{\sigma \in S_n} \text{sgn}(\sigma) u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(n)}.
\]

As the right hand side is an alternating map, we get a unique linear map \( \bigwedge^n \eta: \bigwedge^n (E) \to E^\otimes n \) making the following diagram commute:

\[
\begin{array}{ccc}
E^n & \xrightarrow{\iota^\wedge} & \bigwedge^n (E) \\
\downarrow{\eta} & & \downarrow{\bigwedge^n \eta} \\
E^\otimes n & & \\
\end{array}
\]

Clearly, \( \bigwedge^n \eta(\bigwedge^n (E)) \) is the set of antisymmetrized tensors in \( E^\otimes n \). If we consider the map \( A = (\bigwedge^n \eta) \circ \pi: E^\otimes n \to E^\otimes n \), it is easy to check that \( A \circ A = A \). Therefore, \( A \) is a projection, and by linear algebra, we know that

\[
E^\otimes n = A(E^\otimes n) \oplus \text{Ker} A = \bigwedge^n \eta(\bigwedge^n (A)) \oplus \text{Ker} A.
\]
It turns out that $\text{Ker } A = E \otimes n \cap I_a = \text{Ker } \pi$, where $I_a$ is the two-sided ideal of $T(E)$ generated by all tensors of the form $u \otimes u \in E \otimes^2$ (for example, see Knapp [106], Appendix A). Therefore, $\wedge^n \eta$ is injective,

$$E \otimes^n = \bigwedge^n \eta(\bigwedge (E)) \oplus (E \otimes^n \cap I_a) = \bigwedge^n \eta(\bigwedge (E)) \oplus \text{Ker } \pi,$$

and the exterior tensor power $\wedge^n (E)$ is naturally embedded into $E \otimes^n$.

### 22.5 Exterior Algebras

As in the case of symmetric tensors, we can pack together all the exterior powers $\wedge^n (V)$ into an algebra

$$\bigwedge (V) = \bigoplus_{m \geq 0} \wedge^m (V),$$

called the **exterior algebra (or Grassmann algebra)** of $V$. We mimic the procedure used for symmetric powers. If $I_a$ is the two-sided ideal generated by all tensors of the form $u \otimes u \in V \otimes^2$, we set

$$\bigwedge (V) = T(V) / I_a.$$

Then, $\bigwedge (V)$ automatically inherits a multiplication operation, called **wedge product**, and since $T(V)$ is graded, that is

$$T(V) = \bigoplus_{m \geq 0} V \otimes^m,$$

we have

$$\bigwedge (V) = \bigoplus_{m \geq 0} V \otimes^m / (I_a \cap V \otimes^m).$$

However, it is easy to check that

$$\bigwedge^m (V) \cong V \otimes^m / (I_a \cap V \otimes^m),$$

so

$$\bigwedge (V) \cong \bigwedge (V).$$

When $V$ has finite dimension $d$, we actually have a finite direct sum (coproduct)

$$\bigwedge (V) = \bigoplus_{m=0}^d \bigwedge^m (V),$$
and since each $\bigwedge^m(V)$ has dimension $\binom{d}{m}$, we deduce that

$$\dim(\bigwedge(V)) = 2^d = 2^{\dim(V)}.$$  

The multiplication, $\wedge: \bigwedge^m(V) \times \bigwedge^n(V) \to \bigwedge^{m+n}(V)$, is skew-symmetric in the following precise sense:

**Proposition 22.10.** For all $\alpha \in \bigwedge^m(V)$ and all $\beta \in \bigwedge^n(V)$, we have

$$\beta \wedge \alpha = (-1)^{mn} \alpha \wedge \beta.$$  

**Proof.** Since $v \wedge u = -u \wedge v$ for all $u, v \in V$, Proposition 22.10 follows by induction. \qed

Since $\alpha \wedge \alpha = 0$ for every simple (also called decomposable) tensor $\alpha = u_1 \wedge \cdots \wedge u_n$, it seems natural to infer that $\alpha \wedge \alpha = 0$ for every tensor $\alpha \in \bigwedge(V)$. If we consider the case where $\dim(V) \leq 3$, we can indeed prove the above assertion. However, if $\dim(V) \geq 4$, the above fact is generally false! For example, when $\dim(V) = 4$, if $u_1, u_2, u_3, u_4$ are a basis for $V$, for $\alpha = u_1 \wedge u_2 + u_3 \wedge u_4$, we check that

$$\alpha \wedge \alpha = 2u_1 \wedge u_2 \wedge u_3 \wedge u_4,$$

which is nonzero. However, if $\alpha \in \bigwedge^m E$ with $m$ odd, since $m^2$ is also odd, we have

$$\alpha \wedge \alpha = (-1)^{m^2} \alpha \wedge \alpha = -\alpha \wedge \alpha,$$

so indeed $\alpha \wedge \alpha = 0$ (if $K$ is not a field of characteristic 2).

The above discussion suggests that it might be useful to know when an alternating tensor is simple (decomposable). We will show in Section 22.7 that for tensors $\alpha \in \bigwedge^2(V)$, $\alpha \wedge \alpha = 0$ iff $\alpha$ is simple.

A general criterion for decomposability can be given in terms of some operations known as left hook and right hook (also called interior products); see Section 22.7.

It is easy to see that $\bigwedge(V)$ satisfies the following universal mapping property:

**Proposition 22.11.** Given any $K$-algebra $A$, for any linear map $f: V \to A$, if $(f(v))^2 = 0$ for all $v \in V$, then there is a unique $K$-algebra homomorphism $\overline{f}: \bigwedge(V) \to A$ so that

$$f = \overline{f} \circ i,$$

as in the diagram below:

$$\begin{array}{ccc}
V & \xrightarrow{i} & \bigwedge(V) \\
\downarrow{f} & & \downarrow{\overline{f}} \\
A & \end{array}$$
When $E$ is finite-dimensional, recall the isomorphism $\mu: \bigwedge^n(E^*) \rightarrow \text{Alt}^n(E; K)$, defined as the linear extension of the map given by

$$\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(v_j^*(u_i)).$$

Now, we have also a multiplication operation $\bigwedge^m(E^*) \times \bigwedge^n(E^*) \rightarrow \bigwedge^{m+n}(E^*)$. The following question then arises:

Can we define a multiplication $\text{Alt}^m(E; K) \times \text{Alt}^n(E; K) \rightarrow \text{Alt}^{m+n}(E; K)$ directly on alternating multilinear forms, so that the following diagram commutes:

$$
\begin{array}{ccc}
\bigwedge^m(E^*) \times \bigwedge^n(E^*) & \overset{\wedge}{\longrightarrow} & \bigwedge^{m+n}(E^*) \\
\mu_m \times \mu_n & & \mu_{m+n} \\
\text{Alt}^m(E; K) \times \text{Alt}^n(E; K) & \overset{\wedge}{\longrightarrow} & \text{Alt}^{m+n}(E; K).
\end{array}
$$

As in the symmetric case, the answer is yes! The solution is to define this multiplication such that, for $f \in \text{Alt}^m(E; K)$ and $g \in \text{Alt}^n(E; K)$,

$$(f \wedge g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \text{shuffle}(m,n)} \text{sgn}(\sigma) f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}) g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}), \quad (**)$$

where shuffle$(m,n)$ consists of all $(m,n)$-“shuffles;” that is, permutations $\sigma$ of $\{1, \ldots, m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$. For example, when $m = n = 1$, we have

$$(f \wedge g)(u,v) = f(u)g(v) - g(u)f(v).$$

When $m = 1$ and $n \geq 2$, check that

$$(f \wedge g)(u_1, \ldots, u_{m+1}) = \sum_{i=1}^{m+1} (-1)^{i-1} f(u_i)g(u_1, \ldots, \hat{u}_i, \ldots, u_{m+1}),$$

where the hat over the argument $u_i$ means that it should be omitted.

Here is another explicit example. Suppose $m = 2$ and $n = 1$. Given $v_1^*, v_2^*, v_3^* \in E^*$, the multiplication structure on $\bigwedge(E^*)$ implies that $(v_1^* \wedge v_2^*) \cdot v_3^* = v_1^* \wedge v_2^* \wedge v_3^* \in \bigwedge^3(E^*)$. Furthermore, for $u_1, u_2, u_3, \in E$,

$$\mu_3(v_1^* \wedge v_2^* \wedge v_3^*)(u_1, u_2, u_3) = \sum_{\sigma \in \Theta_3} \text{sgn}(\sigma) v_{\sigma(1)}^*(u_1)v_{\sigma(2)}^*(u_2)v_{\sigma(3)}^*(u_3)$$

$$= v_1^*(u_1)v_2^*(u_2)v_3^*(u_3) - v_1^*(u_1)v_3^*(u_2)v_2^*(u_3) - v_2^*(u_1)v_1^*(u_2)v_3^*(u_3) + v_2^*(u_1)v_3^*(u_2)v_1^*(u_3) + v_3^*(u_1)v_1^*(u_2)v_2^*(u_3) - v_3^*(u_1)v_2^*(u_2)v_1^*(u_3).$$
Now the $(2,1)$-shuffles of $\{1,2,3\}$ are the following three permutations, namely
\[
\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 3 & 2 \\ 2 & 3 & 1 \end{pmatrix}.
\]
If $f \cong \mu_2(v_1^* \wedge v_2^*)$ and $g \cong \mu_1(v_3^*)$, then (**) implies that
\[
(f \cdot g)(u_1, u_2, u_3) = \sum_{\sigma \in \text{shuffle}(2,1)} \text{sgn}(\sigma) f(u_{\sigma(1)}, u_{\sigma(2)}) g(u_{\sigma(3)})
\]
\[
= f(u_1, u_2) g(u_3) - f(u_1, u_3) g(u_2) + f(u_2, u_3) g(u_1)
\]
\[
= \mu_2(v_1^* \wedge v_2^*)(u_1, u_2) \mu_1(v_3^*)(u_3) - \mu_2(v_1^* \wedge v_2^*)(u_1, u_3) \mu_1(v_3^*)(u_2)
\]
\[
+ \mu_2(v_1^* \wedge v_2^*)(u_2, u_3) \mu_1(v_3^*)(u_1)
\]
\[
= (v_1^*(u_1)v_2^*(u_2) - v_1^*(u_1)v_3^*(u_2))v_3^*(u_3)
\]
\[
- (v_1^*(u_1)v_2^*(u_3) - v_2^*(u_1)v_3^*(u_3))v_3^*(u_2)
\]
\[
+ (v_1^*(u_2)v_2^*(u_3) - v_2^*(u_2)v_3^*(u_3))v_3^*(u_1)
\]
\[
= \mu_3(v_1^* \wedge v_2^* \wedge v_3^*)(u_1, u_2, u_3).
\]

As a result of all this, the direct sum
\[
\text{Alt}(E) = \bigoplus_{n \geq 0} \text{Alt}^n(E; K)
\]
is an algebra under the above multiplication, and this algebra is isomorphic to $\wedge(E^*)$. For the record we state

**Proposition 22.12.** When $E$ is finite dimensional, the maps $\mu \colon \wedge^n(E^*) \to \text{Alt}^n(E; K)$ induced by the linear extensions of the maps given by
\[
\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \det(v_i^*(u_i))
\]
yield a canonical isomorphism of algebras $\mu \colon \wedge(E^*) \to \text{Alt}(E)$, where the multiplication in $\text{Alt}(E)$ is defined by the maps $\wedge \colon \text{Alt}^m(E; K) \times \text{Alt}^n(E; K) \to \text{Alt}^{m+n}(E; K)$, with
\[
(f \wedge g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \text{shuffle}(m,n)} \text{sgn}(\sigma) f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}) g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}),
\]
where shuffle$(m,n)$ consists of all $(m,n)$-“shuffles,” that is, permutations $\sigma$ of $\{1, \ldots, m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.

**Remark:** The algebra $\wedge(E)$ is a graded algebra. Given two graded algebras $E$ and $F$, we can make a new tensor product $E \otimes F$, where $E \otimes F$ is equal to $E \otimes F$ as a vector space, but with a skew-commutative multiplication given by
\[
(a \otimes b) \wedge (c \otimes d) = (-1)^{\deg(b)\deg(c)}(ac) \otimes (bd),
\]
where $a \in E^m, b \in F^p, c \in E^n, d \in F^q$. Then, it can be shown that
\[
\wedge(E \oplus F) \cong \wedge(E) \otimes \wedge(F).
\]
22.6 The Hodge $\ast$-Operator

In order to define a generalization of the Laplacian that applies to differential forms on a Riemannian manifold, we need to define isomorphisms

$$\wedge^k V \to \wedge^{n-k} V,$$

for any Euclidean vector space $V$ of dimension $n$ and any $k$, with $0 \leq k \leq n$. If $\langle - , - \rangle$ denotes the inner product on $V$, we define an inner product on $\wedge^k V$, also denoted $\langle - , - \rangle$, by setting

$$\langle u_1 \wedge \cdots \wedge u_k, v_1 \wedge \cdots \wedge v_k \rangle = \det(\langle u_i, v_j \rangle),$$

for all $u_i, v_i \in V$, and extending $\langle - , - \rangle$ by bilinearity.

It is easy to show that if $(e_1, \ldots, e_n)$ is an orthonormal basis of $V$, then the basis of $\wedge^k V$ consisting of the $e_I$ (where $I = \{i_1, \ldots, i_k\}$, with $1 \leq i_1 < \cdots < i_k \leq n$) is an orthonormal basis of $\wedge^k V$. Since the inner product on $V$ induces an inner product on $V^*$ (recall that $\langle \omega_1, \omega_2 \rangle = \langle \omega_1^\flat, \omega_2^\flat \rangle$, for all $\omega_1, \omega_2 \in V^*$), we also get an inner product on $\wedge^k V^*$.

Recall that an orientation of a vector space $V$ of dimension $n$ is given by the choice of some basis $(e_1, \ldots, e_n)$. We say that a basis $(u_1, \ldots, u_n)$ of $V$ is positively oriented iff $\det(u_1, \ldots, u_n) > 0$ (where $\det(u_1, \ldots, u_n)$ denotes the determinant of the matrix whose $j$th column consists of the coordinates of $u_j$ over the basis $(e_1, \ldots, e_n)$), otherwise it is negatively oriented. An oriented vector space is a vector space $V$ together with an orientation of $V$. If $V$ is oriented by the basis $(e_1, \ldots, e_n)$, then $V^*$ is oriented by the dual basis $(e_1^\ast, \ldots, e_n^\ast)$. If $\sigma$ is any permutation of $\{1, \ldots, n\}$, then the basis $(e_{\sigma(1)}, \ldots, e_{\sigma(n)})$ has positive orientation iff the signature $\text{sgn}(\sigma)$ of the permutation $\sigma$ is even.

If $V$ is an oriented vector space of dimension $n$, then we can define a linear isomorphism

$$\ast : \wedge^k V \to \wedge^{n-k} V,$$

called the Hodge $\ast$-operator. The existence of this operator is guaranteed by the following proposition.

**Proposition 22.13.** Let $V$ be any oriented Euclidean vector space whose orientation is given by some chosen orthonormal basis $(e_1, \ldots, e_n)$. For any alternating tensor $\alpha \in \wedge^k V$, there is a unique alternating tensor $\ast \alpha \in \wedge^{n-k} V$ such that

$$\alpha \wedge \beta = (\ast \alpha, \beta) e_1 \wedge \cdots \wedge e_n$$

for all $\beta \in \wedge^{n-k} V$. The alternating tensor $\ast \alpha$ is independent of the choice of the positive orthonormal basis $(e_1, \ldots, e_n)$.
22.6. THE HODGE $*$-OPERATOR

Proof. Since $\bigwedge^n V$ has dimension 1, the alternating tensor $e_1 \wedge \cdots \wedge e_n$ is a basis of $\bigwedge^n V$. It follows that for any fixed $\alpha \in \bigwedge^k V$, the linear map $\lambda_\alpha$ from $\bigwedge^{n-k} V$ to $\bigwedge^n V$ given by

$$\lambda_\alpha(\beta) = \alpha \wedge \beta$$

is of the form

$$\lambda_\alpha(\beta) = f_\alpha(\beta) e_1 \wedge \cdots \wedge e_n$$

for some linear form $f_\alpha \in \left( \bigwedge^{n-k} V \right)^\ast$. But then, by the duality induced by the inner product $\langle -, - \rangle$ on $\bigwedge^{n-k} V$, there is a unique vector $\ast \alpha \in \bigwedge^{n-k} V$ such that

$$f_\lambda(\beta) = \langle \ast \alpha, \beta \rangle$$

for all $\beta \in \bigwedge^{n-k} V$,

which implies that

$$\alpha \wedge \beta = \lambda_\alpha(\beta) = f_\alpha(\beta) e_1 \wedge \cdots \wedge e_n = \langle \ast \alpha, \beta \rangle e_1 \wedge \cdots \wedge e_n,$$

as claimed. If $(e'_1, \ldots, e'_n)$ is any other positively oriented orthonormal basis, by Proposition 22.2 $e'_1 \wedge \cdots \wedge e'_n = \det(P) e_1 \wedge \cdots \wedge e_n = e_1 \wedge \cdots \wedge e_n$, since $\det(P) = 1$ where $P$ is the change of basis from $(e_1, \ldots, e_n)$ to $(e'_1, \ldots, e'_n)$ and both bases are positively oriented. □

The operator $*$ from $\bigwedge^k V$ to $\bigwedge^{n-k} V$ defined by Proposition 22.13 is obviously linear. It is called the Hodge $*$-operator.

The Hodge $*$-operator is defined in terms of the orthonormal basis elements of $\bigwedge V$ as follows: For any increasing sequence $(i_1, \ldots, i_k)$ of elements $i_p \in \{1, \ldots, n\}$, if $(j_1, \ldots, j_{n-k})$ is the increasing sequence of elements $j_q \in \{k_1, \ldots, k_n\}$ such that

$$\{i_1, \ldots, i_k\} \cup \{j_1, \ldots, j_{n-k}\} = \{1, \ldots, n\},$$

then

$$*(e_1 \wedge \cdots \wedge e_i) = \text{sign}(i_1, \ldots, i_k, j_1, \ldots, j_{n-k}) e_{j_1} \wedge \cdots \wedge e_{j_{n-k}}.$$

In particular, for $k = 0$ and $k = n$, we have

$$*(1) = e_1 \wedge \cdots \wedge e_n$$

$$*(e_1 \wedge \cdots \wedge e_n) = 1.$$

For example, for $n = 3$, we have

$$*e_1 = e_2 \wedge e_3$$
$$*e_2 = -e_1 \wedge e_3$$
$$*e_3 = e_1 \wedge e_2$$
$$*(e_1 \wedge e_2) = e_3$$
$$*(e_1 \wedge e_3) = -e_2$$
$$*(e_2 \wedge e_3) = e_1.$$
The Hodge \(+\)-operators \(\ast : \bigwedge^k V \to \bigwedge^{n-k} V\) induces a linear map \(\ast : \bigwedge(V) \to \bigwedge(V)\). We also have Hodge \(+\)-operators \(\ast : \bigwedge^k V^* \to \bigwedge^{n-k} V^*\).

The following proposition shows that the linear map \(\ast : \bigwedge(V) \to \bigwedge(V)\) is an isomorphism.

**Proposition 22.14.** If \(V\) is any oriented vector space of dimension \(n\), for every \(k\) with \(0 \leq k \leq n\), we have

(i) \(\ast^2 = (-\text{id})^{k(n-k)}\).

(ii) \(\langle x, y \rangle = \ast(x \wedge y) = \ast(y \wedge x)\), for all \(x, y \in \bigwedge^k V\).

**Proof.** (1) It is enough to check the identity on basis elements. We have

\[
\ast(e_{i_1} \wedge \cdots \wedge e_{i_k}) = \text{sign}(i_1, \ldots, i_k, j_1, \ldots, j_{n-k}) e_{j_1} \wedge \cdots \wedge e_{j_{n-k}}
\]

and

\[
\ast^2(e_{i_1} \wedge \cdots \wedge e_{i_k}) = \text{sign}(i_1, \ldots, i_k, j_1, \ldots, j_{n-k}) \ast(e_{j_1} \wedge \cdots \wedge e_{j_{n-k}})
\]

\[
= \text{sign}(i_1, \ldots, i_k, j_1, \ldots, j_{n-k}) \text{sign}(j_1, \ldots, j_{n-k}, i_1, \ldots, i_k) e_{i_1} \wedge \cdots \wedge e_{i_k}.
\]

Now, it is easy to see that

\[
\text{sign}(i_1, \ldots, i_k, j_1, \ldots, j_{n-k}) \text{sign}(j_1, \ldots, j_{n-k}, i_1, \ldots, i_k) = (-1)^{k(n-k)},
\]

which yields

\[
\ast^2(e_{i_1} \wedge \cdots \wedge e_{i_k}) = (-1)^{k(n-k)} e_{i_1} \wedge \cdots \wedge e_{i_k},
\]

as claimed.

(ii) These identities are easily checked on basis elements; see Jost [99], Chapter 2, Lemma 2.1.1. \(\square\)

In Section 27.2, we will need to express \(\ast(1)\) in terms of any basis (not necessarily orthonormal) of \(V\). If \((e_1, \ldots, e_n)\) is an orthonormal basis of \(V\) and \((v_1, \ldots, v_n)\) is any other basis of \(V\), then

\[
\langle v_1 \wedge \cdots \wedge v_n, v_1 \wedge \cdots \wedge v_n \rangle = \det(\langle v_i, v_j \rangle),
\]

and since

\[
v_1 \wedge \cdots \wedge v_n = \det(A) e_1 \wedge \cdots \wedge e_n
\]

where \(A\) is the matrix expressing the \(v_j\) in terms of the \(e_i\), we have

\[
\langle v_1 \wedge \cdots \wedge v_n, v_1 \wedge \cdots \wedge v_n \rangle = \det(A)^2 \langle e_1 \wedge \cdots \wedge e_n, e_1 \wedge \cdots \wedge e_n \rangle = \det(A)^2.
\]

As a consequence, \(\det(A) = \sqrt{\det(\langle v_i, v_j \rangle)}\), and

\[
v_1 \wedge \cdots \wedge v_n = \sqrt{\det(\langle v_i, v_j \rangle)} e_1 \wedge \cdots \wedge e_n,
\]

from which it follows that

\[
\ast(1) = \frac{1}{\sqrt{\det(\langle v_i, v_j \rangle)}} v_1 \wedge \cdots \wedge v_n
\]

(see Jost [99], Chapter 2, Lemma 2.1.3).
22.7 Testing Decomposability; Left and Right Hooks

In this section all vector spaces are assumed to have finite dimension. Say \( \text{dim}(E) = n \). Using our nonsingular pairing

\[
\langle -, - \rangle : \bigwedge^p E^* \times \bigwedge^p E \rightarrow K \quad (1 \leq p \leq n)
\]
defined on generators by

\[
\langle u_1^* \wedge \cdots \wedge u_p^*, v_1 \wedge \cdots \wedge u_p \rangle = \det(u_i^*(v_j)),
\]

we define various contraction operations (partial evaluation operators)

\[
\downarrow : \bigwedge^p E \times \bigwedge^{p+q} E^* \rightarrow \bigwedge^q E^* \quad \text{(left hook)}
\]

and

\[
\downarrow : \bigwedge^q E^* \times \bigwedge^{p+q} E \rightarrow \bigwedge^q E^* \quad \text{(right hook)},
\]
as well as the versions obtained by replacing \( E \) by \( E^* \) and \( E^{**} \) by \( E \). We begin with the left interior product or left hook, \( \downarrow \).

Let \( u \in \bigwedge^p E \). For any \( q \) such that \( p + q \leq n \), multiplication on the right by \( u \) is a linear map

\[
\wedge_R(u) : \bigwedge^q E \rightarrow \bigwedge^{p+q} E
\]
given by

\[
v \mapsto v \wedge u
\]
where \( v \in \bigwedge^q E \). The transpose of \( \wedge_R(u) \) yields a linear map

\[
(\wedge_R(u))^\top : \left( \bigwedge^{p+q} E \right)^* \rightarrow \left( \bigwedge^q E \right)^*,
\]

which, using the isomorphisms \( \left( \bigwedge^{p+q} E \right)^* \cong \bigwedge^{p+q} E^* \) and \( \left( \bigwedge^q E \right)^* \cong \bigwedge^q E^* \), can be viewed as a map

\[
(\wedge_R(u))^\top : \bigwedge^p E^* \rightarrow \bigwedge^q E^*
\]
given by

\[
z^* \mapsto z^* \circ \wedge_R(u),
\]
where \( z^* \in \bigwedge^{p+q} E^* \). We denote \( z^* \circ \wedge_R(u) \) by \( u \downarrow z^* \). In terms of our pairing, the adjoint \( u \downarrow \) of \( \wedge_R(u) \) defined by

\[
\langle u \downarrow z^*, v \rangle = \langle z^*, \wedge_R(u)(v) \rangle;
\]
that is, the \( q \)-vector \( u \bowtie z^* \) is uniquely determined by

\[
\langle u \bowtie z^*, v \rangle = \langle z^*, v \wedge u \rangle, \quad \text{for all } u \in \bigwedge^p E, v \in \bigwedge^q E \text{ and } z^* \in \bigwedge^{p+q} E^*.
\]

Note that to be precise the operator

\[
\bowtie : \bigwedge^p E \times \bigwedge^{p+q} E^* \to \bigwedge^q E^*
\]
depends of \( p, q \), so we really defined a family of operators \( \bowtie_{p,q} \). This family of operators \( \bowtie_{p,q} \) induces a map

\[
\bowtie : \bigwedge E \times \bigwedge E^* \to \bigwedge E^*,
\]

with

\[
\bowtie_{p,q} : \bigwedge^p E \times \bigwedge^{p+q} E^* \to \bigwedge^q E^*
\]
as defined before. The common practice is to omit the subscripts of \( \bowtie \). It is immediately verified that

\[
(u \wedge v) \bowtie z^* = u \bowtie (v \bowtie z^*),
\]

for all \( u \in \bigwedge^k E, v \in \bigwedge^{p-k} E, z^* \in \bigwedge^{p+q} E^* \) since

\[
\langle (u \wedge v) \bowtie z^*, w \rangle = \langle z^*, w \wedge u \wedge v \rangle = \langle v \bowtie z^*, w \wedge u \rangle = \langle u \bowtie (v \bowtie z^*), w \rangle,
\]

whenever \( w \in \bigwedge^q E \). This means that

\[
\bowtie : \bigwedge E \times \bigwedge E^* \to \bigwedge E^*
\]
is a left action of the (noncommutative) ring \( \bigwedge E \) with multiplication \( \wedge \) on \( \bigwedge E^* \), which makes \( \bigwedge E^* \) into a left \( \bigwedge \)-module.

By interchanging \( E \) and \( E^* \) and using the isomorphism

\[
\left( \bigwedge^k F \right)^* \cong \bigwedge^k F^*,
\]

we can also define some maps

\[
\bowtie : \bigwedge^p E^* \times \bigwedge^{p+q} E \to \bigwedge^q E.
\]

In terms of our pairing, \( u^* \bowtie z \) is uniquely defined by

\[
\langle v^* \wedge u^*, z \rangle = \langle v^*, u^* \bowtie z \rangle, \quad \text{for all } u^* \in \bigwedge^p E^*, v^* \in \bigwedge^q E^* \text{ and } z \in \bigwedge^{p+q} E.
\]

As for the previous version, we have a family of operators \( \bowtie_{p,q} \) which define an operator

\[
\bowtie : \bigwedge E^* \times \bigwedge E \to \bigwedge E.
\]
We easily verify that
\[(u^* \land v^*) \lor z = u^* \lor (v^* \lor z),\]
so this version of \( \lor \) is a left action of the ring \( \land E^* \) on \( \land E \) which makes \( \land E \) into a left \( \land E^* \)-module.

In order to proceed any further we need some combinatorial properties of the basis of \( \land^p E \) constructed from a basis \((e_1, \ldots, e_n)\) of \( E \). Recall that for any (nonempty) subset \( I \subseteq \{1, \ldots, n\} \), we let
\[e_I = e_{i_1} \land \cdots \land e_{i_p},\]
where \( I = \{i_1, \ldots, i_p\} \) with \( i_1 < \cdots < i_p \). We also let \( e_\emptyset = 1 \).

Given any two nonempty subsets \( H, L \subseteq \{1, \ldots, n\} \) both listed in increasing order, say \( H = \{h_1 < \cdots < h_p\} \) and \( L = \{\ell_1 < \cdots < \ell_q\} \), if \( H \) and \( L \) are disjoint, let \( H \cup L \) be union of \( H \) and \( L \) considered as the ordered sequence
\[(h_1, \ldots, h_p, \ell_1, \ldots, \ell_q).\]
Then let
\[\rho_{H,L} = \begin{cases} 0 & \text{if } H \cap L \neq \emptyset, \\ (-1)^\nu & \text{if } H \cap L = \emptyset, \end{cases}\]
where
\[\nu = |\{(h, l) \mid (h, l) \in H \times L, h > l\}|.\]
Observe that when \( H \cap L = \emptyset \), \( |H| = p \) and \( |L| = q \), the number \( \nu \) is the number of inversions of the sequence
\[(h_1, \cdots, h_p, \ell_1, \cdots, \ell_q),\]
where an inversion is a pair \((h_i, \ell_j)\) such that \( h_i > \ell_j \).

Unless \( p + q = n \), the function whose graph is given by
\[
\begin{pmatrix} 1 & \cdots & p & p+1 & \cdots & p+q \\ h_1 & \cdots & h_p & \ell_1 & \cdots & \ell_q \end{pmatrix}
\]
is not a permutation of \( \{1, \ldots, n\} \). We can view \( \nu \) as a slight generalization of the notion of the number of inversions of a permutation.

**Proposition 22.15.** For any basis \((e_1, \ldots, e_n)\) of \( E \) the following properties hold:

1. If \( H \cap L = \emptyset \), \( |H| = p \), and \( |L| = q \), then
   \[\rho_{H,L} \rho_{L,H} = (-1)^{pq}.\]

2. For \( H, L \subseteq \{1, \ldots, m\} \) listed in increasing order, we have
   \[e_H \land e_L = \rho_{H,L} e_{H \cup L}.\]

   Similarly,
   \[e_H^* \land e_L^* = \rho_{H,L} e_{H \cup L}^*.\]
(3) For the left hook
\[ \iota : \bigwedge^{p} E \times \bigwedge^{p+q} E^{\ast} \longrightarrow \bigwedge^{q} E^{\ast}, \]
we have
\[ e_{H} \iota e_{L}^{\ast} = 0 \quad \text{if } H \not\subseteq L \]
\[ e_{H} \iota e_{L}^{\ast} = \rho_{L-H,H} e_{L-H}^{\ast} \quad \text{if } H \subseteq L. \]

(4) For the left hook
\[ \iota : \bigwedge^{p} E^{\ast} \times \bigwedge^{p+q} E \longrightarrow \bigwedge^{q} E, \]
we have
\[ e_{H}^{\ast} \iota e_{L} = 0 \quad \text{if } H \not\subseteq L \]
\[ e_{H}^{\ast} \iota e_{L} = \rho_{L-H,H} e_{L-H} \quad \text{if } H \subseteq L. \]

Proof. These are proved in Bourbaki [26] (Chapter III, §11, Section 11), but the proofs of (3) and (4) are very concise. We elaborate on the proofs of (2) and (4), the proof of (3) being similar.

In (2) if \( H \cap L \neq \emptyset \), then \( e_{H} \iota e_{L} \) contains some vector twice and so \( e_{H} \iota e_{L} = 0 \). Otherwise, \( e_{H} \iota e_{L} \) consists of
\[ e_{h_{1}} \wedge \cdots \wedge e_{h_{p}} \wedge e_{\ell_{1}} \wedge \cdots \wedge e_{\ell_{q}}, \]
and to order the sequence of indices in increasing order we need to transpose any two indices \((h_{i}, \ell_{j})\) corresponding to an inversion, which yields \( \rho_{H,L} e_{H \cup L} \).

Let us now consider (4). We have \( |L| = p + q \) and \( |H| = p \), and the \( q \)-vector \( e_{H}^{\ast} \iota e_{L} \) is characterized by
\[ \langle v^{\ast}, e_{H}^{\ast} \iota e_{L} \rangle = \langle v^{\ast} \wedge e_{H}^{\ast}, e_{L} \rangle \]
for all \( v^{\ast} \in \bigwedge^{q} E^{\ast} \). For \( v^{\ast} = e_{L-H}^{\ast} \), by (2) we have
\[ \langle e_{L-H}^{\ast}, e_{H}^{\ast} \iota e_{L} \rangle = \langle e_{L-H}^{\ast} \wedge e_{H}^{\ast}, e_{L} \rangle = \langle \rho_{L-H,H} e_{L-H}^{\ast}, e_{L} \rangle = \rho_{L-H,H}, \]
which yields
\[ \langle e_{L-H}^{\ast}, e_{H}^{\ast} \iota e_{L} \rangle = \rho_{L-H,H}. \]
The \( q \)-vector \( e_{H}^{\ast} \iota e_{L} \) can be written as a linear combination \( e_{H}^{\ast} \iota e_{L} = \sum_{J} \lambda_{J} e_{J} \) with \( |J| = q \) so
\[ \langle e_{L-H}^{\ast}, e_{H}^{\ast} \iota e_{L} \rangle = \sum_{J} \lambda_{J} \langle e_{L-H}^{\ast}, e_{J} \rangle. \]
By definition of the pairing, \( \langle e_{L-H}^{\ast}, e_{J} \rangle = 0 \) unless \( J = L - H \), which means that
\[ \langle e_{L-H}^{\ast}, e_{H}^{\ast} \iota e_{L} \rangle = \lambda_{L-H} \langle e_{L-H}^{\ast}, e_{L-H} \rangle = \lambda_{L-H}, \]
so \( \lambda_{L-H} = \rho_{L-H} \), as claimed. \( \blacksquare \)
Using Proposition 22.15, we have the

**Proposition 22.16.** For the left hook

\[ \triangleright : E \times \bigwedge^{q+1} E^* \longrightarrow \bigwedge^q E^* , \]

for every \( u \in E , \ x^* \in \bigwedge^{q+1-s} E^* , \) and \( y^* \in \bigwedge^s E^* , \) we have

\[ u \triangleright (x^* \wedge y^*) = (-1)^s(u \triangleright x^*) \wedge y^* + x^* \wedge (u \triangleright y^*) . \]

**Proof.** We can prove the above identity assuming that \( x^* \) and \( y^* \) are of the form \( e^*_i \) and \( e^*_j \) using Proposition 22.15, but this is rather tedious.

Thus, \( \triangleright : E \times \bigwedge^{q+1} E^* \longrightarrow \bigwedge^q E^* \) is almost an anti-derivation, except that the sign \((-1)^s\) is applied to the wrong factor.

We have a similar identity for the other version of the left hook

\[ \triangleright : E^* \times \bigwedge^{q+1} E \longrightarrow \bigwedge^q E , \]

namely

\[ u^* \triangleright (x \wedge y) = (-1)^s(u^* \triangleright x) \wedge y + x \wedge (u^* \triangleright y) \]

for every \( u^* \in E^* \), \( x \in \bigwedge^{q+1-s} E \), and \( y \in \bigwedge^s E \).

An application of this formula when \( q = 3 \) and \( s = 2 \) yields an interesting equation. In this case, \( u^* \in E^* \) and \( x, y \in \bigwedge^2 E \), so we get

\[ u^* \triangleright (x \wedge y) = (u^* \triangleright x) \wedge y + x \wedge (u^* \triangleright y) . \]

In particular, for \( x = y \), since \( x \in \bigwedge^2 E \) and \( u^* \triangleright x \in E \), Proposition 22.10 implies that \( (u^* \triangleright x) \wedge x = x \wedge (u^* \triangleright x) \), and we obtain

\[ u^* \triangleright (x \wedge x) = 2((u^* \triangleright x) \wedge x) . \]

As a consequence, \( (u^* \triangleright x) \wedge x = 0 \) iff \( u^* \triangleright (x \wedge x) = 0 \). We will use this identity together with Proposition 22.20 to prove that a 2-vector \( x \in \bigwedge^2 E \) is decomposable iff \( x \wedge x = 0 \).

It is also possible to define a right interior product or right hook \( \llcorner \), using multiplication on the left rather than multiplication on the right. Then we have maps

\[ \llcorner : \bigwedge^{p+q} E^* \times \bigwedge^p E \longrightarrow \bigwedge^q E^* \]

given by

\[ \langle z^* \llcorner u, v \rangle = \langle z^*, u \wedge v \rangle , \quad \text{for all} \ u \in \bigwedge^p E , \ v \in \bigwedge^q E , \ \text{and} \ z^* \in \bigwedge^{p+q} E^* . \]
This time we can prove that
\[ z^\ast \lhd (u \wedge v) = (z^\ast \lhd u) \lhd v, \]
so the family of operators \( \lhd_{p,q} \) defines a right action
\[ \lhd : \bigwedge E^* \times \bigwedge E \to \bigwedge E^* \]
of the ring \( \bigwedge E \) on \( \bigwedge E^* \) which makes \( \bigwedge E^* \) into a right \( \bigwedge E \)-module.

Similarly, we have maps
\[ \lhd : \bigwedge E \times \bigwedge E^* \to \bigwedge E \]
given by
\[ \langle u^\ast \wedge v^\ast, z \rangle = \langle v^\ast, z \lhd u^\ast \rangle, \quad \text{for all } u^\ast \in \bigwedge^p E^*, \ v^\ast \in \bigwedge^q E^*, \text{ and } z \in \bigwedge^{p+q} E. \]
We can prove that
\[ z \lhd (u^\ast \wedge v^\ast) = (z \lhd u^\ast) \lhd v^\ast, \]
so the family of operators \( \lhd_{p,q} \) defines a right action
\[ \lhd : \bigwedge E \times \bigwedge E^* \to \bigwedge E \]
of the ring \( \bigwedge E^* \) on \( \bigwedge E \) which makes \( \bigwedge E \) into a right \( \bigwedge E^* \)-module.

Since the left hook \( \lhd : \bigwedge^p E \times \bigwedge^{p+q} E^* \to \bigwedge^q E^* \) is defined by
\[ \langle u \lhd z^\ast, v \rangle = \langle z^\ast, v \wedge u \rangle, \quad \text{for all } u \in \bigwedge^p E, \ v \in \bigwedge^q E \text{ and } z^\ast \in \bigwedge^{p+q} E^*, \]
the right hook
\[ \lhd : \bigwedge E^* \times \bigwedge E \to \bigwedge E^* \]
by
\[ \langle z^\ast \lhd u, v \rangle = \langle z^\ast, u \wedge v \rangle, \quad \text{for all } u \in \bigwedge^p E, \ v \in \bigwedge^q E, \text{ and } z^\ast \in \bigwedge^{p+q} E^*, \]
and \( v \wedge u = (-1)^{pq} u \wedge v \), we conclude that
\[ z^\ast \lhd u = (-1)^{pq} u \lhd z^\ast. \]

Similarly, since
\[ \langle v^\ast \wedge u^\ast, z \rangle = \langle v^\ast, u^\ast \lhd z \rangle, \quad \text{for all } u^\ast \in \bigwedge^p E^*, \ v^\ast \in \bigwedge^q E^* \text{ and } z \in \bigwedge^{p+q} E \]
\[ \langle u^\ast \wedge v^\ast, z \rangle = \langle v^\ast, z \lhd u^\ast \rangle, \quad \text{for all } u^\ast \in \bigwedge^p E^*, \ v^\ast \in \bigwedge^q E^* \text{ and } z \in \bigwedge^{p+q} E, \]
and \( v^* \land u^* = (-1)^{pq} u^* \land v^* \), we have
\[
z \downarrow u^* = (-1)^{pq} u^* \downarrow z.
\]
Therefore the left and right hooks are not independent, and in fact each one determines the other. As a consequence, we can restrict our attention to only one of the hooks, for example the left hook, but there are a few situations where it is nice to use both, for example in Proposition 22.19.

A version of Proposition 22.15 holds for right hooks, but beware that the indices in \( \rho_{L-H,H} \) are permuted. This permutation has to do with the fact that the left hook and the right hook are related via a sign factor.

**Proposition 22.17.** For any basis \((e_1, \ldots, e_n)\) of \(E\) the following properties hold:

1. For the right hook
   \[
   \downarrow : \bigwedge^p E \times \bigwedge^q E^* \rightarrow \bigwedge^E
   \]
   we have
   \[
   e_L \downarrow e_H^* = 0 \quad \text{if } H \not\subseteq L
   
   e_L \downarrow e_H^* = \rho_{H,L-H} e_{L-H} \quad \text{if } H \subseteq L.
   \]

2. For the right hook
   \[
   \downarrow : \bigwedge^p E^* \times \bigwedge^q E \rightarrow \bigwedge^E^*
   \]
   we have
   \[
   e_L^* \downarrow e_H = 0 \quad \text{if } H \not\subseteq L
   
   e_L^* \downarrow e_H = \rho_{H,L-H} e_{L-H}^* \quad \text{if } H \subseteq L.
   \]

**Remark:** Our definition of left hooks as left actions \( \downarrow : \bigwedge^p E \times \bigwedge^{p+q} E^* \rightarrow \bigwedge^q E^* \) and \( \downarrow : \bigwedge^p E^* \times \bigwedge^{p+q} E \rightarrow \bigwedge^q E \) and right hooks as right actions \( \downarrow : \bigwedge^{p+q} E^* \times \bigwedge^p E \rightarrow \bigwedge^q E^* \) and \( \downarrow : \bigwedge^{p+q} E \times \bigwedge^p E^* \rightarrow \bigwedge^q E \) is identical to the definition found in Fulton and Harris [70] (Appendix B). However, the reader should be aware that this is not a universally accepted notation. In fact, the left hook \( u^* \downarrow z \) defined in Bourbaki [26] is our right hook \( z \downarrow u^* \), up to the sign \((-1)^{p(p-1)/2}\). This has to do with the fact that Bourbaki uses a different pairing which also involves an extra sign, namely
\[
\langle v^*, u^* \downarrow z \rangle = (-1)^{p(p-1)/2}\langle u^* \land v^*, z \rangle.
\]
One of the side-effects of this choice is that Bourbaki’s version of formula (3) of Proposition 22.15 (Bourbaki [26], Chapter III, page 168) is

\[ e^*_H \cdot e_L = 0 \quad \text{if} \quad H \not\subseteq L \]
\[ e^*_H \cdot e_L = (-1)^{(p-1)/2} \rho_{H,L-H} e_{L-H} \quad \text{if} \quad H \subseteq L, \]

where \(|H| = p\) and \(|L| = p + q\). This correspond to formula (1) of Proposition 22.17 up to the sign factor \((-1)^{(p-1)/2}\), which we find horribly confusing. Curiously, an older edition of Bourbaki (1958) uses the same pairing as Fulton and Harris [70]. The reason (and the advantage) for this change of sign convention is not clear to us.

We also have the following version of Proposition 22.16 for the right hook:

**Proposition 22.18.** For the right hook

\[ \ll : \bigwedge^{q+1} E^* \times E \rightarrow \bigwedge^q E^*, \]

for every \(u \in E\), \(x^* \in \bigwedge^r E^*\), and \(y^* \in \bigwedge^{q+1-r} E^*\), we have

\[ (x^* \wedge y^*) \ll u = (x^* \ll u) \wedge y^* + (-1)^r x^* \wedge (y^* \ll u). \]

**Proof.** A proof involving determinants can be found in Warner [175], Chapter 2. \(\square\)

Thus, \(\ll : \bigwedge^{q+1} E^* \times E \rightarrow \bigwedge^q E^*\) is an anti-derivation. A similar formula holds for the right hook \(\ll : \bigwedge^{q+1} E \times E^* \rightarrow \bigwedge^q E\), namely

\[ (x \wedge y) \ll u^* = (x \ll u^*) \wedge y + (-1)^r x \wedge (y \ll u^*), \]

for every \(u^* \in E\), \(\in \bigwedge^r E\), and \(y \in \bigwedge^{q+1-r} E\). This formula is used by Shafarevitch [161] to define a hook, but beware that Shafarevitch use the left hook notation \(u^* \cdot x\) rather than the right hook notation. Shafarevitch uses the terminology *convolution*, which seems very unfortunate.

For \(u \in E\), the right hook \(z^* \ll u\) is also denoted \(i(u)z^*\), and called *insertion operator* or *interior product*. This operator plays an important role in differential geometry. If we view \(z^* \in \bigwedge^{n+1}(E^*)\) as an alternating multilinear map in \(\text{Alt}^{n+1}(E; K)\), then \(i(u)z^* \in \text{Alt}^n(E; K)\) is given by

\[ (i(u)z^*)(v_1, \ldots, v_n) = z^*(u, v_1, \ldots, v_n). \]

Using the left hook \(\cdot\) and the right hook \(\ll\) we can define linear maps \(\gamma : \bigwedge^p E \rightarrow \bigwedge^{n-p} E^*\) and \(\delta : \bigwedge^p E^* \rightarrow \bigwedge^{n-p} E\). For any basis \((e_1, \ldots, e_n)\) of \(E\), if we let \(M = \{1, \ldots, n\}\), \(e = e_1 \wedge \cdots \wedge e_n\), and \(e^* = e_1^\ast \wedge \cdots \wedge e_n^\ast\), then

\[ \gamma(u) = u \cdot e^* \quad \text{and} \quad \delta(v^*) = e \ll v^*, \]

for all \(u \in \bigwedge^p E\) and all \(v^* \in \bigwedge^p E^*\).
22.7. TESTING DECOMPOSABILITY; LEFT AND RIGHT HOOKS

Proposition 22.19. The linear maps $\gamma: \bigwedge^p E \rightarrow \bigwedge^{n-p} E^*$ and $\delta: \bigwedge^p E^* \rightarrow \bigwedge^{n-p} E$ are isomorphisms, and $\gamma^{-1} = \delta$. The isomorphisms $\gamma$ and $\delta$ map decomposable vectors to decomposable vectors. Furthermore, if $z \in \bigwedge^p E$ is decomposable, say $z = u_1 \wedge \cdots \wedge u_p$ for some $u_i \in E$, then $\gamma(z) = v_1^* \wedge \cdots \wedge v_{n-p}^*$ for some $v_j^* \in E^*$, and $v_j^*(u_i) = 0$ for all $i, j$. A similar property holds for $z$ composable vectors. Furthermore, if $\gamma$ isomorphisms, and $\gamma = \gamma^{-1} = \delta$. The linear maps $\gamma$ and $\delta$ are the corresponding isomorphisms, then $\gamma' = \lambda \gamma$ and $\delta' = \lambda^{-1} \delta$ for some nonzero $\lambda \in K$.

Proof. Using Propositions 22.15 and 22.17, for any subset $J \subseteq \{1, \ldots, n\} = M$ such that $|J| = p$, we have

$$\gamma(e_J) = e_J \wedge e^* = \rho_{M-J,J} e^*_{M-J}$$ and $$\delta(e^*_{M-J}) = e \wedge e_{M-J} = \rho_{M-J,J} e_J.$$

Thus,$$\delta \circ \gamma = \rho_{M-J,J} \rho_{M-J,J} e_J = e_J,$$

since $\rho_{M-J,J} = \pm 1$. A similar result holds for $\gamma \circ \delta$. This implies that

$$\delta \circ \gamma = \text{id} \quad \text{and} \quad \gamma \circ \delta = \text{id}.$$

Thus, $\gamma$ and $\delta$ are inverse isomorphisms.

If $z \in \bigwedge^p E$ is decomposable, then $z = u_1 \wedge \cdots \wedge u_p$ where $u_1, \ldots, u_p$ are linearly independent since $z \neq 0$, and we can pick a basis of $E$ of the form $(u_1, \ldots, u_n)$. Then the above formulae show that

$$\gamma(z) = \pm u_1^* \wedge \cdots \wedge u_n^*.$$

Since $(u_1^*, \ldots, u_n^*)$ is the dual basis of $(u_1, \ldots, u_n)$, we have $u_i^*(u_j) = \delta_{ij}$. If $(e_1', \ldots, e_n')$ is any other basis of $E$, because $\bigwedge^n E$ has dimension 1, we have

$$e_1' \wedge \cdots \wedge e_n' = \lambda e_1 \wedge \cdots \wedge e_n$$

for some nonzero $\lambda \in K$, and the rest is trivial. \hfill \square

Applying Proposition 22.19 to the case where $p = n - 1$, the isomorphism $\gamma: \bigwedge^{n-1} E \rightarrow \bigwedge^1 E^*$ maps indecomposable vectors in $\bigwedge^{n-1} E$ to indecomposable vectors in $\bigwedge^1 E^* = E^*$. But every vector in $E^*$ is decomposable, so every vector in $\bigwedge^{n-1} E$ is decomposable.

We are now ready to tackle the problem of finding criteria for decomposability. We need a few preliminary results.

Proposition 22.20. Given $z \in \bigwedge^p E$ with $z \neq 0$, the smallest vector space $W \subseteq E$ such that $z \in \bigwedge^p W$ is generated by the vectors of the form

$$u^* \wedge z, \quad \text{with} \quad u^* \in \bigwedge^{p-1} E^*.$$
CHAPTER 22. EXTERIOR TENSOR POWERS AND EXTERIOR ALGEBRAS

Proof. First let $W$ be any subspace such that $z \in \bigwedge^p(W)$ and let $(e_1, \ldots, e_r, e_{r+1}, \ldots, e_n)$ be a basis of $E$ such that $(e_1, \ldots, e_r)$ is a basis of $W$. Then, $u^* = \sum I \lambda_I e_I^*$, where $I \subseteq \{1, \ldots, n\}$ and $|I| = p - 1$, and $z = \sum J \mu_J e_J$, where $J \subseteq \{1, \ldots, r\}$ and $|J| = p \leq r$. It follows immediately from the formula of Proposition 22.15 (4), namely

\[ e_I^* \hook e_J = \rho_{J-I} e_{J-I}, \]

that $u^* \hook z \in W$, since $J - I \subseteq \{1, \ldots, r\}$.

Next we prove that if $W$ is the smallest subspace of $E$ such that $z \in \bigwedge^p(W)$, then $W$ is generated by the vectors of the form $u^* \hook z$, where $u^* \in \bigwedge^{p-1} E^*$. Suppose not, then the vectors $u^* \hook z$ with $u^* \in \bigwedge^{p-1} E^*$ span a proper subspace $U$ of $W$. We prove that for every subspace $W'$ of $W$ with $\dim(W') = \dim(W) - 1 = r - 1$, it is not possible that $u^* \hook z \in W'$ for all $u^* \in \bigwedge^{p-1} E^*$. But then, as $U$ is a proper subspace of $W$, it is contained in some subspace $W''$ with $\dim(W'') = r - 1$, and we have a contradiction.

Let $w \in W - W'$ and pick a basis of $W$ formed by a basis $(e_1, \ldots, e_{r-1})$ of $W'$ and $w$. Any $z \in \bigwedge^p(W)$ can be written as $z = z' + w \wedge z''$, where $z' \in \bigwedge^p W'$ and $z'' \in \bigwedge^{p-1} W'$, and since $W$ is the smallest subspace containing $z$, we have $z'' \neq 0$. Consequently, if we write $z'' = \sum I \lambda_I e_I$ in terms of the basis $(e_1, \ldots, e_{r-1})$ of $W'$, there is some $e_I$, with $I \subseteq \{1, \ldots, r-1\}$ and $|I| = p - 1$, so that the coefficient $\lambda_I$ is nonzero. Now, using any basis of $E$ containing $(e_1, \ldots, e_{r-1}, w)$, by Proposition 22.15 (4), we see that

\[ e_I^* \hook (w \wedge e_I) = \lambda w, \quad \lambda = \pm 1. \]

It follows that

\[ e_I^* \hook z = e_I^* \hook (z' + w \wedge z'') = e_I^* \hook z' + e_I^* \hook (w \wedge z'') = e_I^* \hook z' + \lambda w, \]

with $e_I^* \hook z' \in W'$, which shows that $e_I^* \hook z \notin W'$. Therefore, $W$ is indeed generated by the vectors of the form $u^* \hook z$, where $u^* \in \bigwedge^{p-1} E^*$. \qed

To help understand Proposition 22.20, let $E$ be the vector space with basis $\{e_1, e_2, e_3, e_4\}$ and $z = e_1 \wedge e_2 + e_2 \wedge e_3$. Note that $z \in \bigwedge^2 E$. To find the smallest vector space $W \subseteq E$ such that $z \in \bigwedge^2 W$, we calculate $u^* \hook z$, where $u^* \in \bigwedge^1 E^*$. The multilinearity of $\hook$ implies it is enough to calculate $u^* \hook z$ for $u^* \in \{e_1^*, e_2^*, e_3^*, e_4^*\}$. Proposition 22.15 (4) implies that

\[
\begin{align*}
e_1^* \hook z &= e_1^* \hook (e_1 \wedge e_2 + e_2 \wedge e_3) = e_1^* \hook e_1 \wedge e_2 = e_2, \\
e_2^* \hook z &= e_2^* \hook (e_1 \wedge e_2 + e_2 \wedge e_3) = e_2 - e_3, \\
e_3^* \hook z &= e_3^* \hook (e_1 \wedge e_2 + e_2 \wedge e_3) = e_3 \hook e_2 \wedge e_3 = e_2, \\
e_4^* \hook z &= e_4^* \hook (e_1 \wedge e_2 + e_2 \wedge e_3) = 0.
\end{align*}
\]

Thus $W$ is the two-dimensional vector space generated by the basis $\{e_2, e_1 + e_3\}$. This is not surprising since $z = e_2 \wedge (e_1 + e_3)$ and is in fact decomposable. As this example demonstrates, the action of the left hook provides a way of extracting a basis of $W$ from $z$.

Proposition 22.20 implies the following corollary.
Corollary 22.21. Any nonzero \( z \in \bigwedge^p E \) is decomposable iff the smallest subspace \( W \) of \( E \) such that \( z \in \bigwedge^p W \) has dimension \( p \). Furthermore, if \( z = u_1 \wedge \cdots \wedge u_p \) is decomposable, then \( (u_1, \ldots, u_p) \) is a basis of the smallest subspace \( W \) of \( E \) such that \( z \in \bigwedge^p W \).

Proof. If \( \dim(W) = p \), then for any basis \( (e_1, \ldots, e_p) \) of \( W \) we know that \( \bigwedge^p W \) has \( e_1 \wedge \cdots \wedge e_p \) has a basis, and thus has dimension 1. Since \( z \in \bigwedge^p W \), we have \( z = \lambda e_1 \wedge \cdots \wedge e_p \) for some nonzero \( \lambda \), so \( z \) is decomposable.

Conversely assume that \( z \in \bigwedge^p W \) is nonzero and decomposable. Then, \( z = u_1 \wedge \cdots \wedge u_p \), and since \( z \neq 0 \), by Proposition 22.20, \( (u_1, \ldots, u_p) \) are linearly independent. Then, for any \( v_i = u_1^* \wedge \cdots \wedge u_{i-1}^* \wedge u_{i+1}^* \wedge \cdots \wedge u_p^* \) (where \( u_i^* \) is omitted), we have

\[
v_i \bullet z = (u_1^* \wedge \cdots \wedge u_{i-1}^* \wedge u_{i+1}^* \wedge \cdots \wedge u_p^*) \wedge (u_1 \wedge \cdots \wedge u_p) = \pm u_i,
\]

so by Proposition 22.20 we have \( u_i \in W \) for \( i = 1, \ldots, p \). This shows that \( \dim(W) \geq p \), but since \( z = u_1 \wedge \cdots \wedge u_p \), we have \( \dim(W) = p \), which means that \( (u_1, \ldots, u_p) \) is a basis of \( W \).

Finally we are ready to state and prove the criterion for decomposability with respect to left hooks.

Proposition 22.22. Any nonzero \( z \in \bigwedge^p E \) is decomposable iff

\[
(u^* \bullet z) \wedge z = 0, \quad \text{for all } u^* \in \bigwedge^{p-1} E^*.
\]

Proof. First assume that \( z \in \bigwedge^p E \) is decomposable. If so, the smallest subspace \( W \) of \( E \) such that \( z \in \bigwedge^p W \) has dimension \( p \), so we have \( z = e_1 \wedge \cdots \wedge e_p \) where \( e_1, \ldots, e_p \) form a basis of \( W \). By Proposition 22.20, for every \( u^* \in \bigwedge^{p-1} E^* \), we have \( u^* \bullet z \in W \), so each \( u^* \bullet z \) is a linear combination of the \( e_i \)'s, say

\[
u^* \bullet z = \alpha_1 e_1 + \cdots + \alpha_p e_p,
\]

and

\[
(u^* \bullet z) \wedge z = \sum_{i=1}^{p} \alpha_i e_i \wedge e_1 \wedge \cdots \wedge e_i \wedge \cdots \wedge e_p = 0.
\]

Now assume that \((u^* \bullet z) \wedge z = 0\) for all \( u^* \in \bigwedge^{p-1} E^* \), and that \( \dim(W) = m > p \), where \( W \) is the smallest subspace of \( E \) such that \( z \in \bigwedge^p W \). If \( e_1, \ldots, e_m \) is a basis of \( W \), then we have \( z = \sum_I \lambda_I e_I \), where \( I \subseteq \{1, \ldots, m\} \) and \( |I| = p \). Recall that \( z \neq 0 \), and so, some \( \lambda_I \) is nonzero. By Proposition 22.20, each \( e_i \) can be written as \( u^* \bullet z \) for some \( u^* \in \bigwedge^{p-1} E^* \), and since \((u^* \bullet z) \wedge z = 0\) for all \( u^* \in \bigwedge^{p-1} E^* \), we get

\[
e_j \wedge z = 0 \quad \text{for } \quad j = 1, \ldots, m.
\]

By wedging \( z = \sum_I \lambda_I e_I \) with each \( e_j \), as \( m > p \), we deduce \( \lambda_I = 0 \) for all \( I \), so \( z = 0 \), a contradiction. Therefore, \( m = p \) and \( z \) is decomposable.
As a corollary of Proposition 22.22 we obtain the following fact that we stated earlier without proof.

**Proposition 22.23.** Given any vector space $E$ of dimension $n$, a vector $x \in \bigwedge^2 E$ is decomposable iff $x \wedge x = 0$.

**Proof.** Recall that as an application of Proposition 22.16 we proved the formula ((†)), namely

$$u^* \lrcorner (x \wedge x) = 2((u^* \lrcorner x) \wedge x)$$

for all $x \in \bigwedge^2 E$ and all $u^* \in E^*$. As a consequence, $(u^* \lrcorner x) \wedge x = 0$ iff $u^* \lrcorner (x \wedge x) = 0$. By Proposition 22.22, the 2-vector $x$ is decomposable iff $u^* \lrcorner (x \wedge x) = 0$ for all $u^* \in E^*$ iff $x \wedge x = 0$. Therefore, a 2-vector $x$ is decomposable iff $x \wedge x = 0$. \qed

As an application, assume that $\dim(E) = 3$ and that $(e_1, e_2, e_3)$ is a basis of $E$. Then any 2-vector $x \in \bigwedge^2 E$ is of the form

$$x = \alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3.$$  

We have

$$x \wedge x = (\alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3) \wedge (\alpha e_1 \wedge e_2 + \beta e_1 \wedge e_2 + \gamma e_2 \wedge e_3) = 0,$$

because all the terms involved are of the form $ce_{i_1} \wedge e_{i_2} \wedge e_{i_3} \wedge e_{i_4}$ with $i_1, i_2, i_3, i_4 \in \{1, 2, 3\}$, and so at least two of these indices are identical. Therefore, every 2-vector $x = \alpha e_1 \wedge e_2 + \beta e_1 \wedge e_3 + \gamma e_2 \wedge e_3$ is decomposable, although this not obvious at first glance. For example,

$$e_1 \wedge e_2 + e_1 \wedge e_3 + e_2 \wedge e_3 = (e_1 + e_2) \wedge (e_2 + e_3).$$

We now show that Proposition 22.22 yields an equational criterion for the decomposability of an alternating tensor $z \in \bigwedge^p E$.

### 22.8 The Grassmann-Plücker’s Equations and Grassmannian Manifolds

We follow an argument adapted from Bourbaki [26] (Chapter III, §11, Section 13).

Let $E$ be a vector space of dimensions $n$, let $(e_1, \ldots, e_n)$ be a basis of $E$, and let $(e_1^*, \ldots, e_n^*)$ be its dual basis. Our objective is to determine whether a nonzero vector $z \in \bigwedge^p E$ is decomposable. By Proposition 22.22, the vector $z$ is decomposable iff $(u^* \lrcorner z) \wedge z = 0$ for all $u^* \in \bigwedge^{p-1} E^*$. We can let $u^*$ range over a basis of $\bigwedge^{p-1} E^*$, and then the conditions are

$$(e_H^* \lrcorner z) \wedge z = 0$$
for all $H \subseteq \{1, \ldots, n\}$, with $|H| = p - 1$. Since $(e_H^* \cup z) \wedge z \in \bigwedge^{p+1} E$, this is equivalent to

$$\langle e_J^*, (e_H^* \cup z) \wedge z \rangle = 0$$

for all $H, J \subseteq \{1, \ldots, n\}$, with $|H| = p - 1$ and $|J| = p + 1$. Then, for all $I, I' \subseteq \{1, \ldots, n\}$ with $|I| = |I'| = p$, Formulae (2) and (4) of Proposition 22.15 (4) show that

$$\langle e_J^*, (e_H^* \cup e_I) \wedge e_{I'} \rangle = 0,$$

unless there is some $i \in \{1, \ldots, n\}$ such that $I - H = \{i\}$, $J - I' = \{i\}$. In this case, $I = H \cup \{i\}$ and $I' = J - \{i\}$, and using Formulae (2) and (4) of Proposition 22.15 (4), we have

$$\langle e_J^*, (e_H^* \cup e_{H \cup \{i\}}) \wedge e_{J - \{i\}} \rangle = \rho_{\{i\}, H} \rho_{\{i\}, J - \{i\}}.$$

If we let

$$\epsilon_{i,j,H} = \rho_{\{i\}, H} \rho_{\{i\}, J - \{i\}},$$

we have $\epsilon_{i,j,H} = +1$ if the parity of the number of $j \in J$ such that $j < i$ is the same as the parity of the number of $h \in H$ such that $h < i$, and $\epsilon_{i,j,H} = -1$ otherwise.

Finally, we obtain the following criterion in terms of quadratic equations (Plücker’s equations) for the decomposability of an alternating tensor:

**Proposition 22.24. (Grassmann-Plücker’s Equations)** For $z = \sum I \lambda_I e_I \in \bigwedge^p E$, the conditions for $z \neq 0$ to be decomposable are

$$\sum_{i \in J - H} \epsilon_{i,j,H} \lambda_{H \cup \{i\}} \lambda_{J - \{i\}} = 0,$$

with $\epsilon_{i,j,H} = \rho_{\{i\}, H} \rho_{\{i\}, J - \{i\}}$, for all $H, J \subseteq \{1, \ldots, n\}$ such that $|H| = p - 1$, $|J| = p + 1$, and all $i \in J - H$.

Using the above criterion, it is a good exercise to reprove that if $\text{dim}(E) = n$, then every tensor in $\bigwedge^{n-1}(E)$ is decomposable. We already proved this fact as a corollary of Proposition 22.19.

Given any $z = \sum I \lambda_I e_I \in \bigwedge^p E$ where $\text{dim}(E) = n$, the family of scalars $(\lambda_I)$ (with $I = \{i_1 < \cdots < i_p\} \subseteq \{1, \ldots, n\}$ listed in increasing order) is called the Plücker coordinates of $z$. The Grassmann-Plücker’s Equations give necessary and sufficient conditions for any nonzero $z$ to be decomposable.

For example, when $\text{dim}(E) = n = 4$ and $p = 2$, these equations reduce to the single equation

$$\lambda_{12} \lambda_{34} - \lambda_{13} \lambda_{24} + \lambda_{14} \lambda_{23} = 0.$$
However, it should be noted that the equations given by Proposition 22.24 are not independent in general.

We are now in the position to prove that the Grassmannian $G(p, n)$ can be embedded in the projective space $\mathbb{P}^{(p)-1}$, a fact that we stated in Section 5.3 without proof.

For any $n \geq 1$ and any $k$ with $1 \leq p \leq n$, recall that the Grassmannian $G(p, n)$ is the set of all linear $p$-dimensional subspaces of $\mathbb{R}^n$ (also called $p$-planes). Any $p$-dimensional subspace $U$ of $\mathbb{R}^n$ is spanned by $p$ linearly independent vectors $u_1, \ldots, u_p$ in $\mathbb{R}^n$; write $U = \text{span}(u_1, \ldots, u_k)$. By Proposition 22.7, $(u_1, \ldots, u_p)$ are linearly independent iff $u_1 \wedge \cdots \wedge u_p \neq 0$. If $(v_1, \ldots, v_p)$ are any other linearly independent vectors spanning $U$, then we have

$$v_i = \sum_{j=1}^{p} a_{ij} u_j, \quad 1 \leq i \leq p,$$

for some $a_{ij} \in \mathbb{R}$, and by Proposition 22.2

$$v_1 \wedge \cdots \wedge v_p = \det(A) \ u_1 \wedge \cdots \wedge u_p,$$

where $A = (a_{ij})$. As a consequence, we can define a map $i_G: G(p, n) \rightarrow \mathbb{P}^{(p)-1}$ such that for any $k$-plane $U$, for any basis $(u_1, \ldots, u_p)$ of $U$,

$$i_G(U) = [u_1 \wedge \cdots \wedge u_p],$$

the point of $\mathbb{P}^{(p)-1}$ given by the one-dimensional subspace of $\mathbb{P}^{(p)}$ spanned by $u_1 \wedge \cdots \wedge u_p$.

**Proposition 22.25.** The map $i_G: G(p, n) \rightarrow \mathbb{P}^{(p)-1}$ is injective.

**Proof.** Let $U$ and $V$ be any two $p$-planes and assume that $i_G(U) = i_G(V)$. This means that there is a basis $(u_1, \ldots, u_p)$ of $U$ and a basis $(v_1, \ldots, v_p)$ of $V$ such that

$$v_1 \wedge \cdots \wedge v_p = c \ u_1 \wedge \cdots \wedge u_p,$$

for some nonzero $c \in \mathbb{R}$. The above implies that the smallest subspaces $W$ and $W'$ of $\mathbb{R}^n$ such that $u_1 \wedge \cdots \wedge u_p \in \text{span}W$ and $v_1 \wedge \cdots \wedge v_p \in \text{span}W'$ are identical, so $W = W'$. By Corollary 22.21, this smallest subspace $W$ has both $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_p)$ as bases, so the $v_j$ are linear combinations of the $u_i$ (and vice-versa), and $U = V$. \hfill $\Box$

Since any nonzero $z \in \bigwedge^p \mathbb{R}^n$ can be uniquely written as

$$z = \sum_{I} \lambda_{I} e_I$$

in terms of its Plücker coordinates $(\lambda_I)$, every point of $\mathbb{P}^{(p)-1}$ is defined by the Plücker coordinates $(\lambda_I)$ viewed as homogeneous coordinates. The points of $\mathbb{P}^{(p)-1}$ corresponding to one-dimensional spaces associated with decomposable alternating $p$-tensors are the
points whose coordinates satisfy the Grassmann-Plücker’s equations of Proposition 22.24. Therefore, the map $i_G$ embeds the Grassmannian $G(p, n)$ as an algebraic variety in $\mathbb{RP}^{(n) - 1}$ defined by equations of degree 2.

We can replace the field $\mathbb{R}$ by $\mathbb{C}$ in the above reasoning and we obtain an embedding of the complex Grassmannian $G_C(p, n)$ as an algebraic variety in $\mathbb{CP}^{(n) - 1}$ defined by equations of degree 2.

In particular, if $n = 4$ and $p = 2$, the equation

$$\lambda_{12}\lambda_{34} - \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} = 0$$

is the homogeneous equation of a quadric in $\mathbb{CP}^5$ known as the *Klein quadric*. The points on this quadric are in one-to-one correspondence with the lines in $\mathbb{CP}^3$.

There is also a simple algebraic criterion to decide whether the smallest subspaces $U$ and $V$ associated with two nonzero decomposable vector $u_1 \wedge \cdots \wedge u_p$ and $v_1 \wedge \cdots \wedge v_q$ have a nontrivial intersection.

**Proposition 22.26.** Let $E$ be any $n$-dimensional vector space over a field $K$, and let $U$ and $V$ be the smallest subspaces of $E$ associated with two nonzero decomposable vector $u = u_1 \wedge \cdots \wedge u_p \in \bigwedge^p U$ and $v = v_1 \wedge \cdots \wedge v_q \in \bigwedge^q V$. The following properties hold:

1. We have $U \cap V = (0)$ iff $u \wedge v \neq 0$.

2. If $U \cap V = (0)$, then $U + V$ is the least subspace associated with $u \wedge v$.

**Proof.** Assume $U \cap V = (0)$. We know by Corollary 22.21 that $(u_1, \ldots, u_p)$ is a basis of $U$ and $(v_1, \ldots, v_q)$ is a basis of $V$. Since $U \cap V = (0)$, $(u_1, \ldots, u_p, v_1, \ldots, v_q)$ is a basis of $U + V$, and by Proposition 22.7, we have

$$u \wedge v = u_1 \wedge \cdots \wedge u_p \wedge v_1 \wedge \cdots \wedge v_q \neq 0.$$  

This also proves (2).

Conversely, assume that $\dim(U \cap V) \geq 1$. Pick a basis $(w_1, \ldots, w_r)$ of $W$, and extend this basis to a basis $(w_1, \ldots, w_r, w_{r+1}, \ldots, w_p)$ of $U$ and to a basis $(w_1, \ldots, w_r, w_{p+1}, \ldots, w_{p+q-r})$ of $V$. By Corollary 22.21, $(u_1, \ldots, u_p)$ is also basis of $U$, so

$$u_1 \wedge \cdots \wedge u_p = a w_1 \wedge \cdots \wedge w_r \wedge w_{r+1} \wedge \cdots \wedge w_p$$

for some $a \in K$, and $(v_1, \ldots, v_q)$ is also basis of $V$, so

$$v_1 \wedge \cdots \wedge v_q = b w_1 \cdots \wedge w_r \wedge w_{p+1} \wedge \cdots \wedge w_{p+q-r}$$

for some $b \in K$, and thus

$$u \wedge v = u_1 \wedge \cdots \wedge u_p \wedge v_1 \wedge \cdots \wedge v_q = 0$$

since it contains some repeated $w_i$, with $1 \leq i \leq r$. $\square$
CHAPTER 22. EXTERIOR TENSOR POWERS AND EXTERIOR ALGEBRAS

As an application of Proposition 22.26, consider two projective lines \( D_1 \) and \( D_2 \) in \( \mathbb{RP}^3 \), which means that \( D_1 \) and \( D_2 \) correspond to two 2-planes in \( \mathbb{R}^4 \), and thus to two points in \( \mathbb{RP}^{(2)} = \mathbb{RP}^5 \). These two points correspond to the 2-vectors

\[
z = a_{1,2}e_1 \wedge e_2 + a_{1,3}e_1 \wedge e_3 + a_{1,4}e_1 \wedge e_4 + a_{2,3}e_2 \wedge e_3 + a_{2,4}e_2 \wedge e_4 + a_{3,4}e_3 \wedge e_4
\]

and

\[
z' = a'_{1,2}e_1 \wedge e_2 + a'_{1,3}e_1 \wedge e_3 + a'_{1,4}e_1 \wedge e_4 + a'_{2,3}e_2 \wedge e_3 + a'_{2,4}e_2 \wedge e_4 + a'_{3,4}e_3 \wedge e_4
\]

whose Plücker coordinates satisfy the equation

\[
\lambda_{12}\lambda_{34} - \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} = 0
\]

of the Klein quadric, and \( D_1 \) and \( D_2 \) intersect iff \( z \wedge z' = 0 \) iff

\[
a_{1,2}a'_{3,4} - a_{1,3}a'_{3,4} + a_{1,4}a'_{2,3} + a_{2,3}a'_{1,4} - a_{2,4}a'_{1,3} + a_{3,4}a'_{1,2} = 0.
\]

Observe that for \( D_1 \) fixed, this is a linear condition. This fact is very helpful for solving problems involving intersections of lines. A famous problem is to find how many lines in \( \mathbb{RP}^3 \) meet four given lines in general position. The answer is at most 2.

### 22.9 Vector-Valued Alternating Forms

The purpose of this section is to present the technical background needed for Sections 23.4 and 23.5 on vector-valued differential forms, in particular in the case of Lie groups where differential forms taking their values in a Lie algebra arise naturally.

In this section the vector space \( E \) is assumed to have finite dimension. We know that there is a canonical isomorphism \( \bigwedge^n(E^*) \cong \text{Alt}^n(E; K) \) between alternating \( n \)-forms and alternating multilinear maps. As in the case of general tensors, the isomorphisms provided by Propositions 22.5, 21.16, and 22.8, namely

\[
\text{Alt}^n(E; F) \cong \text{Hom}\left(\bigwedge^n(E), F\right)
\]

\[
\text{Hom}\left(\bigwedge^n(E), F\right) \cong \left(\bigwedge^n(E)\right)^* \otimes F
\]

\[
\left(\bigwedge^n(E)\right)^* \cong \bigwedge^n(E^*)
\]

yield a canonical isomorphism

\[
\text{Alt}^n(E; F) \cong \left(\bigwedge^n(E^*)\right) \otimes F.
\]
Note that $F$ may have infinite dimension. This isomorphism allows us to view the tensors in $\bigwedge^n(E^*) \otimes F$ as \textit{vector-valued alternating forms}, a point of view that is useful in differential geometry. If $(f_1, \ldots, f_r)$ is a basis of $F$, every tensor $\omega \in \bigwedge^n(E^*) \otimes F$ can be written as some linear combination

$$\omega = \sum_{i=1}^{r} \alpha_i \otimes f_i,$$

with $\alpha_i \in \bigwedge^n(E^*)$. We also let

$$\bigwedge(E; F) = \bigoplus_{n=0}^{\infty} \big(\bigwedge^n(E^*) \otimes F\big).$$

Given three vector spaces, $F, G, H$, if we have some bilinear map $\Phi: F \times G \to H$, then we can define a multiplication operation

$$\wedge \Phi: \bigwedge(E; F) \times \bigwedge(E; G) \to \bigwedge(E; H)$$

as follows: For every pair $(m, n)$, we define the multiplication

$$\wedge \Phi: \left(\left(\bigwedge^m(E^*) \otimes F\right) \times \left(\bigwedge^n(E^*) \otimes G\right)\right) \to \left(\bigwedge^{m+n}(E^*)\right) \otimes H$$

by

$$(\alpha \otimes f) \wedge \Phi (\beta \otimes g) = (\alpha \wedge \beta) \otimes \Phi(f, g).$$

As in Section 22.5 (following H. Cartan [37]) we can also define a multiplication

$$\wedge \Phi: \text{Alt}^m(E; F) \times \text{Alt}^n(E; G) \to \text{Alt}^{m+n}(E; H)$$

directly on alternating multilinear maps as follows: For $f \in \text{Alt}^m(E; F)$ and $g \in \text{Alt}^n(E; G)$,

$$(f \wedge \Phi g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \text{shuffle}(m, n)} \text{sgn}(\sigma) \Phi(f(u_{\sigma(1)}, \ldots, u_{\sigma(m)}), g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)})�$$

where \text{shuffle}(m, n) consists of all $(m, n)$-“shuffles;” that is, permutations $\sigma$ of $\{1, \ldots, m+n\}$ such that $\sigma(1) < \cdots < \sigma(m)$ and $\sigma(m+1) < \cdots < \sigma(m+n)$.

A special case of interest is the case where $F = G = H$ is a Lie algebra and $\Phi(a, b) = [a, b]$ is the Lie bracket of $F$. In this case, using a basis $(f_1, \ldots, f_r)$ of $F$, if we write $\omega = \sum_i \alpha_i \otimes f_i$ and $\eta = \sum_j \beta_j \otimes f_j$, we have

$$\omega \wedge \Phi \eta = [\omega, \eta] = \sum_{i,j} \alpha_i \wedge \beta_j \otimes [f_i, f_j].$$
It is customary to denote $\omega \wedge \Phi \eta$ by $[\omega, \eta]$ (unfortunately, the bracket notation is overloaded). Consequently, 

$$[\eta, \omega] = (-1)^{mn+1}[\omega, \eta].$$

In general not much can be said about $\wedge \Phi$, unless $\Phi$ has some additional properties. In particular, $\wedge \Phi$ is generally not associative.

We now use vector-valued alternating forms to generalize both the $\mu$ map of Proposition 22.12 and generalize Proposition 21.16 by defining the map

$$\mu_F: \left(\wedge^n(E^*)\right) \otimes F \rightarrow \text{Alt}^n(E; F)$$

on generators by

$$\mu_F((v_1^* \wedge \cdots \wedge v_n^*) \otimes f)(u_1, \ldots, u_n) = (\det(v_j^*(u_i)))f,$$

with $v_1^*, \ldots, v_n^* \in E^*$, $u_1, \ldots, u_n \in E$, and $f \in F$.

**Proposition 22.27.** The map

$$\mu_F: \left(\wedge^n(E^*)\right) \otimes F \rightarrow \text{Alt}^n(E; F)$$

defined as above is a canonical isomorphism for every $n \geq 0$. Furthermore, given any three vector spaces, $F, G, H$, and any bilinear map $\Phi: F \times G \rightarrow H$, for all $\omega \in (\wedge^n(E^*)) \otimes F$ and all $\eta \in (\wedge^n(E^*)) \otimes G$,

$$\mu_H(\omega \wedge \Phi \eta) = \mu_F(\omega) \wedge \Phi \mu_G(\eta).$$

**Proof.** Since we already know that $(\wedge^n(E^*)) \otimes F$ and $\text{Alt}^n(E; F)$ are isomorphic, it is enough to show that $\mu_F$ maps some basis of $(\wedge^n(E^*)) \otimes F$ to linearly independent elements. Pick some bases $(e_1, \ldots, e_p)$ in $E$ and $(f_j)_{j \in J}$ in $F$. Then, we know that the vectors $e_I^* \otimes f_j$, where $I \subseteq \{1, \ldots, p\}$ and $|I| = n$, form a basis of $(\wedge^n(E^*)) \otimes F$. If we have a linear dependence

$$\sum_{I,j} \lambda_{I,j} \mu_F(e_I^* \otimes f_j) = 0,$$

applying the above combination to each $(e_{i_1}, \ldots, e_{i_n})$ ($I = \{i_1, \ldots, i_n\}$, $i_1 < \cdots < i_n$), we get the linear combination

$$\sum_j \lambda_{I,j} f_j = 0,$$

and by linear independence of the $f_j$’s, we get $\lambda_{I,j} = 0$ for all $I$ and all $j$. Therefore, the $\mu_F(e_I^* \otimes f_j)$ are linearly independent, and we are done. The second part of the proposition is checked using a simple computation. \qed
The following proposition will be useful in dealing with vector-valued differential forms:

**Proposition 22.28.** If \((e_1, \ldots, e_p)\) is any basis of \(E\), then every element \(\omega \in (\bigwedge^n (E^*)) \otimes F\) can be written in a unique way as

\[
\omega = \sum_I e^*_I \otimes f_I, \quad f_I \in F,
\]

where the \(e^*_I\) are defined as in Section 22.2.

**Proof.** Since, by Proposition 22.6, the \(e^*_I\) form a basis of \(\bigwedge^n (E^*)\), elements of the form \(e^*_I \otimes f\) span \((\bigwedge^n (E^*)) \otimes F\). Now, if we apply \(\mu_F(\omega)\) to \((e_i_1, \ldots, e_{i_n})\), where \(I = \{i_1, \ldots, i_n\} \subseteq \{1, \ldots, p\}\), we get

\[
\mu_F(\omega)(e_{i_1}, \ldots, e_{i_n}) = \mu_F(e^*_I \otimes f_I)(e_{i_1}, \ldots, e_{i_n}) = f_I.
\]

Therefore, the \(f_I\) are uniquely determined by \(f\).

Proposition 22.28 can also be formulated in terms of alternating multilinear maps, a fact that will be useful to deal with differential forms.

Define the product \(\cdot : \text{Alt}^n(E; \mathbb{R}) \times F \to \text{Alt}^n(E; F)\) as follows: For all \(\omega \in \text{Alt}^n(E; \mathbb{R})\) and all \(f \in F\),

\[
(\omega \cdot f)(u_1, \ldots, u_n) = \omega(u_1, \ldots, u_n)f,
\]

for all \(u_1, \ldots, u_n \in E\). Then, it is immediately verified that for every \(\omega \in (\bigwedge^n (E^*)) \otimes F\) of the form

\[
\omega = u_1^* \wedge \cdots \wedge u_n^* \otimes f,
\]

we have

\[
\mu_F(u_1^* \wedge \cdots \wedge u_n^* \otimes f) = \mu_F(u_1^* \wedge \cdots \wedge u_n^*) \cdot f.
\]

Then Proposition 22.28 yields

**Proposition 22.29.** If \((e_1, \ldots, e_p)\) is any basis of \(E\), then every element \(\omega \in \text{Alt}^n(E; F)\) can be written in a unique way as

\[
\omega = \sum_I e^*_I \cdot f_I, \quad f_I \in F,
\]

where the \(e^*_I\) are defined as in Section 22.2.
The Pfaffian Polynomial

The results of this section will be needed to define the Euler class of a real orientable rank $2n$ vector bundle; see Section 29.7.

Let $\mathfrak{so}(2n)$ denote the vector space (actually, Lie algebra) of $2n \times 2n$ real skew-symmetric matrices. It is well-known that every matrix $A \in \mathfrak{so}(2n)$ can be written as

$$A = PDP^\top,$$

where $P$ is an orthogonal matrix and where $D$ is a block diagonal matrix

$$D = \begin{pmatrix} D_1 & & \\ & D_2 & \\ & & \ddots \\ & & & D_n \end{pmatrix},$$

consisting of $2 \times 2$ blocks of the form

$$D_i = \begin{pmatrix} 0 & -a_i \\ a_i & 0 \end{pmatrix}.$$

For a proof, see Horn and Johnson [94] (Corollary 2.5.14), Gantmacher [74] (Chapter IX), or Gallier [72] (Chapter 11).

Since $\det(D_i) = a_i^2$ and $\det(A) = \det(PDP^\top) = \det(D) = \det(D_1) \cdots \det(D_n)$, we get

$$\det(A) = (a_1 \cdots a_n)^2.$$

The Pfaffian is a polynomial function $\text{Pf}(A)$ in skew-symmetric $2n \times 2n$ matrices $A$ (a polynomial in $(2n-1)n$ variables) such that

$$\text{Pf}(A)^2 = \det(A),$$

and for every arbitrary matrix $B$,

$$\text{Pf}(BAB^\top) = \text{Pf}(A) \det(B).$$

The Pfaffian shows up in the definition of the Euler class of a vector bundle. There is a simple way to define the Pfaffian using some exterior algebra. Let $(e_1, \ldots, e_{2n})$ be any basis of $\mathbb{R}^{2n}$. For any matrix $A \in \mathfrak{so}(2n)$, let

$$\omega(A) = \sum_{i<j} a_{ij} e_i \wedge e_j,$$

where $A = (a_{ij})$. Then, $\bigwedge^n \omega(A)$ is of the form $Ce_1 \wedge e_2 \wedge \cdots \wedge e_{2n}$ for some constant $C \in \mathbb{R}$. 


**Definition 22.3.** For every skew symmetric matrix \( A \in \mathfrak{so}(2n) \), the Pfaffian polynomial or Pfaffian, is the degree \( n \) polynomial \( \text{Pf}(A) \) defined by
\[
\bigwedge^n \omega(A) = n! \text{Pf}(A) e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}.
\]

Clearly, \( \text{Pf}(A) \) is independent of the basis chosen. If \( A \) is the block diagonal matrix \( D \), a simple calculation shows that
\[
\omega(D) = -(a_1 e_1 \wedge e_2 + a_2 e_3 \wedge e_4 + \cdots + a_{n} e_{2n-1} \wedge e_{2n})
\]
and that
\[
\bigwedge^n \omega(D) = (-1)^n n! a_1 \cdots a_n e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}.
\]
and so
\[
\text{Pf}(D) = (-1)^n a_1 \cdots a_n.
\]
Since \( \text{Pf}(D)^2 = (a_1 \cdots a_n)^2 = \det(A) \), we seem to be on the right track.

**Proposition 22.30.** For every skew symmetric matrix \( A \in \mathfrak{so}(2n) \) and every arbitrary matrix \( B \), we have:

(i) \( \text{Pf}(A)^2 = \det(A) \)

(ii) \( \text{Pf}(BAB^\top) = \text{Pf}(A) \det(B) \).

**Proof.** If we assume that (ii) is proved then, since we can write \( A = PDP^\top \) for some orthogonal matrix \( P \) and some block diagonal matrix \( D \) as above, as \( \det(P) = \pm 1 \) and \( \text{Pf}(D)^2 = \det(A) \), we get
\[
\text{Pf}(A)^2 = \text{Pf}(PDP^\top)^2 = \text{Pf}(D)^2 \det(P)^2 = \det(A),
\]
which is (i). Therefore, it remains to prove (ii).

Let \( f_i = Be_i \) for \( i = 1, \ldots, 2n \), where \((e_1, \ldots, e_{2n})\) is any basis of \( \mathbb{R}^{2n} \). Since \( f_i = \sum_k b_{ki} e_k \), we have
\[
\tau = \sum_{i,j} a_{ij} f_i \wedge f_j = \sum_{i,j} \sum_{k,l} b_{ki} a_{ij} b_{lj} e_k \wedge e_l = \sum_{k,l} (BAB^\top)_{kl} e_k \wedge e_l,
\]
and so, as \( BAB^\top \) is skew symmetric and \( e_k \wedge e_l = -e_l \wedge e_k \), we get
\[
\tau = 2\omega(BAB^\top).
\]
Consequently,
\[
\bigwedge^n \tau = 2^n \bigwedge^n \omega(BAB^\top) = 2^n n! \text{Pf}(BAB^\top) e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}.
\]
Now,
\[ \bigwedge^n \tau = C f_1 \wedge f_2 \wedge \cdots \wedge f_{2n}, \]
for some \( C \in \mathbb{R} \). If \( B \) is singular, then the \( f_i \) are linearly dependent, which implies that \( f_1 \wedge f_2 \wedge \cdots \wedge f_{2n} = 0 \), in which case
\[ \text{Pf}(BAB^\top) = 0, \]
as \( e_1 \wedge e_2 \wedge \cdots \wedge e_{2n} \neq 0 \). Therefore, if \( B \) is singular, \( \det(B) = 0 \) and
\[ \text{Pf}(BAB^\top) = 0 = \text{Pf}(A) \det(B). \]
If \( B \) is invertible, as \( \tau = \sum_{i,j} a_{ij} f_i \wedge f_j = 2 \sum_{i<j} a_{ij} f_i \wedge f_j \), we have
\[ \bigwedge^n \tau = 2^n n! \text{Pf}(A) f_1 \wedge f_2 \wedge \cdots \wedge f_{2n}. \]
However, as \( f_i = B e_i \), we have
\[ f_1 \wedge f_2 \wedge \cdots \wedge f_{2n} = \det(B) e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}, \]
so
\[ \bigwedge^n \tau = 2^n n! \text{Pf}(B) \text{det}(B) e_1 \wedge e_2 \wedge \cdots \wedge e_{2n} \]
and as
\[ \bigwedge^n \tau = 2^n n! \text{Pf}(BAB^\top) e_1 \wedge e_2 \wedge \cdots \wedge e_{2n}, \]
we get
\[ \text{Pf}(BAB^\top) = \text{Pf}(A) \det(B), \]
as claimed. \( \square \)

**Remark:** It can be shown that the polynomial \( \text{Pf}(A) \) is the unique polynomial with integer coefficients such that \( \text{Pf}(A)^2 = \det(A) \) and \( \text{Pf}(\text{diag}(S, \ldots, S)) = +1 \), where
\[ S = \left( \begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array} \right); \]
see Milnor and Stasheff [129] (Appendix C, Lemma 9). There is also an explicit formula for \( \text{Pf}(A) \), namely:
\[ \text{Pf}(A) = \frac{1}{2^n n!} \sum_{\sigma \in \mathfrak{S}_{2n}} \text{sgn}(\sigma) \prod_{i=1}^n a_{\sigma(2i-1)} \sigma(2i). \]
For example, if
\[ A = \begin{pmatrix} 0 & -a \\ a & 0 \end{pmatrix}, \]
then \( \text{Pf}(A) = -a \), and if
\[
A = \begin{pmatrix}
0 & a_1 & a_2 & a_3 \\
-a_1 & 0 & a_4 & a_5 \\
-a_2 & -a_4 & 0 & a_6 \\
-a_3 & -a_5 & -a_6 & 0
\end{pmatrix},
\]
then
\[
\text{Pf}(A) = a_1a_6 - a_2a_5 + a_4a_3.
\]

It is easily checked that
\[
\text{det}(A) = (\text{Pf}(A))^2 = (a_1a_6 - a_2a_5 + a_4a_3)^2.
\]

\[\text{W}\]

Beware, some authors use a different sign convention and require the Pfaffian to have the value +1 on the matrix \( \text{diag}(S', \ldots, S') \), where
\[
S' = \begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix}.
\]

For example, if \( \mathbb{R}^{2n} \) is equipped with an inner product \( \langle - , - \rangle \), then some authors define \( \omega(A) \) as
\[
\omega(A) = \sum_{i<j} \langle Ae_i, e_j \rangle e_i \wedge e_j,
\]
where \( A = (a_{ij}) \). But then, \( \langle Ae_i, e_j \rangle = a_{ji} \) and not \( a_{ij} \), and this Pfaffian takes the value +1 on the matrix \( \text{diag}(S', \ldots, S') \). This version of the Pfaffian differs from our version by the factor \( (-1)^n \). In this respect, Madsen and Tornehave [119] seem to have an incorrect sign in Proposition B6 of Appendix C.

We will also need another property of Pfaffians. Recall that the ring \( M_n(\mathbb{C}) \) of \( n \times n \) matrices over \( \mathbb{C} \) is embedded in the ring \( M_{2n}(\mathbb{R}) \) of \( 2n \times 2n \) matrices with real coefficients, using the injective homomorphism that maps every entry \( z = a + ib \in \mathbb{C} \) to the \( 2 \times 2 \) matrix
\[
\begin{pmatrix}
a & -b \\
b & a
\end{pmatrix}.
\]

If \( A \in M_n(\mathbb{C}) \), let \( A_\mathbb{R} \in M_{2n}(\mathbb{R}) \) denote the real matrix obtained by the above process. Observe that every skew Hermitian matrix \( A \in \mathfrak{u}(n) \) \( (i.e., \) with \( A^* = -A^\top = -A \) yields a matrix \( A_\mathbb{R} \in \mathfrak{so}(2n) \).

**Proposition 22.31.** For every skew Hermitian matrix \( A \in \mathfrak{u}(n) \), we have
\[
\text{Pf}(A_\mathbb{R}) = i^n \text{det}(A).
\]
Proof. It is well-known that a skew Hermitian matrix can be diagonalized with respect to a unitary matrix $U$ and that the eigenvalues are pure imaginary or zero, so we can write

$$A = U \mathrm{diag}(ia_1, \ldots, ia_n)U^*,$$

for some reals $a_j \in \mathbb{R}$. Consequently, we get

$$A_{\mathbb{R}} = U_{\mathbb{R}} \mathrm{diag}(D_1, \ldots, D_n)U_{\mathbb{R}}^T,$$

where

$$D_j = \begin{pmatrix} 0 & -a_j \\ a_j & 0 \end{pmatrix}$$

and

$$\mathrm{Pf}(A_{\mathbb{R}}) = \mathrm{Pf}(\mathrm{diag}(D_1, \ldots, D_n)) = (-1)^n a_1 \cdots a_n,$$

as we saw before. On the other hand,

$$\det(A) = \det(\mathrm{diag}(ia_1, \ldots, ia_n)) = i^n a_1 \cdots a_n,$$

and as $(-1)^n = i^n i^n$, we get

$$\mathrm{Pf}(A_{\mathbb{R}}) = i^n \det(A),$$

as claimed. □

Madsen and Tornehave [119] state Proposition 22.31 using the factor $(-i)^n$, which is wrong.
Chapter 23

Differential Forms

23.1 Differential Forms on Subsets of $\mathbb{R}^n$ and de Rham Cohomology

The theory of differential forms is one of the main tools in geometry and topology. This theory has a surprisingly large range of applications, and it also provides a relatively easy access to more advanced theories such as cohomology. For all these reasons, it is really an indispensable theory, and anyone with more than a passible interest in geometry should be familiar with it.

The theory of differential forms was initiated by Poincaré and further elaborated by Elie Cartan at the end of the nineteenth century. Differential forms have two main roles:

1. Describe various systems of partial differential equations on manifolds.

2. To define various geometric invariants reflecting the global structure of manifolds or bundles. Such invariants are obtained by integrating certain differential forms.

As we will see shortly, as soon as one tries to define integration on higher-dimensional objects, such as manifolds, one realizes that it is not functions that are integrated, but instead differential forms. Furthermore, as by magic, the algebra of differential forms handles changes of variables automatically and yields a neat form of “Stokes formula.”

Our goal is to define differential forms on manifolds, but we begin with differential forms on open subsets of $\mathbb{R}^n$ in order to build up intuition.

Differential forms are smooth functions on open subsets $U$ of $\mathbb{R}^n$, taking as values alternating tensors in some exterior power $\bigwedge^p(\mathbb{R}^n)^*$. Recall from Sections 22.4 and 22.5, in particular Proposition 22.12, that for every finite-dimensional vector space $E$, the isomorphisms $\mu: \bigwedge^n(E^*) \rightarrow \text{Alt}^n(E; \mathbb{R})$ induced by the linear extensions of the maps given by

$$
\mu(v_1^* \wedge \cdots \wedge v_n^*)(u_1, \ldots, u_n) = \begin{vmatrix}
    v_1^*(u_1) & \cdots & v_1^*(u_n) \\
     \vdots & \ddots & \vdots \\
    v_n^*(u_1) & \cdots & v_n^*(u_n)
\end{vmatrix} = \det(v_j^*(u_i))
$$

yield a canonical isomorphism of algebras \( \mu : \bigwedge(E^*) \to \text{Alt}(E) \), where

\[
\text{Alt}(E) = \bigoplus_{n \geq 0} \text{Alt}^n(E; \mathbb{R}),
\]

and where \( \text{Alt}^n(E; \mathbb{R}) \) is the vector space of alternating multilinear maps on \( \mathbb{R}^n \). Recall that multiplication on alternating multilinear forms is defined such that, for \( f \in \text{Alt}^m(E; K) \) and \( g \in \text{Alt}^n(E; K) \),

\[
(f \wedge g)(u_1, \ldots, u_{m+n}) = \sum_{\sigma \in \text{shuffle}(m,n)} \text{sgn}(\sigma) f(u_{\sigma(1)}, \ldots, u_{\sigma(m)})g(u_{\sigma(m+1)}, \ldots, u_{\sigma(m+n)}),
\]

where \( \text{shuffle}(m, n) \) consists of all \((m, n)\) “shuffles;” that is, permutations \( \sigma \) of \( \{1, \ldots, m+n\} \) such that \( \sigma(1) < \cdots < \sigma(m) \) and \( \sigma(m+1) < \cdots < \sigma(m+n) \). The isomorphism \( \mu \) has the property that

\[
\mu(\omega \wedge \eta) = \mu(\omega) \wedge \mu(\eta), \quad \omega, \eta \in \bigwedge(E^*),
\]

where the wedge operation on the left is the wedge on the exterior algebra \( \bigwedge(E^*) \), and the wedge on the right is the multiplication on \( \text{Alt}(E) \) defined in (**)..

In view of these isomorphisms, we will identify \( \omega \) and \( \mu(\omega) \) for any \( \omega \in \bigwedge^n(E^*) \), and we will write \( \omega(u_1, \ldots, u_n) \) as an abbreviation for \( \mu(\omega)(u_1, \ldots, u_n) \).

Because \( \text{Alt}(\mathbb{R}^n) \) is an algebra under the wedge product, differential forms also have a wedge product. However, the power of differential forms stems from the exterior differential \( d \), which is a skew-symmetric version of the usual differentiation operator.

**Definition 23.1.** Given any open subset \( U \) of \( \mathbb{R}^n \), a smooth differential \( p \)-form on \( U \), for short a \( p \)-form on \( U \), is any smooth function \( \omega : U \to \bigwedge^p(\mathbb{R}^n)^* \). The vector space of all \( p \)-forms on \( U \) is denoted \( \mathcal{A}^p(U) \). The vector space \( \mathcal{A}^*(U) = \bigoplus_{p \geq 0} \mathcal{A}^p(U) \) is the set of differential forms on \( U \).

Observe that \( \mathcal{A}^0(U) = C^\infty(U, \mathbb{R}) \), the vector space of smooth functions on \( U \), and \( \mathcal{A}^1(U) = C^\infty(U, (\mathbb{R}^n)^*) \), the set of smooth functions from \( U \) to the set of linear forms on \( \mathbb{R}^n \). Also, \( \mathcal{A}^p(U) = \{0\} \) for \( p > n \).

**Remark:** The space \( \mathcal{A}^*(U) \) is also denoted \( \mathcal{A}^*(U) \). Other authors use \( \Omega^p(U) \) instead of \( \mathcal{A}^p(U) \), but we prefer to reserve \( \Omega^p \) for holomorphic forms.

Recall from Section 22.2 that if \((e_1, \ldots, e_n)\) is any basis of \( \mathbb{R}^n \) and \((e_1^*, \ldots, e_n^*)\) is its dual basis, then the alternating tensors

\[
e_i^* = e_i^* \wedge \cdots \wedge e_p^*
\]

form basis of \( \bigwedge^p(\mathbb{R}^n)^* \), where \( I = \{i_1, \ldots, i_p\} \subseteq \{1, \ldots, n\} \), with \( i_1 < \cdots < i_p \). Thus, with respect to the basis \((e_1, \ldots, e_n)\), every \( p \)-form \( \omega \) can be uniquely written

\[
\omega(x) = \sum_I f_I(x) e_{i_1}^* \wedge \cdots \wedge e_{i_p}^* = \sum_I f_I(x) e_I^* \quad x \in U,
\]
where each $f_I$ is a smooth function on $U$. For example, if $U = \mathbb{R}^2 - \{0\}$, then

$$\omega(x, y) = \frac{-y}{x^2 + y^2} e_1^* + \frac{x}{x^2 + y^2} e_2^*$$

is a 1-form on $U$ (with $e_1 = (1, 0)$ and $e_2 = (0, 1)$).

We often write $\omega_x$ instead of $\omega(x)$. Now, not only is $\mathcal{A}^*(U)$ a vector space, it is also an algebra.

**Definition 23.2.** The wedge product on $\mathcal{A}^*(U)$ is defined as follows: For all $p, q \geq 0$, the wedge product $\wedge: \mathcal{A}^p(U) \times \mathcal{A}^q(U) \to \mathcal{A}^{p+q}(U)$ is given by

$$(\omega \wedge \eta)_x = \omega_x \wedge \eta_x, \quad x \in U.$$

For example, if $\omega$ and $\eta$ are one-forms, then

$$(\omega \wedge \eta)_x(u, v) = \omega_x(u)\eta_x(v) - \omega_x(v)\eta_x(u).$$

In particular, if $U \subseteq \mathbb{R}^3$ and $\omega_x = a_1 e_1^* + a_3 e_3^*$ and $\eta_x = b_1 e_1^* + b_2 e_2^*$, for $u = (u_1, u_2, u_3) \in \mathbb{R}^3$ and $v = (v_1, v_2, v_3) \in \mathbb{R}^3$, the preceding line implies

$$\omega_x(u)\eta_x(v) - \omega_x(v)\eta_x(u)$$

= $(a_1 b_2 e_1^* \wedge e_2^* - a_3 b_1 e_1^* \wedge e_3^* - a_3 b_2 e_2^* \wedge e_3^*)(u, v)$

= $(a_1 b_2 e_1^* \wedge e_2^* + a_3 b_1 e_1^* \wedge e_2^* + a_3 b_2 e_2^* \wedge e_3^*)(u, v)$

= $((a_1 e_1^* + a_3 e_3^*)(u) \wedge (b_1 e_1^* + b_2 e_2^*))(u, v)$

= $(\omega \wedge \eta)_x(u, v)$,

since $e_1^* \wedge e_2^* = 0$ and $e_i^* \wedge e_j^* = -e_j^* \wedge e_i^*$ for all $1 \leq i < j \leq 3$.

For $f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R})$ and $\omega \in \mathcal{A}^p(U)$, we have $f \wedge \omega = f \omega$. Thus, the algebra $\mathcal{A}^*(U)$ is also a $C^\infty(U, \mathbb{R})$-module,

**Proposition 22.10** immediately yields

**Proposition 23.1.** For all forms $\omega \in \mathcal{A}^p(U)$ and $\eta \in \mathcal{A}^q(U)$, we have

$$\eta \wedge \omega = (-1)^{pq}\omega \wedge \eta.$$
We now come to the crucial operation of exterior differentiation. First recall that if $f: U \to V$ is a smooth function from $U \subseteq \mathbb{R}^n$ to a (finite-dimensional) normed vector space $V$, the derivative $f': U \to \text{Hom}(\mathbb{R}^n, V)$ of $f$ (also denoted $Df$) is a function with domain $U$, with $f'(x)$ a linear map in $\text{Hom}(\mathbb{R}^n, V)$ for every $x \in U$, such that if $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$, $(u_1, \ldots, u_m)$ is a basis of $V$, and if $f(x) = f_1(x)u_1 + \cdots + f_m(x)u_m$, then

$$f'(x)(y_1 e_1 + \cdots + y_n e_n) = \sum_{i=1}^m \left( \sum_{j=1}^n \frac{\partial f_i}{\partial x_j}(x) y_j \right) u_i.$$ 

The $m \times n$ matrix

$$\left( \frac{\partial f_i}{\partial x_j}(x) \right)$$

is the Jacobian matrix of $f$ at $x$, and if we write

$$z_1 u_1 + \cdots + z_m u_m = f'(x)(y_1 e_1 + \cdots + y_n e_n),$$

then in matrix form, we have

$$\begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$ 

We also write $f'_x(u)$ for $f'(x)(u)$. Observe that since a $p$-form is a smooth map $\omega: U \to \wedge^p(\mathbb{R}^n)^*$, its derivative is a map

$$\omega': U \to \text{Hom}\left( \mathbb{R}^n, \bigwedge^p(\mathbb{R}^n)^* \right)$$

such that $\omega'_x$ is a linear map from $\mathbb{R}^n$ to $\wedge^p(\mathbb{R}^n)^*$ for every $x \in U$. By the isomorphism $\wedge^p(\mathbb{R}^n)^* \cong \text{Alt}^p(\mathbb{R}^n; \mathbb{R})$, we can view $\omega'_x$ as a linear map $\omega'_x: \mathbb{R}^n \to \text{Alt}^p(\mathbb{R}^n; \mathbb{R})$, or equivalently as a multilinear form $\omega'_x: (\mathbb{R}^n)^{p+1} \to \mathbb{R}$ which is alternating in its last $p$ arguments. The exterior derivative $(d\omega)_x$ is obtained by making $\omega'_x$ into an alternating map in all of its $p + 1$ arguments.

To make things more concrete, let us pick a basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, so that the $n \choose p$ tensors $e_I^*$ form a basis of $\wedge^p(\mathbb{R}^n)^*$, where $I$ is any subset $I = \{i_1, \ldots, i_p\} \subseteq \{1, \ldots, n\}$ such that $i_1 < \cdots < i_p$. Then every $p$-form $\omega$ can be uniquely written as

$$\omega_x = \sum_I f_I(x) e_I^*, \quad x \in U,$$

where each $f_I$ is a smooth function on $U$, and for any $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$,

$$\omega'_x(v) = \sum_I f'_I(x)(v) e_I^* = \sum_I \sum_{j=1}^n \frac{\partial f_I}{\partial x_j}(x) v_j e_I^* = \sum_I \langle \text{grad}(f_I)_x \cdot v \rangle e_I^*.$$
23.1. DIFFERENTIAL FORMS ON $\mathbb{R}^N$ AND DE RHAM COHOMOLOGY

where $\cdot$ is the standard Euclidean inner product.

**Remark:** Observe that $\omega'_x$ is given by the $\binom{n}{p} \times n$ Jacobian matrix

$$\begin{pmatrix}
\frac{\partial f_I}{\partial x_1}(x) \\
\vdots \\
\frac{\partial f_I}{\partial x_n}(x)
\end{pmatrix}
$$

and that the product of the $I$th row of the above matrix by $v$

$$\begin{pmatrix}
\frac{\partial f_I}{\partial x_1}(x) & \cdots & \frac{\partial f_I}{\partial x_n}(x)
\end{pmatrix}
\begin{pmatrix}
v_1 \\
\vdots \\
v_n
\end{pmatrix}
$$

gives the coefficient $\text{grad}(f_I)_x \cdot v$ of $e^*_I$.

**Definition 23.3.** For every $p \geq 0$, the exterior differential $d: \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$ is given by

$$(d\omega)_x(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1} \omega'_x(u_i)(u_1, \ldots, \hat{u}_i, \ldots, u_{p+1}),$$

for all $\omega \in \mathcal{A}^p(U)$, all $x \in U$, and all $u_1, \ldots, u_{p+1} \in \mathbb{R}^n$, where the hat over the argument $u_i$ means that it should be omitted.

In terms of a basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, if $\omega_x = \sum_I f_I(x) e^*_I$, then

$$(d\omega)_x(u_1, \ldots, u_{p+1}) = \sum_I f'_I(x) (u_i) e^*_I(u_1, \ldots, \hat{u}_i, \ldots, u_{p+1})$$

$$= \sum_I \text{grad}(f_I)_x \cdot u_i e^*_I(u_1, \ldots, \hat{u}_i, \ldots, u_{p+1}).$$

One should check that $(d\omega)_x$ is indeed alternating, but this is easy. If necessary to avoid confusion, we write $d^p: \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$ instead of $d: \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$.

**Remark:** Definition 23.3 is the definition adopted by Cartan [36, 37] and Madsen and Tornehave [119]. Some authors use a different approach often using Propositions 23.2 and 23.3 as a starting point, but we find the approach using Definition 23.3 more direct. Furthermore, this approach extends immediately to the case of vector-valued forms.

For any smooth function, $f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R})$, we get

$$df_x(u) = f'_x(u).$$

---

1We warn the reader that a few typos have crept up in the English translation, Cartan [37], of the original version Cartan [36].
Therefore, for smooth functions, the exterior differential \( df \) coincides with the usual derivative \( f' \) (we identify \( \Lambda^1(\mathbb{R}^n)^* \) and \( (\mathbb{R}^n)^* \)). For any 1-form \( \omega \in \mathcal{A}^1(U) \), we have

\[
d\omega_x(u, v) = \omega_x'(u)(v) - \omega_x'(v)(u).
\]

It follows that the map

\[
(u, v) \mapsto \omega_x'(u)(v)
\]

is symmetric iff \( d\omega = 0 \).

For a concrete example of exterior differentiation, consider

\[
\omega_{(x, y)} = \frac{-y}{x^2 + y^2} e_1^* + \frac{x}{x^2 + y^2} e_2^* = f_1(x, y)e_1^* + f_2(x, y)e_2^*.
\]

Since

\[
\text{grad}(f_1)^\top (x, y) = \left( \frac{2xy}{(x^2 + y^2)^2} \frac{y^2 - x^2}{(x^2 + y^2)^2} \right)
\]

\[
\text{grad}(f_2)^\top (x, y) = \left( \frac{y^2 - x^2}{(x^2 + y^2)^2} \frac{-2xy}{(x^2 + y^2)^2} \right)
\]

if we write \( u_1 = \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} \) and \( u_2 = \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix} \), then we have

\[
\omega'_{(x, y)}(u_1)(u_2) = (\text{grad}(f_1)_{(x, y)} \cdot u_1)u_1^*(u_2) + (\text{grad}(f_2)_{(x, y)} \cdot u_1)e_2^*(u_2)
\]

\[
= \left( \frac{2xy}{(x^2 + y^2)^2} \frac{y^2 - x^2}{(x^2 + y^2)^2} \right) \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} e_1^* \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}
\]

\[
+ \left( \frac{y^2 - x^2}{(x^2 + y^2)^2} \frac{-2xy}{(x^2 + y^2)^2} \right) \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix} e_2^* \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}
\]

\[
= \frac{2xy(u_{11}u_{21} - u_{12}u_{22}) + (y^2 - x^2)(u_{12}u_{21} + u_{11}u_{22})}{(x^2 + y^2)^2}.
\]

A similar computation shows that

\[
\omega'_{(x, y)}(u_2)(u_1) = \frac{2xy(u_{11}u_{21} - u_{12}u_{22}) + (y^2 - x^2)(u_{12}u_{21} + u_{11}u_{22})}{(x^2 + y^2)^2}
\]

\[
= \omega'_{(x, y)}(u_1)(u_2),
\]

and so

\[
d\omega_{(x, y)}(u_1, u_2) = \omega'_{(x, y)}(u_1)(u_2) - \omega'_{(x, y)}(u_2)(u_1) = 0.
\]

Therefore \( d\omega_{(x, y)} = 0 \) for all \( (x, y) \in U \), that is, \( d\omega = 0 \).

The following observation is quite trivial but it will simplify notation: On \( \mathbb{R}^n \), we have the projection function \( pr_i : \mathbb{R}^n \to \mathbb{R} \) with \( pr_i(u_1, \ldots, u_n) = u_i \). Note that \( pr_i = e_i^* \), where
(e_1, \ldots, e_n) is the canonical basis of \( \mathbb{R}^n \). Let \( x_i : U \to \mathbb{R} \) be the restriction of \( pr_i \) to \( U \). Then, note that \( x'_i \) is the constant map given by

\[ x'_i(x) = pr_i, \quad x \in U. \]

It follows that \( dx_i = x'_i \) is the constant function with value \( pr_i = e_i^* \). Now, since every \( p \)-form \( \omega \) can be uniquely expressed as

\[ \omega_x = \sum_I f_I(x) e^*_i \wedge \cdots \wedge e^*_p = \sum_I f_I(x) e^*_I, \quad x \in U, \]

using Definition 23.2, we see immediately that \( \omega \) can be uniquely written in the form

\[ \omega = \sum_I f_I(x) \, dx_i \wedge \cdots \wedge dx_p, \quad (\ast) \]

where the \( f_I \) are smooth functions on \( U \).

Observe that for \( f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R}) \), we have

\[ df_x = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x) e^*_i \quad \text{and} \quad df = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \, dx_i. \]

**Proposition 23.2.** For every \( p \)-form \( \omega \in \mathcal{A}^p(U) \) with \( \omega = f \, dx_i \wedge \cdots \wedge dx_p \), we have

\[ d\omega = df \wedge dx_i \wedge \cdots \wedge dx_p. \]

**Proof.** Recall that \( \omega_x = f e^*_i \wedge \cdots \wedge e^*_p = f e^*_I \), so

\[ \omega'_x(u) = f'_x(u) e^*_I = df_x(u) e^*_I, \]

and by Definition 23.3, we get

\[ d\omega_x(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1} df_x(u_i) e^*_I(u_1, \ldots, \hat{u}_i, \ldots, u_{p+1}) = (df_x \wedge e^*_I)(u_1, \ldots, u_{p+1}), \]

where the last equation is an instance of the equation stated just before Proposition 22.12.

In practice we use Proposition 23.2 to compute \( d\omega \). For example, if we take the previous example of

\[ \omega = \frac{-y}{x^2 + y^2} \, dx + \frac{x}{x^2 + y^2} \, dy, \]
Proposition 23.2 implies that
\[ d\omega = d\left(\frac{-y}{x^2 + y^2}\right) \land dx + d\left(\frac{x}{x^2 + y^2}\right) \land dy \]
\[ = \left(\frac{2xy}{(x^2 + y^2)^2} dx + \frac{y^2 - x^2}{(x^2 + y^2)^2} dy\right) \land dx + \left(\frac{y^2 - x^2}{(x^2 + y^2)^2} dx - \frac{2xy}{(x^2 + y^2)^2} dy\right) \land dy \]
\[ = \frac{y^2 - x^2}{(x^2 + y^2)^2} dy \land dx + \frac{y^2 - x^2}{(x^2 + y^2)^2} dx \land dy = 0. \]

We can now prove

**Proposition 23.3.** For all \( \omega \in \mathcal{A}^p(U) \) and all \( \eta \in \mathcal{A}^q(U) \),
\[ d(\omega \land \eta) = d\omega \land \eta + (-1)^q \omega \land d\eta. \]

**Proof.** In view of the unique representation \((\ast)\), it is enough to prove the proposition when \( \omega = f e_i^* \) and \( \eta = g e_j^* \). In this case, as \( \omega \land \eta = f g e_i^* \land e_j^* \), by Proposition 23.2, we have
\[ d(\omega \land \eta) = d(fg) \land e_i^* \land e_j^* \]
\[ = ((df)g + f(df)) \land e_i^* \land e_j^* \]
\[ = (df)g \land e_i^* \land e_j^* + f(df) \land e_i^* \land e_j^* \]
\[ = df \land e_i^* \land ge_j^* + (-1)^pg e_i^* \land dg \land e_j^* \]
\[ = d\omega \land \eta + (-1)^p \omega \land d\eta \]
since by Proposition 23.2, \( d\omega = df \land e_i^* \) and \( d\eta = g \land e_j^* \).

We say that \( d \) is an *anti-derivation of degree* \(-1\). Finally, here is the crucial and almost magical property of \( d \):

**Proposition 23.4.** For every \( p \geq 0 \), the composition \( \mathcal{A}^p(U) \xrightarrow{d} \mathcal{A}^{p+1}(U) \xrightarrow{d} \mathcal{A}^{p+2}(U) \) is identically zero; that is
\[ d \circ d = 0, \]
which is an abbreviation for \( d^{p+1} \circ d^p = 0 \).

**Proof.** It is enough to prove the proposition when \( \omega = f e_i^* \). We have
\[ d\omega_x = df_x \land e_i^* = \frac{\partial f}{\partial x_1}(x) e_1^* \land e_i^* + \cdots + \frac{\partial f}{\partial x_n}(x) e_n^* \land e_i^*. \]
As \( e_i^* \land e_j^* = -e_j^* \land e_i^* \) and \( e_i^* \land e_i^* = 0 \), we get
\[ (d \circ d)\omega = \sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(x) e_i^* \land e_j^* \land e_i^* \]
\[ = \sum_{i<j} \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(x) - \frac{\partial^2 f}{\partial x_j \partial x_i}(x) \right) e_i^* \land e_j^* \land e_i^* = 0, \]
since partial derivatives commute (as \( f \) is smooth).

\[ \Box \]
Propositions 23.2, 23.3 and 23.4 can be summarized by saying that $\mathcal{A}^*(U)$ together with the product $\wedge$ and the differential $d$ is a differential graded algebra. As $\mathcal{A}^*(U) = \bigoplus_{p \geq 0} \mathcal{A}^p(U)$ and $d^p: \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$, we can view $d = (d^p)$ as a linear map $d: \mathcal{A}^*(U) \to \mathcal{A}^*(U)$ such that

$$d \circ d = 0.$$  

**Definition 23.4.** The diagram

$$\mathcal{A}^0(U) \xrightarrow{d} \mathcal{A}^1(U) \xrightarrow{} \cdots \xrightarrow{} \mathcal{A}^{p-1}(U) \xrightarrow{d} \mathcal{A}^p(U) \xrightarrow{d} \mathcal{A}^{p+1}(U) \xrightarrow{} \cdots$$

is called the de Rham complex of $U$. It is a cochain complex.

Let us consider one more example. Assume $n = 3$ and consider any function $f \in \mathcal{A}^0(U)$. We have

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz,$$

and the vector

$$\left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$$

is the gradient of $f$. Next let

$$\omega = Pdx + Qdy + Rdz$$

be a 1-form on some open $U \subseteq \mathbb{R}^3$. An easy calculation yields

$$d\omega = dP \wedge dx + dQ \wedge dy + dR \wedge dz$$

$$= \left( \frac{\partial P}{\partial y} dx + \frac{\partial P}{\partial z} dy + \frac{\partial P}{\partial z} dz \right) \wedge dx + \left( \frac{\partial Q}{\partial x} dx + \frac{\partial Q}{\partial y} dy + \frac{\partial Q}{\partial z} dz \right) \wedge dy + \left( \frac{\partial R}{\partial x} dx + \frac{\partial R}{\partial y} dy + \frac{\partial R}{\partial z} dz \right) \wedge dz$$

$$= \left( \frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z} \right) dy \wedge dz + \left( \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x} \right) dz \wedge dx + \left( \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dx \wedge dy.$$  

The vector field given by

$$\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \quad \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \quad \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}$$

is the curl of the vector field given by $(P,Q,R)$. Now, if

$$\eta = Ady \wedge dz + Bdz \wedge dx + Cdx \wedge dy$$
is a 2-form on \( \mathbb{R}^3 \), we get

\[
\begin{align*}
\,d\eta &= dA \wedge dy \wedge dz + dB \wedge dz \wedge dx + dC \wedge dx \wedge dy \\
&= \left( \frac{\partial A}{\partial x} \, dx + \frac{\partial A}{\partial y} \, dy + \frac{\partial A}{\partial z} \, dz \right) \wedge dy \wedge dz \\
&\quad + \left( \frac{\partial B}{\partial x} \, dx + \frac{\partial B}{\partial y} \, dy + \frac{\partial B}{\partial z} \, dz \right) \wedge dz \wedge dx \\
&\quad + \left( \frac{\partial C}{\partial x} \, dx + \frac{\partial C}{\partial y} \, dy + \frac{\partial C}{\partial z} \, dz \right) \wedge dx \wedge dy \\
&= \frac{\partial A}{\partial x} \, dx \wedge dy \wedge dz + \frac{\partial B}{\partial y} \, dy \wedge dz \wedge dx + \frac{\partial C}{\partial z} \, dz \wedge dx \wedge dy \\
&\quad = \left( \frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z} \right) \, dx \wedge dy \wedge dz.
\end{align*}
\]

The real number \( \frac{\partial A}{\partial x} + \frac{\partial B}{\partial y} + \frac{\partial C}{\partial z} \) is called the divergence of the vector field \((A, B, C)\).

When is there a smooth field \((P, Q, R)\) whose curl is given by a prescribed smooth field \((A, B, C)\)? Equivalently, when is there a 1-form \(\omega = P \, dx + Q \, dy + R \, dz\) such that

\[d\omega = \eta = Ady \wedge dz + Bdz \wedge dx + Cdx \wedge dy?\]

By Proposition 23.4 it is necessary that \(d\eta = 0\); that is, \((A, B, C)\) has zero divergence. However, this condition is not sufficient in general; it depends on the topology of \(U\). If \(U\) is star-like, Poincaré’s Lemma (to be considered shortly) says that this condition is sufficient.

**Definition 23.5.** A differential form \(\omega\) is closed iff \(d\omega = 0\); exact iff \(\omega = d\eta\) for some differential form \(\eta\). For every \(p \geq 0\), let

\[Z^p(U) = \{ \omega \in \mathcal{A}^p(U) \mid d\omega = 0 \} = \text{Ker} \, d : \mathcal{A}^p(U) \rightarrow \mathcal{A}^{p+1}(U)\]

be the vector space of closed \(p\)-forms, also called \(p\)-cocycles, and for every \(p \geq 1\), let

\[B^p(U) = \{ \omega \in \mathcal{A}^p(U) \mid \exists \eta \in \mathcal{A}^{p-1}(U), \omega = d\eta \} = \text{Im} \, d : \mathcal{A}^{p-1}(U) \rightarrow \mathcal{A}^p(U)\]

be the vector space of exact \(p\)-forms, also called \(p\)-coboundaries. Set \(B^0(U) = (0)\). Forms in \(\mathcal{A}^p(U)\) are also called \(p\)-cochains. As \(B^p(U) \subseteq Z^p(U)\) (by Proposition 23.4), for every \(p \geq 0\), we define the \(p\)th de Rham cohomology group of \(U\) as the quotient space

\[H^p_{\text{DR}}(U) = Z^p(U)/B^p(U);\]

This is an abelian group under addition of cosets. An element of \(H^p_{\text{DR}}(U)\) is called a cohomology class and is denoted \([\omega]\), where \(\omega \in Z^p(U)\) is a cocycle. The real vector space \(H^*_{\text{DR}}(U) = \bigoplus_{p \geq 0} H^p_{\text{DR}}(U)\) is called the de Rham cohomology algebra of \(U\).
23.1. DIFFERENTIAL FORMS ON $\mathbb{R}^n$ AND DE RHAM COHOMOLOGY

We often drop the subscript $\text{DR}$ and write $H^p(U)$ for $H^p_{\text{DR}}(U)$ (resp. $H^\bullet(U)$ for $H^\bullet_{\text{DR}}(U)$), when no confusion arises. Proposition 23.4 shows that every exact form is closed, but the converse is false in general. Measuring the extent to which closed forms are not exact is the object of de Rham cohomology.

For example, if we consider the form
\[ \omega(x,y) = -\frac{y}{x^2+y^2} \, dx + \frac{x}{x^2+y^2} \, dy, \]
on $U = \mathbb{R}^2 - \{0\}$, we have $d\omega = 0$. Yet, it is not hard to show (using integration, see Madsen and Tornehave [119], Chapter 1) that there is no smooth function $f$ on $U$ such that $df = \omega$. Thus, $\omega$ is a closed form which is not exact. This is because $U$ is punctured.

Observe that $H^0(U) = Z^0(U) = \{ f \in C^\infty(U, \mathbb{R}) \mid df = 0 \}$; that is, $H^0(U)$ is the space of locally constant functions on $U$, equivalently, the space of functions that are constant on the connected components of $U$. Thus, the cardinality of $H^0(U)$ gives the number of connected components of $U$. For a large class of open sets (for example, open sets that can be covered by finitely many convex sets), the cohomology groups $H^p(U)$ are finite dimensional.

Going back to Definition 23.5, we define the vector spaces $Z^*(U)$ and $B^*(U)$ by
\[ Z^*(U) = \bigoplus_{p \geq 0} Z^p(U) \quad \text{and} \quad B^*(U) = \bigoplus_{p \geq 0} B^p(U). \]
Now, $\mathcal{A}^*(U)$ is a graded algebra with multiplication $\wedge$. Observe that $Z^*(U)$ is a subalgebra of $\mathcal{A}^*(U)$, since
\[ d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^p \omega \wedge d\eta, \]
so $d\omega = 0$ and $d\eta = 0$ implies $d(\omega \wedge \eta) = 0$. Furthermore, $B^*(U)$ is an ideal in $Z^*(U)$, because if $\omega = d\eta$ and $d\tau = 0$, then
\[ d(\eta \tau) = d\eta \wedge \tau + (-1)^{p-1} \eta \wedge d\tau = \omega \wedge \tau, \]
with $\eta \in \mathcal{A}^{p-1}(U)$. Therefore, $H^\bullet_{\text{DR}} = Z^*(U)/B^*(U)$ inherits a graded algebra structure from $\mathcal{A}^*(U)$. Explicitly, the multiplication in $H^\bullet_{\text{DR}}$ is given by
\[ [\omega] [\eta] = [\omega \wedge \eta]. \]

It turns out that Propositions 23.3 and 23.4 together with the fact that $d$ coincides with the derivative on $\mathcal{A}^0(U)$ characterize the differential $d$.

**Theorem 23.5.** There is a unique linear map $d: \mathcal{A}^*(U) \to \mathcal{A}^*(U)$ with $d = (d^p)$ and $d^p: \mathcal{A}^p(U) \to \mathcal{A}^{p+1}(U)$ for every $p \geq 0$, such that

1. $df = f'$, for every $f \in \mathcal{A}^0(U) = C^\infty(U, \mathbb{R})$,
2. $d \circ d = 0$. 

(3) For every \( \omega \in \mathcal{A}^p(U) \) and every \( \eta \in \mathcal{A}^q(U) \),
\[
d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^p \omega \wedge d\eta.
\]

Proof. Existence has already been shown, so we only have to prove uniqueness. Let \( \delta \) be another linear map satisfying (1)–(3). By (1), \( df = \delta f = f' \) if \( f \in \mathcal{A}^0(U) \). In particular, this holds when \( f = x_i \), with \( x_i : U \rightarrow \mathbb{R} \) the restriction of \( pr_i \) to \( U \). In this case, we know that \( \delta x_i = e_i^* \), the constant function \( e_i^* = pr_i \). By (2), \( \delta e_i^* = 0 \). Using (3), we get \( \delta e_i^* = 0 \) for every nonempty subset \( I \subseteq \{1, \ldots, n\} \). If \( \omega = fe_i^* \), by (3), we get
\[
\delta \omega = \delta f \wedge e_i^* + f \wedge \delta e_i^* = \delta f \wedge e_i^* = df \wedge e_i^* = d\omega.
\]

Finally, since every differential form is a linear combination of special forms \( f_ie_i^* \), we conclude that \( \delta = d \).

We now consider the action of smooth maps \( \varphi : U \rightarrow U' \) on differential forms in \( \mathcal{A}^*(U') \). We will see that \( \varphi \) induces a map from \( \mathcal{A}^*(U') \) to \( \mathcal{A}^*(U) \) called a pull-back map. This corresponds to a change of variables.

Recall Proposition 22.9 which states that if \( f : E \rightarrow F \) is any linear map between two finite-dimensional vector spaces \( E \) and \( F \), then
\[
\mu\left(\bigwedge^p f^\top(\omega)\right)(u_1, \ldots, u_p) = \mu(\omega)(f(u_1), \ldots, f(u_p)), \quad \omega \in \bigwedge^p F^*, \ u_1, \ldots, u_p \in E.
\]

We apply this proposition with \( E = \mathbb{R}^n \), \( F = \mathbb{R}^m \), and \( f = \varphi_x \ (x \in U) \), and get
\[
\mu\left(\bigwedge^p (\varphi_x')^\top(\omega_{\varphi(x)})\right)(u_1, \ldots, u_p) = \mu(\omega_{\varphi(x)})(\varphi'_x(u_1), \ldots, \varphi'_x(u_p)), \quad \omega \in \mathcal{A}^p(V), \ u_i \in \mathbb{R}^n.
\]

This gives us the behavior of \( \bigwedge^p (\varphi_x')^\top \) under the identification of \( \bigwedge^p (\mathbb{R})^* \) and \( \text{Alt}^n(\mathbb{R}^n; \mathbb{R}) \) via the isomorphism \( \mu \). Consequently, denoting \( \bigwedge^p (\varphi_x')^\top \) by \( \varphi^* \), we make the following definition:

**Definition 23.6.** Let \( U \subseteq \mathbb{R}^n \) and \( V \subseteq \mathbb{R}^m \) be two open subsets. For every smooth map \( \varphi : U \rightarrow V \), for every \( p \geq 0 \), we define the map \( \varphi^* : \mathcal{A}^p(V) \rightarrow \mathcal{A}^p(U) \) by
\[
\varphi^*(\omega)(u_1, \ldots, u_p) = \omega_{\varphi(x)}(\varphi'_x(u_1), \ldots, \varphi'_x(u_p)),
\]
for all \( \omega \in \mathcal{A}^p(V) \), all \( x \in U \), and all \( u_1, \ldots, u_p \in \mathbb{R}^n \). We say that \( \varphi^*(\omega) \) (for short, \( \varphi^* \omega \)) is the pull-back of \( \omega \) by \( \varphi \).

As \( \varphi \) is smooth, \( \varphi^* \omega \) is a smooth \( p \)-form on \( U \). The maps \( \varphi^* : \mathcal{A}^p(V) \rightarrow \mathcal{A}^p(U) \) induce a map also denoted \( \varphi^* : \mathcal{A}^*(V) \rightarrow \mathcal{A}^*(U) \). Using the chain rule we check immediately that
\[
id^* = \text{id},
(\psi \circ \varphi)^* = \varphi^* \circ \psi^*.
\]
Here is an example of Definition 23.6. Let $U = [0,1] \times [0,1] \subset \mathbb{R}^2$ and let $V = \mathbb{R}^3$. Define $\varphi : Q \to \mathbb{R}^3$ as $\varphi(u,v) = (\varphi_1(u,v),\varphi_2(u,v),\varphi_3(u,v)) = (x,y,z)$ where

$$x = u + v, \quad y = u - v, \quad z = uv.$$  

Let $w = xdy \wedge dz + ydx \wedge dz$ be a 2-form in $V$. Clearly

$$\varphi'(u,v) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ v & u \end{pmatrix}.$$  

Set $u_1 = \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}$ and $u_2 = \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}$. Definition 23.6 implies that the pull back of $\omega$ into $U$ is

$$\varphi^*(\omega)(u_1,u_2) = \omega_{\varphi(u,v)}(\varphi'(u,v)(u_1),\varphi'(u,v)(u_2))$$

$$= \omega_{\varphi(u,v)} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ v & u \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ v & u \end{pmatrix} \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}$$

$$= \omega_{\varphi(u,v)} \begin{pmatrix} u_{11} + u_{12} \\ u_{11} - u_{12} \\ vu_{11} + uu_{12} \end{pmatrix}, \begin{pmatrix} u_{21} + u_{22} \\ u_{21} - u_{22} \\ vu_{21} + uu_{22} \end{pmatrix}$$

$$= (u+v)dy \wedge dz \begin{pmatrix} u_{11} + u_{12} \\ u_{11} - u_{12} \\ vu_{11} + uu_{12} \end{pmatrix}, \begin{pmatrix} u_{21} + u_{22} \\ u_{21} - u_{22} \\ vu_{21} + uu_{22} \end{pmatrix}$$

$$+ (u-v)dx \wedge dz \begin{pmatrix} u_{11} + u_{12} \\ u_{11} - u_{12} \\ vu_{11} + uu_{12} \end{pmatrix}, \begin{pmatrix} u_{21} + u_{22} \\ u_{21} - u_{22} \\ vu_{21} + uu_{22} \end{pmatrix}$$

$$= (u+v) \begin{vmatrix} u_{11} - u_{12} & u_{21} - u_{22} \\ vu_{11} + uu_{12} & vu_{21} + uu_{22} \end{vmatrix} + (u-v) \begin{vmatrix} u_{11} + u_{12} & u_{21} + u_{22} \\ vu_{11} + uu_{12} & vu_{21} + uu_{22} \end{vmatrix}$$

$$= (u+v)(u+v)(u_{11}u_{22} - u_{21}u_{12}) + (u-v)(u-v)(u_{11}u_{22} - u_{21}u_{12})$$

$$= (u+v)(u+v)du \wedge dv(u_1,u_2) + (u-v)(u-v)du \wedge dv(u_1,u_2)$$

$$= 2(u^2 + v^2)du \wedge dv(u_1,u_2).$$

As the preceding example demonstrates, Definition 23.6 is not convenient for computations, so it is desirable to derive rules that yield a recursive definition of the pull-back.

The first rule has to do with the constant form $\omega = e_i^*$. We claim that

$$\varphi^*e_i^* = d\varphi_i, \quad \text{with } \varphi_i = pr_i \circ \varphi.$$
Proposition 23.6. Let \( \varphi = (\varphi_1)_x e_1 + \cdots + (\varphi_m)_x e_m \) for all \( x \in U \), \( \varphi'(u) = (\varphi_1)_x(u) e_1 + \cdots + (\varphi_m)_x(u) e_m \), and

\[
\varphi^*(e_i^*)_x(u) = \sum_{l=1}^n \frac{\partial \varphi_i}{\partial x_l}(x) u_l = \sum_{l=1}^n \frac{\partial \varphi_i}{\partial x_l}(x) e_l^*(u),
\]

so

\[
\varphi^*(\omega \wedge \eta) = \varphi^* \omega \wedge \varphi^* \eta, \quad \text{for all } \omega \in \mathcal{A}^p(V) \text{ and all } \eta \in \mathcal{A}^q(V).
\]

(i) \( \varphi^*(f) = f \circ \varphi \), for all \( f \in \mathcal{A}^0(V) \).

(ii) \( \varphi^*(\omega \wedge \eta) = \varphi^*(d \omega) \), for all \( \omega \in \mathcal{A}^p(V) \); that is, the following diagram commutes for all \( p \geq 0 \):

\[
\begin{array}{ccc}
\mathcal{A}^p(V) & \xrightarrow{\varphi^*} & \mathcal{A}^p(U) \\
\downarrow{d} & & \downarrow{d} \\
\mathcal{A}^{p+1}(V) & \xrightarrow{\varphi^*} & \mathcal{A}^{p+1}(U).
\end{array}
\]

Proof. (i) (See Madsen and Tornehave [119], Chapter 3). For any \( x \in U \) and any vectors \( u_1, \ldots, u_{p+q} \in \mathbb{R}^n \) (with \( p, q \geq 1 \)), we have

\[
\varphi^*(\omega \wedge \eta)_x(u_1, \ldots, u_{p+q}) = (\omega \wedge \eta)_{\varphi(x)}(\varphi'_x(u_1), \ldots, \varphi'_x(u_{p+q}))
\]

\[
= \sum_{\sigma \in \text{shuffle}(p,q)} \text{sgn}(\sigma) \omega_{\varphi(x)}(\varphi'_x(u_{\sigma(1)}), \ldots, \varphi'_x(u_{\sigma(p)}))
\]

\[
\eta_{\varphi(x)}(\varphi'_x(u_{\sigma(p+1)}), \ldots, \varphi'_x(u_{\sigma(p+q)}))
\]

\[
= \sum_{\sigma \in \text{shuffle}(p,q)} \text{sgn}(\sigma) \varphi^*(\omega)_x(u_{\sigma(1)}, \ldots, u_{\sigma(p)})
\]

\[
\varphi^*(\eta)_x(u_{\sigma(p+1)}, \ldots, u_{\sigma(p+q)})
\]

\[
= (\varphi^*(\omega)_x \wedge \varphi^*(\eta)_x)(u_1, \ldots, u_{p+q}).
\]
If \( p = 0 \) or \( q = 0 \), the proof is similar but simpler. We leave it as exercise to the reader.

(ii) If \( f \in \mathcal{A}^0(V) = C^\infty(V) \), by definition \( \varphi^*(f)_x = f(\varphi(x)) \), which means that \( \varphi^*(f) = f \circ \varphi \).

First we prove (iii) in the case \( \omega \in \mathcal{A}^0(V) \). Using (i) and (ii) and the fact that \( \varphi^* e^*_i = d\varphi_i \), since

\[
df = \sum_{k=1}^m \frac{\partial f}{\partial x_k} e^*_k,
\]

we have

\[
\varphi^*(df) = \sum_{k=1}^m \varphi^* \left( \frac{\partial f}{\partial x_k} \right) \wedge \varphi^*(e^*_k)
\]

\[
= \sum_{k=1}^m \left( \frac{\partial f}{\partial x_k} \circ \varphi \right) \wedge \left( \sum_{l=1}^n \frac{\partial \varphi_k}{\partial x_l} e^*_l \right)
\]

\[
= \sum_{k=1}^m \sum_{l=1}^n \left( \frac{\partial f}{\partial x_k} \circ \varphi \right) \left( \frac{\partial \varphi_k}{\partial x_l} \right) e^*_l
\]

\[
= \sum_{l=1}^n \left( \sum_{k=1}^m \left( \frac{\partial f}{\partial x_k} \circ \varphi \right) \frac{\partial \varphi_k}{\partial x_l} \right) e^*_l
\]

\[
= \sum_{l=1}^n \frac{\partial (f \circ \varphi)}{\partial x_l} e^*_l
\]

\[
= df \circ \varphi
\]

\[
df = d(\varphi^*(f)).
\]

For the case where \( \omega = fe^*_i \), we know that \( d\omega = df \wedge e^*_i \). We claim that

\[d\varphi^*(e^*_i) = 0.\]

To prove this first we show by induction on \( p \) that

\[d\varphi^*(e^*_i) = d(\varphi^*(e^*_i) \wedge \cdots \wedge e^*_p)) = d(\varphi^*(e^*_i) \wedge \cdots \wedge \varphi^*(e^*_p))\]

\[= \sum_{k=1}^p (-1)^{k-1} \varphi^*(e^*_i) \wedge \cdots \wedge d(\varphi^*(e^*_i)) \wedge \cdots \wedge \varphi^*(e^*_p).\]

The base case \( p = 1 \) is trivial. Assuming that the induction hypothesis holds for any \( p \geq 1 \),
Proposition 23.6 implies that $I \subset \{i_1 < i_2 < \cdots < i_{p+1}\}$, using Proposition 23.3, we have

\[
\frac{d\phi^*}{d\varphi^*} (e_i^*) = d(\phi^* (e_i^*) \wedge \varphi^*(e_{i_2}^*) \wedge \cdots \wedge \varphi^*(e_{i_{p+1}}^*))
= d(\phi^* (e_i^*)) \wedge \varphi^*(e_{i_2}^*) \wedge \cdots \wedge \varphi^*(e_{i_{p+1}}^*)
- (-1)^{i+1} \varphi^*(e_i^*) \wedge d(\phi^*(e_{i_2}^*) \wedge \cdots \wedge \varphi^*(e_{i_{p+1}}^*))
= d(\phi^*(e_i^*)) \wedge \varphi^*(e_{i_2}^*) \wedge \cdots \wedge \varphi^*(e_{i_{p+1}}^*)
\]

establishing the induction hypothesis.

As a consequence of the above equation, we have

\[
d\phi^*(e_i^*) = d(\phi^*(e_i^*) \wedge \cdots \wedge \varphi^*(e_{i_p}^*))
= \sum (-1)^{i-1} \varphi^*(e_i^*) \wedge \cdots \wedge d(\phi^*(e_{i_k}^*) \wedge \cdots \wedge \varphi^*(e_{i_{p+1}}^*)) = 0,
\]

since $\varphi^*(e_i^*) = d\varphi_{i_k}$ and $d \circ d = 0$. Consequently,

\[
d(\phi^*(f) \wedge \varphi^*(e_i^*)) = d(\phi^*(f)) \wedge \varphi^*(e_i^*).
\]

Then we have

\[
\phi^*(d\omega) = \phi^*(df) \wedge \phi^*(e_i^*) = d(\phi^*(f)) \wedge \phi^*(e_i^*) = d(\phi^*(f)) \wedge \phi^*(e_i^*) = d(\phi^*(f e_i^*)) = d(\phi^* \omega).
\]

Since every differential form is a linear combination of special forms $f e_i^*$, we are done. \qed

We use Proposition 23.6 to recompute the pull-back of $w = x dy \wedge dz + y dx \wedge dz$. Recall $Q = [0, 1] \times [0, 1] \subset \mathbb{R}^2$ and $\varphi : U \rightarrow \mathbb{R}^3$ was defined via

\[
x = u + v, \quad y = u - v, \quad z = uv.
\]

Proposition 23.6 implies that

\[
\phi^*(\omega) = (u + v)\phi^*(dy) \wedge \phi^*(dz) + (u - v)\phi^*(dx) \wedge \phi^*(dz)
= (u + v)d(\phi^*y) \wedge d(\phi^*z) + (u - v)d(\phi^*x) \wedge d(\phi^*y)
= (u + v)(du - dv) \wedge (vdu + udv) + (u - v)(du + dv) \wedge (vdu + udv)
= 2(u^2 + v^2) du \wedge dv.
\]
We may generalize the techniques of the preceding calculation by using Proposition 23.6 to compute $\varphi^*\omega$ where $\varphi: U \to V$ is a smooth map between two open subsets $U$ and $V$ of $\mathbb{R}^n$ and $\omega = f dy_1 \wedge \cdots \wedge dy_n$ is a $p$-form on $V$. We can write $\varphi = (\varphi_1, \ldots, \varphi_n)$ with $\varphi_i: U \to \mathbb{R}$. By Proposition 23.6, we have

$$\varphi^*\omega = \varphi^*(f) \varphi^*(dy_1) \wedge \cdots \wedge \varphi^*(dy_n)$$

$$= \varphi^*(f)(\varphi^*y_1) \wedge \cdots \wedge \varphi^*(y_n)$$

$$= (f \circ \varphi)(\varphi^*y_1) \wedge \cdots \wedge d(\varphi^*y_n).$$

However, $\varphi^*y_i = \varphi_i$ so we have

$$\varphi^*\omega = (f \circ \varphi)d\varphi_1 \wedge \cdots \wedge d\varphi_n.$$

For any $x \in U$, since

$$d(\varphi_i)_x = \sum_{j=1}^n \frac{\partial \varphi_i}{\partial x_j}(x) \, dx_j$$

we get

$$d\varphi_1 \wedge \cdots \wedge d\varphi_n = \det \left( \frac{\partial \varphi_i}{\partial x_j}(x) \right) \, dx_1 \wedge \cdots \wedge dx_n = J(\varphi)_x \, dx_1 \wedge \cdots \wedge dx_n$$

where

$$J(\varphi)_x = \det \left( \frac{\partial \varphi_i}{\partial x_j}(x) \right)$$

is the Jacobian of $\varphi$ at $x \in U$. It follows that

$$(\varphi^*\omega)_x = \varphi^*(f \, dy_1 \wedge \cdots \wedge dy_n)_x = f(\varphi(x)) \, J(\varphi)_x \, dx_1 \wedge \cdots \wedge dx_n.$$

The fact that $d$ and pull-back commutes is an important fact: It allows us to show that a map $\varphi: U \to V$ induces a map $H^*(\varphi): H^*(V) \to H^*(U)$ on cohomology, and it is crucial in generalizing the exterior differential to manifolds.

To a smooth map $\varphi: U \to V$, we associate the map $H^p(\varphi): H^p(V) \to H^p(U)$ given by

$$H^p(\varphi)([\omega]) = [\varphi^*(\omega)].$$

This map is well defined, because if we pick any representative $\omega + d\eta$ in the cohomology class $[\omega]$ specified by the closed form $\omega$, then

$$d\varphi^*\omega = \varphi^*d\omega = 0,$$

so $\varphi^*\omega$ is closed, and

$$\varphi^*(\omega + d\eta) = \varphi^*\omega + \varphi^*(d\eta) = \varphi^*\omega + d\varphi^*\eta,$$
which shows that $H^p(\varphi)([\omega])$ is well defined. It is also clear that

$$H^{p+q}(\varphi)([\omega][\eta]) = H^p(\varphi)([\omega])H^q(\varphi)([\eta]),$$

which means that $H^\cdot(\varphi)$ is a homomorphism of graded algebras. We often denote $H^\cdot(\varphi)$ by $\varphi^\ast$.

We conclude this section by stating without proof an important result known as the Poincaré Lemma. Recall that a subset $S \subseteq \mathbb{R}^n$ is star-shaped iff there is some point $c \in S$ such that for every point $x \in S$, the closed line segment $[c,x]$ joining $c$ and $x$ is entirely contained in $S$.

**Theorem 23.7. (Poincaré’s Lemma)** If $U \subseteq \mathbb{R}^n$ is any star-shaped open set, then we have $H^p(U) = (0)$ for $p > 0$ and $H^0(U) = \mathbb{R}$. Thus, for every $p \geq 1$, every closed form $\omega \in A^p(U)$ is exact.

**Sketch of proof.** Pick $c$ so that $U$ is star-shaped w.r.t. $c$ and let $g: U \to U$ be the constant function with value $c$. Then, we see that

$$g^\ast \omega = \begin{cases} 0 & \text{if } \omega \in A^p(U), \text{ with } p \geq 1, \\ \omega(c) & \text{if } \omega \in A^0(U), \end{cases}$$

where $\omega(c)$ denotes the constant function with value $\omega(c)$. The trick is to find a family of linear maps $h^p: A^p(U) \to A^{p-1}(U)$, for $p \geq 1$, with $h^0 = 0$, such that

$$d \circ h^p + h^{p+1} \circ d = \text{id} - g^\ast, \quad p > 0,$$

called a *chain homotopy*. Indeed, if $\omega \in A^p(U)$ is closed and $p \geq 1$, we get $dh^p \omega = \omega$, so $\omega$ is exact, and if $p = 0$ we get $h^1d\omega = 0 = \omega - \omega(c)$, so $\omega$ is constant. It remains to find the $h^p$, which is not obvious. A construction of these maps can be found in Madsen and Tornehave [119] (Chapter 3), Warner [175] (Chapter 4), Cartan [37] (Section 2) Morita [133] (Chapter 3).

\[\square\]

In Section 23.2, we promote differential forms to manifolds. As preparation, note that every open subset $U \subseteq \mathbb{R}^n$ is a manifold, and that for every $x \in U$, the tangent space $T_xU$ to $U$ at $x$ is canonically isomorphic to $\mathbb{R}^n$. It follows that the tangent bundle $TU$ and the cotangent bundle $T^\ast U$ are trivial, namely $TU \cong U \times \mathbb{R}^n$ and $T^\ast U \cong U \times (\mathbb{R}^n)^\ast$, so the bundle

$$\bigwedge^k T^\ast U \cong U \times \bigwedge^k (\mathbb{R}^n)^\ast$$

is also trivial. Consequently, we can view $A^k(U)$ as the set of smooth sections of the vector bundle $\bigwedge^k T^\ast(U)$. The generalization to manifolds is then to define the space of differential $p$-forms on a manifold $M$ as the space of smooth sections of the bundle $\bigwedge^k T^\ast M$. 

23.2 Differential Forms on Manifolds

Let $M$ be any smooth manifold of dimension $n$. We define the vector bundle $\bigwedge T^*M$ as the direct sum bundle

$$\bigwedge T^*M = \bigoplus_{k=0}^{n} \bigwedge^k T^*M;$$

see Section 28.3 for details.

**Definition 23.7.** Let $M$ be any smooth manifold of dimension $n$. The set $\mathcal{A}^k(M)$ of smooth differential $k$-forms on $M$ is the set of smooth sections $\Gamma(M, \bigwedge^k T^*M)$ of the bundle $\bigwedge^k T^*M$, and the set $\mathcal{A}^*(M)$ of all smooth differential forms on $M$ is the set of smooth sections $\Gamma(M, \bigwedge T^*M)$ of the bundle $\bigwedge T^*M$.

Recall that a smooth section of the bundle $\bigwedge^k T^*M$ is a smooth function $\omega: M \to \bigwedge^k T^*M$ such that $\omega(p) \in \bigwedge^k T^*_p M$ for all $p \in M$.

Observe that $\mathcal{A}^0(M) \cong \mathcal{C}^\infty(M, \mathbb{R})$, the set of smooth functions on $M$, since the bundle $\bigwedge^0 T^*M$ is isomorphic to $M \times \mathbb{R}$, and smooth sections of $M \times \mathbb{R}$ are just graphs of smooth functions on $M$. We also write $\mathcal{C}^\infty(M)$ for $\mathcal{C}^\infty(M, \mathbb{R})$. If $\omega \in \mathcal{A}^*(M)$, we often write $\omega_p$ for $\omega(p)$.

Definition 23.7 is quite abstract, and it is important to get a more down-to-earth feeling by taking a local view of differential forms, namely with respect to a chart. So, let $(U, \varphi)$ be a local chart on $M$, with $\varphi: U \to \mathbb{R}^n$, and let $x_i = pr_i \circ \varphi$, the $i$th local coordinate $(1 \leq i \leq n)$ (see Section 7.2). Recall that by Proposition 7.7, for any $p \in U$, the vectors

$$\left( \frac{\partial}{\partial x_1} \right)_p, \ldots, \left( \frac{\partial}{\partial x_n} \right)_p$$

form a basis of the tangent space $T_p M$. Furthermore, by Proposition 7.13 and the discussion following Proposition 7.12, the linear forms $(dx_1)_p, \ldots, (dx_n)_p$ form a basis of $T^*_p M$, (where $(dx_i)_p$, the differential of $x_i$ at $p$, is identified with the linear form such that $df_p(v) = v(f)$, for every smooth function $f$ on $U$ and every $v \in T_p M$). Consequently, locally on $U$, every $k$-form $\omega \in \mathcal{A}^k(M)$ can be written uniquely as

$$\omega = \sum_I f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k} = \sum_I f_I dx_I, \quad p \in U,$$

where $I = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, with $i_1 < \ldots < i_k$ and $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_k}$. Furthermore, each $f_I$ is a smooth function on $U$.

**Remark:** We define the set of smooth $(r, s)$-tensor fields as the set $\Gamma(M, T^{r,s}(M))$ of smooth sections of the tensor bundle $T^{r,s}(M) = T^{\otimes r} M \otimes (T^* M)^{\otimes s}$. Then, locally in a chart $(U, \varphi)$, every tensor field $\omega \in \Gamma(M, T^{r,s}(M))$ can be written uniquely as

$$\omega = \sum_{I, J} f_{I,J}^{i_1, \ldots, i_r} \left( \frac{\partial}{\partial x_{i_1}} \right) \otimes \cdots \otimes \left( \frac{\partial}{\partial x_{i_r}} \right) \otimes dx_{j_1} \otimes \cdots \otimes dx_{j_s},$$
The operations on the algebra $\bigwedge T^*M$ yield operations on differential forms using pointwise definitions. If $\omega, \eta \in \mathcal{A}^r(M)$ and $\lambda \in \mathbb{R}$, then for every $x \in M$,

$$(\omega + \eta)_x = \omega_x + \eta_x$$

$$(\lambda \omega)_x = \lambda \omega_x$$

$$(\omega \wedge \eta)_x = \omega_x \wedge \eta_x.$$ 

Actually, it is necessary to check that the resulting forms are smooth, but this is easily done using charts. When $f \in \mathcal{A}^0(M)$, we write $f\omega$ instead of $f \wedge \omega$. It follows that $\mathcal{A}^*(M)$ is a graded real algebra and a $C^\infty(M)$-module.

Proposition 23.1 generalizes immediately to manifolds.

**Proposition 23.8.** For all forms $\omega \in \mathcal{A}^r(M)$ and $\eta \in \mathcal{A}^s(M)$, we have

$$\eta \wedge \omega = (-1)^{pq} \omega \wedge \eta.$$ 

For any smooth map $\varphi: M \to N$ between two manifolds $M$ and $N$, we have the differential map $d\varphi: TM \to TN$, also a smooth map, and for every $p \in M$, the map $d\varphi_p: T_p M \to T_{\varphi(p)} N$ is linear. As in Section 23.1, Proposition 22.9 gives us the formula

$$\mu\left(\left(\bigwedge^k (d\varphi_p)^\top\right) (\omega_{\varphi(p)})(u_1, \ldots, u_k) = \mu(\omega_{\varphi(p)})(d\varphi_p(u_1), \ldots, d\varphi_p(u_k)), \quad \omega \in \mathcal{A}^k(N),$$ 

for all $u_1, \ldots, u_k \in T_pM$. This gives us the behavior of $\bigwedge^k (d\varphi_p)^\top$ under the identification of $\bigwedge^k T^*_p M$ and $\text{Alt}^k(T_pM; \mathbb{R})$ via the isomorphism $\mu$. Here is the extension of Definition 23.6 to differential forms on a manifold.

**Definition 23.8.** For any smooth map $\varphi: M \to N$ between two smooth manifolds $M$ and $N$, for every $k \geq 0$, we define the map $\varphi^*: \mathcal{A}^k(N) \to \mathcal{A}^k(M)$ by

$$\varphi^* (\omega)_p(u_1, \ldots, u_k) = \omega_{\varphi(p)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_k)),$$

for all $\omega \in \mathcal{A}^k(N)$, all $p \in M$, and all $u_1, \ldots, u_k \in T_pM$. We say that $\varphi^* (\omega)$ (for short, $\varphi^* \omega$) is the pull-back of $\omega$ by $\varphi$.

The maps $\varphi^*: \mathcal{A}^k(N) \to \mathcal{A}^k(M)$ induce a map also denoted $\varphi^*: \mathcal{A}^*(N) \to \mathcal{A}^*(M)$. Using the chain rule, we check immediately that

$$\text{id}^* = \text{id},$$

$$(\psi \circ \varphi)^* = \varphi^* \circ \psi^*.$$ 

We need to check that $\varphi^* \omega$ is smooth, and for this it is enough to check it locally on a chart $(U, \varphi)$. On $U$ we know that $\omega \in \mathcal{A}^k(M)$ can be written uniquely as

$$\omega = \sum_f f_i dx_{i_1} \wedge \cdots \wedge dx_{i_k}, \quad p \in U,$$
with \( f_i \) smooth and it is easy to see (using the definition) that

\[
\varphi^* \omega = \sum_i (f_i \circ \varphi) d(x_i \circ \varphi) \wedge \cdots \wedge d(x_k \circ \varphi), \quad (*)
\]

which is smooth.

Line \((*)\) is what we use to efficiently calculate the pull-back of \( \omega \) on \( M \). For example, let \( M = \{(\theta, \varphi) : 0 < \theta < \pi, 0 < \varphi < 2\pi\} \subset \mathbb{R}^2 \), \( N = S^2 \) and \( \psi : M \to N \) the parametrization of \( S^2 \) given by

\[
x = \sin \theta \cos \varphi, \quad y = \sin \theta \sin \varphi, \quad z = \cos \theta.
\]

See Figure 23.1.

![Figure 23.1: The spherical coordinates of \( S^2 \)](image)

Let \( w = x \, dy \) be a form on \( S^2 \). The pull-back of \( \omega \) into \( M \) is calculated via \((*)\) as

\[
\psi^* w = \sin \theta \cos \varphi d(\sin \theta \sin \varphi) \\
= \sin \theta \cos \varphi (\cos \theta \sin \varphi \, d\theta + \sin \theta \cos \varphi \, d\varphi),
\]

where we applied Proposition 23.2 since \( M \subset \mathbb{R}^2 \).

**Remark:** The fact that the pull-back of differential forms makes sense for arbitrary smooth maps \( \varphi: M \to N \), and not just diffeomorphisms, is a major technical superiority of forms over vector fields.

The next step is to define \( d \) on \( \mathcal{A}^*(M) \). There are several ways to proceed, but since we already considered the special case where \( M \) is an open subset of \( \mathbb{R}^n \), we proceed using charts.
Given a smooth manifold $M$ of dimension $n$, let $(U, \varphi)$ be any chart on $M$. For any $\omega \in \mathcal{A}^k(M)$ and any $p \in U$, define $(d\omega)_p$ as follows: If $k = 0$, that is $\omega \in C^\infty(M)$, let
\[
(d\omega)_p = d\omega_p,
\]
the differential of $\omega$ at $p$, and if $k \geq 1$, let
\[
(d\omega)_p = \varphi^*(d((\varphi^{-1})^*\omega)_{\varphi(p)})_p,
\]
where $d$ is the exterior differential on $\mathcal{A}^k(\varphi(U))$. More explicitly, $(d\omega)_p$ is given by
\[
(d\omega)_p(u_1, \ldots, u_{k+1}) = d((\varphi^{-1})^*\omega)_{\varphi(p)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_{k+1})),
\]
for every $p \in U$ and all $u_1, \ldots, u_{k+1} \in T_pM$. Observe that the above formula is still valid when $k = 0$ if we interpret the symbol $d$ in $d((\varphi^{-1})^*\omega)_{\varphi(p)} = d(\omega \circ \varphi^{-1})_{\varphi(p)}$ as the differential.

Since $\varphi^{-1} : \varphi(U) \to U$ is map whose domain is an open subset $W = \varphi(U)$ of $\mathbb{R}^n$, the form $(\varphi^{-1})^*\omega$ is a differential form in $\mathcal{A}^*(W)$, so $d((\varphi^{-1})^*\omega)$ is well-defined.

The formula at line (**) encapsulates the following “natural” three step procedure:

Step 1: Take the form $\omega$ on the manifold $M$ and precompose $\omega$ with the parameterization $\varphi^{-1}$ so that $(\varphi^{-1})^*\omega$ is now a form in $U$, a subset of $\mathbb{R}^m$, where $m$ is the dimension of $M$.

Step 2: Differentiate $(\varphi^{-1})^*\omega$ via Proposition 23.2.

Step 3: Compose the result of Step 2 with the chart map $\varphi$ and pull the differentiate form on $U$ back into $M$.

We need to check that the definition at line (**) does not depend on the chart $(U, \varphi)$.

Proof. For any other chart $(V, \psi)$, with $U \cap V \neq \emptyset$, the map $\theta = \psi \circ \varphi^{-1}$ is a diffeomorphism between the two open subsets $\varphi(U \cap V)$ and $\psi(U \cap V)$, and $\theta = \psi \circ \varphi$. Let $x = \varphi(p)$ and $y = \psi(p)$. We need to check that
\[
d((\varphi^{-1})^*\omega)_x(d\varphi_p(u_1), \ldots, d\varphi_p(u_{k+1})) = d((\psi^{-1})^*\omega)_y(d\psi_p(u_1), \ldots, d\psi_p(u_{k+1})),
\]
for every $p \in U \cap V$ and all $u_1, \ldots, u_{k+1} \in T_pM$. However, $y = \psi(p) = \theta(\varphi(p)) = \theta(x)$, so
\[
d((\psi^{-1})^*\omega)_y(d\psi_p(u_1), \ldots, d\psi_p(u_{k+1})) = d((\varphi^{-1} \circ \theta^{-1})^*\omega)_{\theta(x)}(d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})),
\]
and since
\[
(\varphi^{-1} \circ \theta^{-1})^* = (\theta^{-1})^* \circ (\varphi^{-1})^*
\]
and, by Proposition 23.6 (iii),
\[
d(((\theta^{-1})^* \circ (\varphi^{-1})^*)\omega) = d((\theta^{-1})^*((\varphi^{-1})^*\omega)) = (\theta^{-1})^*(d((\varphi^{-1})^*\omega)),
\]
we get
\[ \begin{align*}
d((\varphi^{-1} \circ \theta^{-1})^* \omega)_{\theta(x)} & (d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})) \\
& = (\theta^{-1})^*(d((\varphi^{-1})^* \omega))_{\theta(x)} (d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})),
\end{align*} \]
and then
\[ \begin{align*}
(\theta^{-1})^*(d((\varphi^{-1})^* \omega))_{\theta(x)} & (d(\theta \circ \varphi)_p(u_1), \ldots, d(\theta \circ \varphi)_p(u_{k+1})) \\
& = d((\varphi^{-1})^* \omega)_{\theta(x)} ((d\theta^{-1})_{\theta(x)}(d(\theta \circ \varphi)_p(u_1)), \ldots, (d\theta^{-1})_{\theta(x)}(d(\theta \circ \varphi)_p(u_{k+1}))).
\end{align*} \]
As \((d\theta^{-1})_{\theta(x)}(d(\theta \circ \varphi)_p(u_i)) = d(\theta^{-1} \circ (\theta \circ \varphi))_p(u_i) = d\varphi_p(u_i)\), by the chain rule, we obtain
\[ d((\psi^{-1})^* \omega)_{\theta(x)}(d\psi_p(u_1), \ldots, d\psi_p(u_{k+1})) = d((\varphi^{-1})^* \omega)_{\theta(x)}(d\varphi_p(u_1), \ldots, d\varphi_p(u_{k+1})), \]
as desired. \qed

Observe that \((d\omega)_p\) is smooth on \(U\), and as our definition of \((d\omega)_p\) does not depend on the choice of a chart, the forms \((d\omega) \upharpoonright U\) agree on overlaps and yield a differential form \(d\omega\) defined on the whole of \(M\). Thus we can make the following definition:

**Definition 23.9.** If \(M\) is any smooth manifold, there is a linear map \(d: A^k(M) \to A^{k+1}(M)\) for every \(k \geq 0\), such that for every \(\omega \in A^k(M)\), for every chart \((U, \varphi)\), for every \(p \in U\), if \(k = 0\), that is \(\omega \in C^\infty(M)\), then
\[ (d\omega)_p = d\omega_p, \quad \text{the differential of } \omega \text{ at } p, \]
else if \(k \geq 1\), then
\[ (d\omega)_p = \varphi^* (d((\varphi^{-1})^* \omega)_{\varphi(p)})_p, \]
where \(d\) is the exterior differential on \(A^k(\varphi(U))\) from Definition 23.3. We obtain a linear map \(d: A^*(M) \to A^*(M)\) called **exterior differentiation**.

To explicitly demonstrate Definition (23.9), we return to our previous example of \(\psi : M \to \mathbb{S}^2\) and \(\omega = xdy\) considered as a one form on \(\mathbb{S}^2\). Note that
\[ \psi(\theta, \varphi) = (\sin \theta \cos \varphi, \sin \theta \sin \varphi \cos \theta) \]
is a parameterization of the \(\mathbb{S}^2\) and hence
\[ \psi^{-1}(x, y, z) = (\cos^{-1}(z), \tan^{-1}(y/x)) \]
provides the structure of a chart on \(\mathbb{S}^2\). We already found that the pull-back of \(\omega\) into \(M\) is
\[ \psi^* \omega = \sin \theta \cos \varphi \cos \theta \sin \varphi \, d\theta + \sin \theta \cos \varphi \sin \theta \cos \varphi \, d\varphi. \]
Proposition 23.2 is now applied \( \psi^* \omega \) to give us
\[
d\psi^* \omega = d(\sin \theta \cos \varphi \cos \theta \sin \varphi \, d\theta) + d(\sin \theta \cos \varphi \sin \theta \cos \varphi \, d\varphi)
\]
\[
= d(\sin \theta \cos \varphi \cos \theta \sin \varphi) \wedge d\theta + d(\sin \theta \cos \varphi \sin \theta \cos \varphi) \wedge d\varphi
\]
\[
= \frac{\partial}{\partial \varphi} (\sin \theta \cos \varphi \cos \theta \sin \varphi) d\varphi \wedge d\theta + \frac{\partial}{\partial \theta} (\sin \theta \cos \varphi \sin \theta \cos \varphi) d\theta \wedge d\varphi
\]
\[
= \sin \theta \cos \theta (-\sin^2 \varphi + \cos^2 \varphi) d\varphi \wedge d\theta + 2 \sin \theta \cos \theta \cos^2 \varphi d\theta \wedge d\varphi
\]
\[
= \sin \theta \cos \theta (\sin^2 \varphi + \cos^2 \varphi) d\theta \wedge d\varphi
\]
\[
= \sin \theta \cos \theta \, d\theta \wedge d\varphi.
\]

It just remains to compose \( d\psi^* \omega \) with \( \psi^{-1} \) to obtain
\[
d\omega = \psi^{-1} (d\psi^* \omega ) = z \sqrt{1 - z^2} (\cos^{-1} z) \wedge d(\tan^{-1} y/x)
\]
\[
= -z \, dz \wedge \left( -\frac{y}{x^2 + y^2} \, dx + \frac{x}{x^2 + y^2} \, dy \right)
\]
\[
= \frac{zy}{x^2 + y^2} \, dz \wedge dx - \frac{zx}{x^2 + y^2} \, dz \wedge dy.
\]

Propositions 23.3, 23.4 and 23.6 generalize to manifolds.

**Proposition 23.9.** Let \( M \) and \( N \) be smooth manifolds and let \( \varphi: M \to N \) be a smooth map.

1. For all \( \omega \in \mathcal{A}^r(M) \) and all \( \eta \in \mathcal{A}^s(M) \),
\[
d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^r \omega \wedge d\eta.
\]

2. For every \( k \geq 0 \), the composition \( \mathcal{A}^k(M) \xrightarrow{d} \mathcal{A}^{k+1}(M) \xrightarrow{d} \mathcal{A}^{k+2}(M) \) is identically zero; that is,
\[
d \circ d = 0.
\]

3. \( \varphi^*(\omega \wedge \eta) = \varphi^* \omega \wedge \varphi^* \eta \), for all \( \omega \in \mathcal{A}^r(N) \) and all \( \eta \in \mathcal{A}^s(N) \).

4. \( \varphi^*(f) = f \circ \varphi \), for all \( f \in \mathcal{A}^0(N) \).

5. \( d\varphi^*(\omega) = \varphi^*(d\omega) \), for all \( \omega \in \mathcal{A}^k(N) \); that is, the following diagram commutes for all \( k \geq 0 \):
\[
\begin{array}{ccc}
\mathcal{A}^k(N) & \xrightarrow{\varphi^*} & \mathcal{A}^k(M) \\
\downarrow d & & \downarrow d \\
\mathcal{A}^{k+1}(N) & \xrightarrow{\varphi^*} & \mathcal{A}^{k+1}(M).
\end{array}
\]
Proof. It is enough to prove these properties in a chart \((U, \varphi)\), which is easy. We only check (2). We have
\[
(d(d\omega))_p = d((\varphi^* (d((\varphi^{-1})*\omega))_{\varphi(p)}))_p \\
= \varphi^* \left[ d\left( (\varphi^{-1})^* \left( \varphi^* (d((\varphi^{-1})*\omega))_{\varphi(p)} \right) \right) \right]_{\varphi(p)} \\
= \varphi^* \left[ d(d((\varphi^{-1})*\omega))_{\varphi(p)} \right]_p \\
= 0,
\]
as \((\varphi^{-1})^* \circ \varphi^* = (\varphi \circ \varphi^{-1})^* = \text{id}^* = \text{id}\) and \(d \circ d = 0\) on forms in \(\mathcal{A}^k(\varphi(U))\), with \(\varphi(U) \subseteq \mathbb{R}^n\).

As a consequence, Definition 23.5 of the de Rham cohomology generalizes to manifolds. For every manifold \(M\), we have the de Rham complex
\[
\mathcal{A}^0(M) \xrightarrow{d} \mathcal{A}^1(M) \rightarrow \cdots \rightarrow \mathcal{A}^{k-1}(M) \xrightarrow{d} \mathcal{A}^k(M) \xrightarrow{d} \mathcal{A}^{k+1}(M) \rightarrow \cdots,
\]
and we can define the cohomology groups \(H^k_{\text{DR}}(M)\) and the graded cohomology algebra \(H^\bullet_{\text{DR}}(M)\). For every \(k \geq 0\), let
\[
Z^k(M) = \{\omega \in \mathcal{A}^k(M) \mid d\omega = 0\} = \text{Ker} \ d : \mathcal{A}^k(M) \rightarrow \mathcal{A}^{k+1}(M)
\]
be the vector space of closed \(k\)-forms, and for every \(k \geq 1\), let
\[
B^k(M) = \{\omega \in \mathcal{A}^k(M) \mid \exists \eta \in \mathcal{A}^{k-1}(M), \omega = d\eta\} = \text{Im} \ d : \mathcal{A}^{k-1}(M) \rightarrow \mathcal{A}^k(M)
\]
be the vector space of exact \(k\)-forms, with \(B^0(M) = (0)\). Then, for every \(k \geq 0\), we define the \(k\)th de Rham cohomology group of \(M\) as the quotient space
\[
H^k_{\text{DR}}(M) = Z^k(M)/B^k(M).
\]
This is an abelian group under addition of cosets. The real vector space \(H^\bullet_{\text{DR}}(M) = \bigoplus_{k \geq 0} H^k_{\text{DR}}(M)\) is called the de Rham cohomology algebra of \(M\). We often drop the subscript \(\text{DR}\) when no confusion arises. Every smooth map \(\varphi : M \rightarrow N\) between two manifolds induces an algebra map \(\varphi^* : H^\bullet(N) \rightarrow H^\bullet(M)\).

Another important property of the exterior differential is that it is a local operator, which means that the value of \(d\omega\) at \(p\) only depends on the values of \(\omega\) near \(p\). Not all operators are local. For example, the operator \(I : C^\infty([a,b]) \rightarrow C^\infty([a,b])\) given by
\[
I(f) = \int_a^b f(t) \, dt,
\]
where \(I(f)\) is the constant function on \([a,b]\) is not local since for any point \(p \in [a,b]\), the calculation of \(I(f)\) requires evaluating \(f\) over \([a,b]\).

More generally, we have the following definition.
**Definition 23.10.** A linear map $D: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ is a **local operator** if for all $k \geq 0$, for any nonempty open subset $U \subseteq M$ and for any two $k$-forms $\omega, \eta \in \mathcal{A}^k(M)$, if $\omega \upharpoonright U = \eta \upharpoonright U$, then $(D\omega) \upharpoonright U = (D\eta) \upharpoonright U$. Since $D$ is linear, the above condition is equivalent to: for any $k$-form $\omega \in \mathcal{A}^k(M)$, if $\omega \upharpoonright U = 0$, then $(D\omega) \upharpoonright U = 0$.

Since property (1) of Proposition 23.9 comes up a lot, we introduce the following definition.

**Definition 23.11.** Given any smooth manifold $M$, a linear map $D: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ is called an **antiderivation** if for all $r,s \geq 0$, for all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$,

$$D(\omega \wedge \eta) = D\omega \wedge \eta + (-1)^r \omega \wedge D\eta.$$  

The antiderivation is of **degree** $m \in \mathbb{Z}$ if $D: \mathcal{A}^p(M) \to \mathcal{A}^{p+m}(M)$ for all $p$ such that $p+m \geq 0$.

By Proposition 23.9, exterior differentiation $d: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ is an antiderivation of degree 1.

**Proposition 23.10.** Let $M$ be a smooth manifold. Any linear antiderivation $D: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ is a local operator.

**Proof.** By linearity, it is enough to show that if $\omega \upharpoonright U = 0$, then $(D\omega) \upharpoonright U = 0$. There is an apparent problem, which is that although $\omega$ is zero on $U$, it may not be zero outside $U$, so it is not obvious that we can conclude that $D\omega$ is zero on $U$. The crucial ingredient to circumvent this difficulty is the existence of “bump functions.” By Proposition 9.2 applied to the constant function with value 1, for every $p \in U$, there some open subset $V \subseteq U$ containing $p$ and a smooth function $f: M \to \mathbb{R}$ such that $\text{supp } f \subseteq U$ and $f \equiv 1$ on $V$. Consequently, $f\omega$ is a smooth differential form which is identically zero, and since $D$ is an antiderivation

$$D(f\omega) = Df \wedge \omega + fD\omega;$$

which, evaluated ap $p$ yields

$$0 = Df_p \wedge \omega_p + 1D\omega_p = Df_p \wedge 0 + 1D\omega_p = D\omega_p;$$

that is, $D\omega_p = 0$, as claimed. \hfill \Box

**Remark:** If $D: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ is a linear map which is a **derivation**, which means that

$$D(\omega \wedge \eta) = D\omega \wedge \eta + \omega \wedge D\eta$$

for all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$, then the proof of Proposition 23.10 still works and shows that $D$ is also a local operator.

By Proposition 23.10, exterior differentiation $d: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ is a local operator. As in the case of differential forms on $\mathbb{R}^n$, the operator $d$ is uniquely determined by the properties of Theorem 23.5.
Theorem 23.11. Let $M$ be a smooth manifold. There is a unique linear operator 
\[ d: \mathcal{A}^r(M) \to \mathcal{A}^r(M), \]
with $d = (d^k)$ and $d^k: \mathcal{A}^k(M) \to \mathcal{A}^{k+1}(M)$ for every $k \geq 0$, such that

1. $(df)_p = df_p$, where $df_p$ is the differential of $f$ at $p \in M$ for every $f \in \mathcal{A}^0(M) = C^\infty(M)$.

2. $d \circ d = 0$.

3. For every $\omega \in \mathcal{A}^r(M)$ and every $\eta \in \mathcal{A}^s(M)$,
\[ d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^{r}\omega \wedge d\eta. \]

Furthermore, any linear operator $d$ satisfying (1)–(3) is a local operator.

Proof. Existence has already been established.

Let $D: \mathcal{A}^r(M) \to \mathcal{A}^r(M)$ be any linear operator satisfying (1)–(3). We need to prove that $D = d$ where $d$ is defined in Definition 23.9. For any $k \geq 0$, pick any $\omega \in \mathcal{A}^k(M)$. For every $p \in M$, we need to prove that $(D\omega)_p = (d\omega)_p$. Let $(U, \varphi)$ be any chart with $p \in U$, and let $x_i = pr_i \circ \varphi$ be the corresponding local coordinate maps. We know that $\omega \in \mathcal{A}^k(M)$ can be written uniquely as
\[ \omega_q = \sum_I f_I(q)dx_{i_1} \wedge \cdots \wedge dx_{i_k} \quad q \in U. \]

Using Proposition 9.2, there is some open subset $V$ of $U$ with $p \in V$ and some functions $\tilde{f}_I$, and $\tilde{x}_{i_1}, \ldots, \tilde{x}_{i_k}$ defined on $M$ and agreeing with $f_I, x_{i_1}, \ldots, x_{i_k}$ on $V$. If we define
\[ \tilde{\omega} = \sum_I \tilde{f}_I d\tilde{x}_{i_1} \wedge \cdots \wedge d\tilde{x}_{i_k} \]
then $\tilde{\omega}$ is defined for all $p \in M$ and
\[ \omega | V = \tilde{\omega} | V. \]

By Proposition 23.10, since $D$ is a linear map satisfying (3), it is a local operator so
\[ D\omega | V = D\tilde{\omega} | V. \]
Since $D$ satisfies (1), we have $D\tilde{x}_{ij} = d\tilde{x}_{ij}$. Then, at $p$, by linearity we have
\[ (D\omega)_p = (D\tilde{\omega})_p = \left(D\left(\sum_I \tilde{f}_I d\tilde{x}_{i_1} \wedge \cdots \wedge d\tilde{x}_{i_k}\right)\right)_p \]
\[ = \left(\sum_I \tilde{f}_I D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k}\right)_p \]
\[ = \sum_I D\left(\tilde{f}_I D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k}\right)_p. \]
As in the proof of Proposition 23.6(iii), using (3) we can show by induction that

\[ D(D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k}) = \sum_{j=1}^{k} (-1)^{j-1} D\tilde{x}_{i_1} \wedge \cdots \wedge DD\tilde{x}_{i_j} \wedge \cdots \wedge D\tilde{x}_{i_k} \]

and since by (2) \( DD\tilde{x}_{ij} = 0 \), we have

\[ D(D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k}) = 0. \]

Then, using the above, by (3) and (1), we get

\[ \sum_I D\left(\tilde{f}_I D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k}\right)_p = \left(\sum_I D\tilde{f}_I \wedge D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k}\right)_p + \left(\sum_I \tilde{f}_I \wedge D(D\tilde{x}_{i_1} \wedge \cdots \wedge D\tilde{x}_{i_k})\right)_p \]
\[ = \left(\sum_I df_I \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}\right)_p \]
\[ = (d\omega)_p. \]

Therefore \((D\omega)_p = (d\omega)_p\), which proves that \( D = d \).

\[ \square \]

**Remark:** A closer look at the proof of Theorem 23.11 shows that it is enough to assume \( DD\omega = 0 \) on forms \( \omega \in \mathcal{A}^0(M) = C^\infty(M) \).

Smooth differential forms can also be defined in terms of alternating \( C^\infty(M) \)-multilinear maps on smooth vector fields. This approach also yields a global formula for the exterior derivative \( d\omega(X_1, \ldots, X_{k+1}) \) of a \( k \)-form \( \omega \) applied to \( k+1 \) vector fields \( X_1, \ldots, X_{k+1} \). This formula is not very useful for computing \( d\omega \) at a given point \( p \) since it requires vector fields as input, but it is quite useful in theoretical investigations.

Let \( \omega \in \mathcal{A}^k(M) \) be any smooth \( k \)-form on \( M \). Then, \( \omega \) induces an alternating multilinear map

\[ \omega: \underbrace{\mathcal{X}(M) \times \cdots \times \mathcal{X}(M)}_{k} \longrightarrow C^\infty(M) \]

as follows: For any \( k \) smooth vector fields \( X_1, \ldots, X_k \in \mathcal{X}(M) \),

\[ \omega(X_1, \ldots, X_k)(p) = \omega_p(X_1(p), \ldots, X_k(p)). \]
This map is obviously alternating and \( \mathbb{R} \)-linear, but it is also \( C^\infty(M) \)-linear, since for every \( f \in C^\infty(M) \),

\[
\omega(X_1, \ldots, fX_i, \ldots X_k)(p) = \omega_p(X_1(p), \ldots, f(p)X_i(p), \ldots, X_k(p))
\]

\[
= f(p)\omega_p(X_1(p), \ldots, X_i(p), \ldots, X_k(p))
\]

\[
= (f\omega)_p(X_1(p), \ldots, X_i(p), \ldots, X_k(p)).
\]

(Recall, that the set of smooth vector fields \( \mathfrak{X}(M) \) is a real vector space and a \( C^\infty(M) \)-module.)

Interestingly, every alternating \( C^\infty(M) \)-multilinear map on smooth vector fields determines a differential form. This is because \( \omega(X_1, \ldots, X_k)(p) \) only depends on the values of \( X_1, \ldots, X_k \) at \( p \).

**Proposition 23.12.** Let \( M \) be a smooth manifold. For every \( k \geq 0 \), there is an isomorphism between the space of \( k \)-forms \( \mathcal{A}^k(M) \) and the space \( \text{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M)) \) of alternating \( C^\infty(M) \)-multilinear maps on smooth vector fields. That is,

\[
\mathcal{A}^k(M) \cong \text{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M)),
\]

viewed as \( C^\infty(M) \)-modules.

**Proof.** We follow the proof in O’Neill [138] (Chapter 2, Lemma 3 and Proposition 2). Let \( \Phi \colon \mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M) \to C^\infty(M) \) be an alternating \( C^\infty(M) \)-multilinear map. First, we prove that for any vector fields \( X_1, \ldots, X_k \) and \( Y_1, \ldots, Y_k \), for every \( p \in M \), if \( X_i(p) = Y_i(p) \), then

\[
\Phi(X_1, \ldots, X_k)(p) = \Phi(Y_1, \ldots, Y_k)(p).
\]

Observe that

\[
\Phi(X_1, \ldots, X_k) - \Phi(Y_1, \ldots, Y_k) = \Phi(X_1 - Y_1, X_2, \ldots, X_k) + \Phi(Y_1, X_2 - Y_2, X_3, \ldots, X_k) + \Phi(Y_1, Y_2, X_3 - Y_3, \ldots, X_k) + \cdots + \Phi(Y_1, \ldots, Y_{k-2}, X_{k-1} - Y_{k-1}, X_k) + \Phi(Y_1, \ldots, Y_{k-1}, X_k - Y_k).
\]

As a consequence, it is enough to prove that if \( X_i(p) = 0 \) for some \( i \), then

\[
\Phi(X_1, \ldots, X_k)(p) = 0.
\]

Without loss of generality, assume \( i = 1 \). In any local chart \( (U, \varphi) \) near \( p \), we can write

\[
X_1 = \sum_{i=1}^n f_i \frac{\partial}{\partial x_i},
\]
and as $X_i(p) = 0$, we have $f_i(p) = 0$ for $i = 1, \ldots, n$. Since the expression on the right-hand side is only defined on $U$, we extend it using Proposition 9.2 once again. There is some open subset $V \subseteq U$ containing $p$ and a smooth function $h: M \to \mathbb{R}$ such that $\text{supp } h \subseteq U$ and $h \equiv 1$ on $V$. Then, we let $h_i = hf_i$, a smooth function on $M$, $Y_i = h_i \frac{\partial}{\partial x_i}$, a smooth vector field on $M$, and we have $h_i \upharpoonright V = f_i \upharpoonright V$ and $Y_i \upharpoonright V = \frac{\partial}{\partial x_i} \upharpoonright V$. Now, since $h^2 = 1$ on $V$, it is obvious that

$$X_1 = \sum_{i=1}^n h_i Y_i + (1 - h^2) X_1$$
on V, so as $\Phi$ is $C^\infty(M)$-multilinear, $h_i(p) = 0$ and $h(p) = 1$, we get

$$\Phi(X_1, X_2, \ldots, X_k)(p) = \Phi\left( \sum_{i=1}^n h_i Y_i + (1 - h^2) X_1, X_2, \ldots, X_k \right)(p)$$

$$= \sum_{i=1}^n h_i(p) \Phi(Y_i, X_2, \ldots, X_k)(p) + (1 - h^2(p)) \Phi(X_1, X_2, \ldots, X_k)(p) = 0,$$

as claimed.

Next, we show that $\Phi$ induces a smooth differential form. For every $p \in M$, for any $u_1, \ldots, u_k \in T_p M$, we can pick smooth functions $f_i$ equal to 1 near $p$ and 0 outside some open near $p$, so that we get smooth vector fields $X_1, \ldots, X_k$ with $X_k(p) = u_k$. We set

$$\omega_p(u_1, \ldots, u_k) = \Phi(X_1, \ldots, X_k)(p).$$

As we proved that $\Phi(X_1, \ldots, X_k)(p)$ only depends on $X_1(p) = u_1, \ldots, X_k(p) = u_k$, the function $\omega_p$ is well defined, and it is easy to check that it is smooth. Therefore, the map $\Phi \mapsto \omega$ just defined is indeed an isomorphism.

\begin{itemize}
\item[(1)] The space $\text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), C^\infty(M))$ of all $C^\infty(M)$-linear maps $\mathfrak{X}(M) \to C^\infty(M)$ is also a $C^\infty(M)$-module, called the dual of $\mathfrak{X}(M)$, and sometimes denoted $\mathfrak{X}^*(M)$. Proposition 23.12 shows that as $C^\infty(M)$-modules,

$$\mathcal{A}^1(M) \cong \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), C^\infty(M)) = \mathfrak{X}^*(M).$$

\item[(2)] A result analogous to Proposition 23.12 holds for tensor fields. Indeed, there is an isomorphism between the set of tensor fields $\Gamma(M, T^{r,s}_p(M))$, and the set of $C^\infty(M)$-multilinear maps

$$\Phi: \underbrace{\mathcal{A}^1(M) \times \cdots \times \mathcal{A}^1(M)}_{r} \times \underbrace{\mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M)}_{s} \to C^\infty(M),$$

where $\mathcal{A}^1(M)$ and $\mathfrak{X}(M)$ are $C^\infty(M)$-modules.
\end{itemize}
Recall from Section 8.1 (Definition 8.2) that for any function \( f \in C^\infty(M) \) and every vector field \( X \in \mathfrak{X}(M) \), the Lie derivative \( X[f] \) (or \( X(f) \)) of \( f \) w.r.t. \( X \) is defined so that
\[
X[f]_p = df_p(X(p)).
\]
Also recall the notion of the Lie bracket \([X, Y]\) of two vector fields (see Definition 8.3). The interpretation of differential forms as \( C^\infty(M) \)-multilinear forms given by Proposition 23.12 yields the following formula for \((d\omega)(X_1, \ldots, X_{k+1})\), where the \( X_i \) are vector fields:

**Proposition 23.13.** Let \( M \) be a smooth manifold. For every \( k \)-form \( \omega \in \mathcal{A}^k(M) \), we have
\[
(d\omega)(X_1, \ldots, X_{k+1}) = \sum_{i=1}^{k+1} (-1)^{i-1} X_i[\omega(X_1, \ldots, \widehat{X_i}, \ldots, X_{k+1})] + \sum_{i<j} (-1)^{i+j} \omega([X_i, X_j], X_1, \ldots, \widehat{X_i}, \ldots, \widehat{X_j}, \ldots, X_{k+1})],
\]
for all vector fields, \( X_1, \ldots, X_{k+1} \in \mathfrak{X}(M) \):

**Proof sketch.** First, one checks that the right-hand side of the formula in Proposition 23.13 is alternating and \( C^\infty(M) \)-multilinear. For this, use Proposition 8.4 (c). Consequently, by Proposition 23.12, this expression defines a \((k+1)\)-form. Second, it is enough to check that both sides of the equation agree on charts \((U, \varphi)\). Then, we know that \( d\omega \) can be written uniquely as
\[
\omega = \sum_I f_I dx_{i_1} \wedge \cdots \wedge dx_{i_k} \quad p \in U.
\]
Also, as differential forms are \( C^\infty(M) \)-multilinear, it is enough to consider vector fields of the form \( X_i = \frac{\partial}{\partial x_i} \). However, for such vector fields, \([X_i, X_j] = 0\), and then it is a simple matter to check that the equation holds. For more details, see Morita [133] (Chapter 2).

In particular, when \( k = 1 \), Proposition 23.13 yields the often used formula:
\[
d\omega(X, Y) = X[\omega(Y)] - Y[\omega(X)] - \omega([X, Y]).
\]
There are other ways of proving the formula of Proposition 23.13, for instance, using Lie derivatives.

Before considering the Lie derivative \( L_X \omega \) of differential forms, we define interior multiplication by a vector field, \( i(X)\omega \). We will see shortly that there is a relationship between \( L_X, i(X) \), and \( d \), known as Cartan’s Formula.

**Definition 23.12.** Let \( M \) be a smooth manifold. For every vector field \( X \in \mathfrak{X}(M) \), for all \( k \geq 1 \), there is a linear map \( i(X) : \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M) \) defined so that, for all \( \omega \in \mathcal{A}^k(M) \), for all \( p \in M \), for all \( u_1, \ldots, u_{k-1} \in T_pM \),
\[
(i(X)\omega)_p(u_1, \ldots, u_{k-1}) = \omega_p(X_p, u_1, \ldots, u_{k-1}).
\]
Obviously, $i(X)$ is $C^\infty(M)$-linear in $X$, namely

$$i(fX)\omega = fi(X)\omega, \quad i(X)(f\omega) = fi(X)\omega,$$

and it is easy to check that $i(X)\omega$ is indeed a smooth $(k-1)$-form. When $k = 0$, we set $i(X)\omega = 0$. Observe that $i(X)\omega$ is also given by

$$(i(X)\omega)_p = i(X_p)\omega_p, \quad p \in M,$$

where $i(X_p)$ is the interior product (or insertion operator) defined in Section 22.7 (with $i(X_p)\omega_p$ equal to our right hook, $\omega_p \hookrightarrow X_p$). As a consequence, by Proposition 22.18, the operator $i(X)$ is an anti-derivation of degree $-1$; that is, we have

$$i(X)(\omega \wedge \eta) = (i(X)\omega) \wedge \eta + (-1)^r \omega \wedge (i(X)\eta),$$

for all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$.

**Remark:** Other authors, including Marsden, use a left hook instead of a right hook, and denote $i(X)\omega$ as $X \bservable \omega$.

### 23.3 Lie Derivatives

We just saw in Section 23.2 that for any function $f \in C^\infty(M)$ and every vector field $X \in \mathfrak{X}(M)$, the Lie derivative $X[f]$ (or $X(f)$) of $f$ w.r.t. $X$ is defined so that

$$X[f]_p = df_p(X_p).$$

Recall from Definition 8.11 and the observation immediately following it that for any manifold $M$, given any two vector fields $X, Y \in \mathfrak{X}(M)$, the *Lie derivative of $X$ with respect to $Y$* is given by

$$(L_X Y)_p = \lim_{t \to 0} \frac{\left(\Phi_t^* Y\right)_p - Y_p}{t} = \frac{d}{dt} \left(\Phi_t^* Y\right)_p \bigg|_{t=0},$$

where $\Phi_t$ is the local one-parameter group associated with $X$ ($\Phi$ is the global flow associated with $X$; see Definition 8.9, Theorem 8.9, and the remarks following it), and $\Phi_t^*$ is the pull-back of the diffeomorphism $\Phi_t$ (see Definition 8.4). Furthermore, to calculate $L_X Y$ recall that

$$L_X Y = [X, Y].$$

We claim that we also have

$$X_p[f] = \lim_{t \to 0} \frac{(\Phi_t^* f)(p) - f(p)}{t} = \frac{d}{dt} (\Phi_t^* f)(p) \bigg|_{t=0},$$

with $\Phi_t^* f = f \circ \Phi_t$ (as usual for functions).
Recall from Section 8.3 that if $\Phi$ is the flow of $X$, then for every $p \in M$, the map $t \mapsto \Phi_t(p)$ is an integral curve of $X$ through $p$; that is

$$\dot{\Phi}_t(p) = X(\Phi_t(p)), \quad \Phi_0(p) = p,$$

in some open set containing $p$. In particular, $\dot{\Phi}_0(p) = X_p$. Then, we have

$$\lim_{t \to 0} \frac{(\Phi^*_t f)(p) - f(p)}{t} = \lim_{t \to 0} \frac{f(\Phi_t(p)) - f(\Phi_0(p))}{t} = \frac{d}{dt} (f \circ \Phi_t(p)) \bigg|_{t=0} = df_p(\dot{\Phi}_0(p)) = df_p(X_p) = X_p[f].$$

We would like to define the Lie derivative of differential forms (and tensor fields). This can be done algebraically or in terms of flows; the two approaches are equivalent, but it seems more natural to give a definition using flows.

**Definition 23.13.** Let $M$ be a smooth manifold. For every vector field $X \in \mathfrak{X}(M)$, for every $k$-form $\omega \in A^k(M)$, the *Lie derivative of $\omega$ with respect to $X$*, denoted $L_X \omega$, is given by

$$(L_X \omega)_p = \lim_{t \to 0} \frac{(\Phi^*_t \omega)_p - \omega_p}{t} = \frac{d}{dt} (\Phi^*_t \omega)_p \bigg|_{t=0},$$

where $\Phi^*_t \omega$ is the pull-back of $\omega$ along $\Phi_t$ (see Definition 23.8).

Obviously, $L_X : A^k(M) \to A^k(M)$ is a linear map, but it has many other interesting properties.

We can also define the Lie derivative on tensor fields as a map $L_X : \Gamma(M, T^{r,s}(M)) \to \Gamma(M, T^{r,s}(M))$, by requiring that for any tensor field

$$\alpha = X_1 \otimes \cdots \otimes X_r \otimes \omega_1 \otimes \cdots \otimes \omega_s,$$

where $X_i \in \mathfrak{X}(M)$ and $\omega_j \in A^1(M)$,

$$\Phi^*_t \alpha = \Phi^*_t X_1 \otimes \cdots \otimes \Phi^*_t X_r \otimes \Phi^*_t \omega_1 \otimes \cdots \otimes \Phi^*_t \omega_s,$$

where $\Phi^*_t X_i$ is the pull-back of the vector field, $X_i$, and $\Phi^*_t \omega_j$ is the pull-back of one-form $\omega_j$, and then setting

$$(L_X \alpha)_p = \lim_{t \to 0} \frac{(\Phi^*_t \alpha)_p - \alpha_p}{t} = \frac{d}{dt} (\Phi^*_t \alpha)_p \bigg|_{t=0}.$$

So, as long we can define the “right” notion of pull-back, the formula giving the Lie derivative of a function, a vector field, a differential form, and more generally a tensor field, is the same.
The Lie derivative of tensors is used in most areas of mechanics, for example in elasticity (the rate of strain tensor) and in fluid dynamics.

We now state, mostly without proofs, a number of properties of Lie derivatives. Most of these proofs are fairly straightforward computations, often tedious, and can be found in most texts, including Warner [175], Morita [133], and Gallot, Hullin and Lafontaine [73].

**Proposition 23.14.** Let $M$ be a smooth manifold. For every vector field $X \in \mathfrak{X}(M)$, the following properties hold:

1. For all $\omega \in \mathcal{A}^r(M)$ and all $\eta \in \mathcal{A}^s(M)$,
   \[ L_X(\omega \wedge \eta) = (L_X \omega) \wedge \eta + \omega \wedge (L_X \eta); \]
   that is, $L_X$ is a derivation.

2. For all $\omega \in \mathcal{A}^k(M)$, for all $Y_1, \ldots, Y_k \in \mathfrak{X}(M)$,
   \[ L_X(\omega(Y_1, \ldots, Y_k)) = (L_X \omega)(Y_1, \ldots, Y_k) + \sum_{i=1}^k \omega(Y_1, \ldots, Y_i, L_X Y_i, Y_{i+1}, \ldots, Y_k). \]

3. The Lie derivative commutes with $d$:
   \[ L_X \circ d = d \circ L_X. \]

**Proof.** We only prove (2). First, we claim that if $\varphi: M \to M$ is a diffeomorphism, then for every $\omega \in \mathcal{A}^k(M)$, for all $X_1, \ldots, X_k \in \mathfrak{X}(M)$,
   \[ (\varphi^* \omega)(X_1, \ldots, X_k) = \varphi^*(\omega((\varphi^{-1})^*X_1, \ldots, (\varphi^{-1})^*X_k)), \]
   where $(\varphi^{-1})^*X_i$ is the pull-back of the vector field $X_i$ (also equal to the push-forward $\varphi_*X_i$ of $X_i$; see Definition 8.4). Recall that
   \[ ((\varphi^{-1})^*Y)_p = d\varphi_{\varphi^{-1}(p)}(Y_{\varphi^{-1}(p)}), \]
   for any vector field $Y$. Then, for every $p \in M$, we have
   \[
   (\varphi^*(\omega(X_1, \ldots, X_k)))(p) = \omega_{\varphi(p)}(d\varphi_p(X_1(p)), \ldots, d\varphi_p(X_k(p)))
   = \omega_{\varphi(p)}(d\varphi_{\varphi^{-1}(\varphi(p))}(X_1(\varphi^{-1}(\varphi(p)))), \ldots, d\varphi_{\varphi^{-1}(\varphi(p))}(X_k(\varphi^{-1}(\varphi(p)))))
   = \omega_{\varphi(p)}(((\varphi^{-1})^*X_1)_{\varphi(p)}, \ldots, ((\varphi^{-1})^*X_k)_{\varphi(p)})
   = (\omega((\varphi^{-1})^*X_1, \ldots, (\varphi^{-1})^*X_k)) \circ \varphi)(p)
   = \varphi^*(\omega((\varphi^{-1})^*X_1, \ldots, (\varphi^{-1})^*X_k))(p),
   \]
   since for any function $g \in C^\infty(M)$, we have $\varphi^*g = g \circ \varphi$. 
23.3. LIE DERIVATIVES

We know that

\[ X_p[f] = \lim_{t \to 0} \frac{(\Phi^*_t f)(p) - f(p)}{t} \]

and for any vector field \( Y \),

\[ [X, Y]_p = (L_X Y)_p = \lim_{t \to 0} \frac{(\Phi^*_t Y)_p - Y_p}{t}. \]

Since the one-parameter group associated with \( -X \) is \( \Phi_{-t} \) (this follows from \( \Phi_{-t} \circ \Phi_t = \text{id} \)), we have

\[ \lim_{t \to 0} \frac{(\Phi^*_{-t} Y)_p - Y_p}{t} = -[X, Y]_p. \]

Now, using \( \Phi_t^{-1} = \Phi_{-t} \) and \( (*) \), we have

\[
(L_X \omega)(Y_1, \ldots, Y_k) = \lim_{t \to 0} \frac{(\Phi^*_t \omega)(Y_1, \ldots, Y_k) - \omega(Y_1, \ldots, Y_k)}{t} \\
= \lim_{t \to 0} \frac{\Phi^*_t (\omega(\Phi^*_{-t} Y_1, \ldots, \Phi^*_{-t} Y_k)) - \omega(Y_1, \ldots, Y_k)}{t} \\
= \lim_{t \to 0} \frac{\Phi^*_t (\omega(\Phi^*_{-t} Y_1, \ldots, \Phi^*_{-t} Y_k)) - \Phi^*_t (\omega(Y_1, \ldots, Y_k))}{t} \\
+ \lim_{t \to 0} \frac{\Phi^*_t (\omega(Y_1, \ldots, Y_k)) - \omega(Y_1, \ldots, Y_k)}{t}.
\]

Call the first term \( A \) and the second term \( B \). Then, as

\[ X_p[f] = \lim_{t \to 0} \frac{(\Phi^*_t f)(p) - f(p)}{t}, \]

we have

\[ B = X[\omega(Y_1, \ldots, Y_k)]. \]

As to \( A \), we have

\[
A = \lim_{t \to 0} \frac{\Phi^*_t (\omega(\Phi^*_{-t} Y_1, \ldots, \Phi^*_{-t} Y_k)) - \Phi^*_t (\omega(Y_1, \ldots, Y_k))}{t} \\
= \lim_{t \to 0} \Phi^*_t \left( \frac{\omega(\Phi^*_{-t} Y_1, \ldots, \Phi^*_{-t} Y_k) - \omega(Y_1, \ldots, Y_k)}{t} \right) \\
= \lim_{t \to 0} \Phi^*_t \left( \frac{\omega(\Phi^*_{-t} Y_1, \ldots, \Phi^*_{-t} Y_k) - \omega(Y_1, \Phi^*_{-t} Y_2, \ldots, \Phi^*_{-t} Y_k)}{t} \right) \\
+ \lim_{t \to 0} \Phi^*_t \left( \frac{\omega(Y_1, \Phi^*_{-t} Y_2, \ldots, \Phi^*_{-t} Y_k) - \omega(Y_1, Y_2, \Phi^*_{-t} Y_3, \ldots, \Phi^*_{-t} Y_k)}{t} \right) \\
+ \cdots + \lim_{t \to 0} \Phi^*_t \left( \frac{\omega(Y_1, \ldots, Y_{k-1}, \Phi^*_{-t} Y_k) - \omega(Y_1, \ldots, Y_k)}{t} \right) \\
= \sum_{i=1}^k \omega(Y_1, \ldots, -[X, Y_i], \ldots, Y_k). \]
When we add up \( A \) and \( B \), we get
\[
A + B = X[\omega(Y_1, \ldots, Y_k)] - \sum_{i=1}^{k} \omega(Y_1, \ldots, [X, Y_i], \ldots, Y_k)
\]
\[
= (L_X \omega)(Y_1, \ldots, Y_k),
\]
which finishes the proof.

Part (2) of Proposition 23.14 shows that the Lie derivative of a differential form can be defined in terms of the Lie derivatives of functions and vector fields:
\[
(L_X \omega)(Y_1, \ldots, Y_k) = L_X(\omega(Y_1, \ldots, Y_k)) - \sum_{i=1}^{k} \omega(Y_1, \ldots, Y_{i-1}, [X, Y_i], Y_{i+1}, \ldots, Y_k).
\]

To best calculate \( L_X \omega \), we use Cartan's Formula:

**Proposition 23.15. (Cartan’s Formula)** Let \( M \) be a smooth manifold. For every vector field \( X \in \mathfrak{X}(M) \), for every \( \omega \in \mathcal{A}^k(M) \), we have
\[
L_X \omega = i(X) d\omega + d(i(X)\omega),
\]
that is, \( L_X = i(X) \circ d + d \circ i(X) \).

**Proof.** If \( k = 0 \), then \( L_X f = X[f] = df(X) \) for a function \( f \), and on the other hand, \( i(X)f = 0 \) and \( i(X)df = df(X) \), so the equation holds. If \( k \geq 1 \), then by Proposition 23.13, we have
\[
(i(X)d\omega)(X_1, \ldots, X_k) = d\omega(X, X_1, \ldots, X_k)
\]
\[
= X[\omega(X_1, \ldots, X_k)] + \sum_{i=1}^{k} (-1)^i X_i[\omega(X, X_1, \ldots, \widehat{X_i}, \ldots, X_k)]
\]
\[
+ \sum_{j=1}^{k} (-1)^j \omega([X, X_j], X_1, \ldots, \widehat{X_j}, \ldots, X_k)
\]
\[
+ \sum_{i<j} (-1)^{i+j} \omega([X_i, X_j], X_1, \ldots, \widehat{X_i}, \ldots, \widehat{X_j}, \ldots, X_k).
\]

On the other hand, again by Proposition 23.13, we have
\[
(d(i(X)\omega))(X_1, \ldots, X_k) = \sum_{i=1}^{k} (-1)^{i-1} X_i[\omega(X, X_1, \ldots, \widehat{X_i}, \ldots, X_k)]
\]
\[
+ \sum_{i<j} (-1)^{i+j} \omega(X, [X_i, X_j], X_1, \ldots, \widehat{X_i}, \ldots, \widehat{X_j}, \ldots, X_k).
\]
Adding up these two equations, we get
\[
(i(X)d\omega + di(X))\omega(X_1, \ldots, X_k) = X[\omega(X_1, \ldots, X_k)] \\
+ \sum_{i=1}^{k} (-1)^i \omega([X, X_i], X_1, \ldots, \hat{X}_i, \ldots, X_k) \\
= X[\omega(X_1, \ldots, X_k)] - \sum_{i=1}^{k} \omega(X_1, \ldots, [X, X_i], \ldots, X_k) = (L_X \omega)(X_1, \ldots, X_k),
\]
as claimed. \qed

Here is an example which demonstrates the usefulness of Cartan’s formula. Consider $S^1$ embedded in $\mathbb{R}^2$ via the parameterization $\psi : (0, 2\pi) \to \mathbb{R}^2$, where $\psi(t) = (\cos t, \sin t)$. Since $\psi'(t) = (-\sin t, \cos t) = (-y, x)$, the vector field $X = -y \partial/\partial x + x \partial/\partial y$ is tangent to $S^1$. Consider $\omega = -y dx + x dy$ as the restriction of the one form in $\mathbb{R}^2$ to $S^1$. We want to calculate $L_X \omega$ on $S^1$ using Cartan’s formula. This means we must first compute
\[
d\omega = d(-y) \wedge dx + d(x) \wedge dy = -dy \wedge dx + dx \wedge dy = 2 \, dx \wedge dy.
\]
Next we compute $i(X)d\omega$ as follows: by definition of $i(X)$,
\[
i(X)dx = dx(X) = dx \left( -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y} \right) = -y, \\
i(X)dy = dy(X) = dx \left( -y \frac{\partial}{\partial x} + x \frac{\partial}{\partial y} \right) = x.
\]
Then the anti-derivation property of $i(X)$ implies that
\[
i(X)d\omega = 2i(X)(dx \wedge dy) = 2 (i(X)dx \wedge dy - dx \wedge i(X)dy) = 2(-ydy - xdx). \quad (\star)
\]
To complete Cartan’s formula, we must calculate $d(i(X)\omega)$. Since $i(X)$ is $C^\infty(S^1)$ linear in $X$, we have
\[
i(X)\omega = -y \, i(X)dx + x \, i(X)dy = y^2 + x^2,
\]
and hence
\[
d(i(X)\omega) = d(y^2 + x^2) = 2y \, dy + 2x \, dx. \quad (\star\star)
\]
Cartan’s formula combines Lines $(\star)$ and $(\star\star)$ to give
\[
L_X\omega = i(X)\omega + d(i(X)\omega) = -2y \, dy - 2x \, dx + 2y \, dy + 2x \, dx = 0.
\]
The following proposition states more useful identities, some of which can be proved using Cartan’s formula:
Proposition 23.16. Let $M$ be a smooth manifold. For all vector fields $X, Y \in \mathfrak{X}(M)$, for all $\omega \in \mathcal{A}^k(M)$, we have

1. $L_X i(Y) - i(Y)L_X = i([X,Y])$.
2. $L_X L_Y \omega - L_Y L_X \omega = L_{[X,Y]} \omega$.
3. $L_X i(X) \omega = i(X)L_X \omega$.
4. $L_f X \omega = fL_X \omega + df \wedge i(X)\omega$, for all $f \in C^\infty(M)$.
5. For any diffeomorphism $\varphi: M \to N$, for all $Z \in \mathfrak{X}(N)$ and all $\beta \in \mathcal{A}^k(N)$,
   \[ \varphi^* L_Z \beta = L_{\varphi^* Z} \varphi^* \beta. \]

Finally here is a proposition about the Lie derivative of tensor fields. Obviously, tensor product and contraction of tensor fields are defined pointwise; that is

\[ (\alpha \otimes \beta)_p = \alpha_p \otimes \beta_p \]
\[ (c_{i,j}\alpha)_p = c_{i,j}\alpha_p, \]
for all $p \in M$, where $c_{i,j}$ is the contraction operator of Definition 21.7.

Proposition 23.17. Let $M$ be a smooth manifold. For every vector field $X \in \mathfrak{X}(M)$, the Lie derivative $L_X : \Gamma(M, T^\ast \otimes \mathcal{X}(M)) \to \Gamma(M, T^\ast \otimes \mathcal{X}(M))$ is the unique linear local operator satisfying the following properties:

1. $L_X f = X[f] = df(X)$, for all $f \in C^\infty(M)$.
2. $L_X Y = [X,Y]$, for all $Y \in \mathfrak{X}(M)$.
3. $L_X (\alpha \otimes \beta) = (L_X \alpha) \otimes \beta + \alpha \otimes (L_X \beta)$, for all tensor fields $\alpha \in \Gamma(M, T^{r_1,s_1}(M))$ and $\beta \in \Gamma(M, T^{r_2,s_2}(M))$; that is, $L_X$ is a derivation.
4. For all tensor fields $\alpha \in \Gamma(M, T^{r,s}(M))$, with $r, s > 0$, for every contraction operator $c_{i,j}$,
   \[ L_X (c_{i,j}(\alpha)) = c_{i,j}(L_X \alpha). \]

The proof of Proposition 23.17 can be found in Gallot, Hulin and Lafontaine [73] (Chapter 1). The following proposition is also useful:

Proposition 23.18. For every $(0,q)$-tensor $S \in \Gamma(M, (T^\ast)^{\otimes q}(M))$, we have

\[ (L_X S)(X_1, \ldots, X_q) = X[S(X_1, \ldots, X_q)] - \sum_{i=1}^q S(X_1, \ldots, [X,X_i], \ldots, X_q), \]
for all $X_1, \ldots, X_q, X \in \mathfrak{X}(M)$.

There are situations in differential geometry where it is convenient to deal with differential forms taking values in a vector space. This happens when we consider connections and the curvature form on vector bundles and principal bundles (see Chapter 29) and when we study Lie groups, where differential forms valued in a Lie algebra occur naturally.
23.4 Vector-Valued Differential Forms

This section contains background material for Chapter 29, especially for Section 29.1. Let us go back for a moment to differential forms defined on some open subset of $\mathbb{R}^n$. In Section 23.1, a differential form is defined as a smooth map $\omega: U \to \bigwedge^p (\mathbb{R}^n)^*$, and since we have a canonical isomorphism

$$\mu: \bigwedge^p (\mathbb{R}^n)^* \cong \text{Alt}^p(\mathbb{R}^n; \mathbb{R}),$$

such differential forms are real-valued. Now, let $F$ be any normed vector space, possibly infinite dimensional. Then, $\text{Alt}^p(\mathbb{R}^n; F)$ is also a normed vector space, and by Proposition 22.27, we have a canonical isomorphism

$$\mu_F: \left(\bigwedge^p (\mathbb{R}^n)^*\right) \otimes F \to \text{Alt}^p(\mathbb{R}^n; F)$$

defined on generators by

$$\mu_F((v_1^* \wedge \cdots \wedge v_p^*) \otimes f)(u_1, \ldots, u_p) = (\det(v_j^*(u_i)))f,$$

with $v_1^*, \ldots, v_p^* \in (\mathbb{R}^n)^*$, $u_1, \ldots, u_p \in \mathbb{R}^n$, and $f \in F$. Then, it is natural to define differential forms with values in $F$ as smooth maps $\omega: U \to \text{Alt}^p(\mathbb{R}^n; F)$. Actually, we can even replace $\mathbb{R}^n$ with any normed vector space, even infinite dimensional as in Cartan [37], but we do not need such generality for our purposes.

**Definition 23.14.** Let $F$ be any normed vector space. Given any open subset $U$ of $\mathbb{R}^n$, a smooth differential $p$-form on $U$ with values in $F$, for short a $p$-form on $U$, is any smooth function $\omega: U \to \text{Alt}^p(\mathbb{R}^n; F)$. The vector space of all $p$-forms on $U$ is denoted $\mathcal{A}^p(U; F)$. The vector space $\mathcal{A}^*(U; F) = \bigoplus_{p \geq 0} \mathcal{A}^p(U; F)$ is the set of differential forms on $U$ with values in $F$.

Observe that $\mathcal{A}^0(U; F) = C^\infty(U, F)$, the vector space of smooth functions on $U$ with values in $F$, and $\mathcal{A}^1(U; F) = C^\infty(U, \text{Hom}(\mathbb{R}^n, F))$, the set of smooth functions from $U$ to the set of linear maps from $\mathbb{R}^n$ to $F$. Also, $\mathcal{A}^p(U; F) = (0)$ for $p > n$.

Of course, we would like to have a “good” notion of exterior differential, and we would like as many properties of “ordinary” differential forms as possible to remain valid. As will see in our somewhat sketchy presentation, these goals can be achieved, except for some properties of the exterior product.

Using the isomorphism

$$\mu_F: \left(\bigwedge^p (\mathbb{R}^n)^*\right) \otimes F \to \text{Alt}^p(\mathbb{R}^n; F)$$
and Proposition 22.28, we obtain a convenient expression for differential forms in $\mathcal{A}^*(U;F)$. If $(e_1,\ldots,e_n)$ is any basis of $\mathbb{R}^n$ and $(e_1^*,\ldots,e_n^*)$ is its dual basis, then every differential $p$-form $\omega \in \mathcal{A}^p(U;F)$ can be written uniquely as

$$\omega(x) = \sum_I e_{i_1}^* \wedge \cdots \wedge e_{i_p}^* \otimes f_I(x) = \sum_I e_{i_1}^* \otimes f_I(x) \quad x \in U,$$

where each $f_I : U \to F$ is a smooth function on $U$.

As explained in Section 22.9, to express the above formula directly on alternating multilinear maps, define the product $\cdot : \text{Alt}^p(\mathbb{R}^n;F) \times F \to \text{Alt}^p(\mathbb{R}^n;F)$ as follows: For all $\omega \in \text{Alt}^p(\mathbb{R}^n;\mathbb{R})$ and all $f \in F$,

$$(\omega \cdot f)(u_1,\ldots,u_p) = \omega(u_1,\ldots,u_p)f,$$

for all $u_1,\ldots,u_p \in \mathbb{R}^n$. Then, it is immediately verified that for every $\omega \in (\wedge^p(\mathbb{R}^n)^*) \otimes F$ of the form

$$\omega = u_1^* \wedge \cdots \wedge u_p^* \otimes f,$$

we have

$$\mu_F(u_1^* \wedge \cdots \wedge u_p^* \otimes f) = \mu_F(u_1^* \wedge \cdots \wedge u_p^*) \cdot f.$$

By Proposition 22.29, the above property can be restated as the fact every differential $p$-form $\omega \in \mathcal{A}^p(U;F)$ can be written uniquely as

$$\omega(x) = \sum_I e_{i_1}^* \wedge \cdots \wedge e_{i_p}^* \cdot f_I(x), \quad x \in U,$$

where each $f_I : U \to F$ is a smooth function on $U$.

In order to define a multiplication on differential forms, we use a bilinear form $\Phi : F \times G \to H$; see Section 22.9. For every pair $(p,q)$, we define the multiplication

$$\wedge_{\Phi} : \left(\wedge^p(\mathbb{R}^n)^*) \otimes F \right) \times \left(\wedge^q(\mathbb{R}^n)^*) \otimes G \right) \to \left(\wedge^{p+q}(\mathbb{R}^n)^*) \otimes H \right)$$

by

$$(\alpha \otimes f) \wedge_{\Phi} (\beta \otimes g) = (\alpha \wedge \beta) \otimes \Phi(f,g).$$

We can also define a multiplication $\wedge_{\Phi}$ directly on alternating multilinear maps as follows: For $f \in \text{Alt}^p(\mathbb{R}^n;F)$ and $g \in \text{Alt}^q(\mathbb{R}^n;G)$,

$$(f \wedge_{\Phi} g)(u_1,\ldots,u_{p+q}) = \sum_{\sigma \in \text{shuffle}(p,q)} \text{sgn}(\sigma) \Phi(f(u_{\sigma(1)},\ldots,u_{\sigma(p)}),g(u_{\sigma(p+1)},\ldots,u_{\sigma(p+q)})),$$

where shuffle$(p,q)$ consists of all $(m,n)$-“shuffles,” that is, permutations $\sigma$ of $\{1,\ldots,p+q\}$ such that $\sigma(1) < \cdots < \sigma(p)$ and $\sigma(p+1) < \cdots < \sigma(p+q)$.
Then, we obtain a multiplication
\[ \wedge \Phi : \mathcal{A}^p(U; F) \times \mathcal{A}^q(U; G) \to \mathcal{A}^{p+q}(U; H), \]
defined so that for any differential forms \( \omega \in \mathcal{A}^p(U; F) \) and \( \eta \in \mathcal{A}^q(U; G) \),
\[ (\omega \wedge \Phi \eta)_x = \omega_x \wedge \Phi \eta_x, \quad x \in U. \]

In general, not much can be said about \( \wedge \Phi \), unless \( \Phi \) has some additional properties. In particular, \( \wedge \Phi \) is generally not associative, and there is no analog of Proposition 23.1. For simplicity of notation, we write \( \wedge \) for \( \wedge \Phi \).

Using \( \Phi \), we can also define a multiplication
\[ \cdot : \mathcal{A}^p(U; F) \times \mathcal{A}^0(U; G) \to \mathcal{A}^p(U; H) \]
given by
\[ (\omega \cdot f)_x(u_1, \ldots, u_p) = \Phi(\omega_x(u_1, \ldots, u_p), f(x)), \]
for all \( x \in U \), all \( f \in \mathcal{A}^0(U; G) = C^\infty(U, G) \), and all \( u_1, \ldots, u_p \in \mathbb{R}^n \). This multiplication will be used in the case where \( F = \mathbb{R} \) and \( G = H \) to obtain a normal form for differential forms.

Generalizing \( d \) is no problem. Observe that since a differential \( p \)-form is a smooth map \( \omega : U \to \text{Alt}^p(\mathbb{R}^n; F) \), its derivative is a map
\[ \omega' : U \to \text{Hom}(\mathbb{R}^n, \text{Alt}^p(\mathbb{R}^n; F)) \]
such that \( \omega'_x \) is a linear map from \( \mathbb{R}^n \) to \( \text{Alt}^p(\mathbb{R}^n; F) \) for every \( x \in U \). We can view \( \omega'_x \) as a multilinear map \( \omega'_x : (\mathbb{R}^n)^{p+1} \to F \) which is alternating in its last \( p \) arguments. As in Section 23.1, the exterior derivative \((d\omega)_x \) is obtained by making \( \omega'_x \) into an alternating map in all of its \( p + 1 \) arguments.

**Definition 23.15.** For every \( p \geq 0 \), the **exterior differential** \( d : \mathcal{A}^p(U; F) \to \mathcal{A}^{p+1}(U; F) \) is given by
\[ (d\omega)_x(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1} \omega'_x(u_i)(u_1, \ldots, \hat{u}_i, \ldots, u_{p+1}), \]
for all \( \omega \in \mathcal{A}^p(U; F) \) and all \( u_1, \ldots, u_{p+1} \in \mathbb{R}^n \), where the hat over the argument \( u_i \) means that it should be omitted.

Note that \( d \) depends on the vector space \( F \), so to be very precise we should denote \( d \) as \( d_F \). To keep notation simple, it is customary to drop the subscript \( F \).

For any smooth function \( f \in \mathcal{A}^0(U; F) = C^\infty(U, F) \), we get
\[ df_x(u) = f'_x(u). \]
Therefore, for smooth functions, the exterior differential \( df \) coincides with the usual derivative \( f' \). The important observation following Definition 23.3 also applies here. If \( x_i: U \to \mathbb{R} \) is the restriction of \( pr_i \) to \( U \), then \( x'_i \) is the constant map given by

\[
x'_i(x) = pr_i, \quad x \in U.
\]

It follows that \( dx_i = x'_i \) is the constant function with value \( pr_i = e^*_i \). As a consequence, every \( p \)-form \( \omega \) can be uniquely written as

\[
\omega_x = \sum_I dx_{i_1} \wedge \cdots \wedge dx_{i_p} \otimes f_I(x),
\]

where each \( f_I: U \to F \) is a smooth function on \( U \). Using the multiplication \( \cdot \) induced by the scalar multiplication in \( F \) \( (\Phi(\lambda, f) = \lambda f, \text{ with } \lambda \in \mathbb{R} \text{ and } f \in F) \), we see that every \( p \)-form \( \omega \) can be uniquely written as

\[
\omega = \sum_I dx_{i_1} \wedge \cdots \wedge dx_{i_p} \cdot f_I.
\]

As for real-valued functions, for any \( f \in \mathcal{A}^0(U; F) = C^\infty(U, F) \), we have

\[
df_x(u) = \sum_{i=1}^n u_i \frac{\partial f}{\partial x_i}(x) = \sum_{i=1}^n e^*_i(u) \frac{\partial f}{\partial x_i}(x),
\]

and so,

\[
df = \sum_{i=1}^n dx_i \cdot \frac{\partial f}{\partial x_i}.
\]

In general, Proposition 23.3 fails, unless \( F \) is finite-dimensional (see below). However for any arbitrary \( F \), a weak form of Proposition 23.3 can be salvaged. Again, let \( \Phi: F \times G \to H \) be a bilinear form, let \( \cdot : \mathcal{A}^p(U; F) \times \mathcal{A}^0(U; G) \to \mathcal{A}^p(U; H) \) be as defined before Definition 23.15, and let \( \wedge_\Phi \) be the wedge product associated with \( \Phi \). The following fact is proved in Cartan [37] (Section 2.4):

**Proposition 23.19.** For all \( \omega \in \mathcal{A}^p(U; F) \) and all \( f \in \mathcal{A}^0(U; G) \), we have

\[
d(\omega \cdot f) = (d\omega) \cdot f + \omega \wedge_\Phi df.
\]

Fortunately, \( d \circ d \) still vanishes, but this requires a completely different proof, since we can’t rely on Proposition 23.2 (see Cartan [37], Section 2.5). Similarly, Proposition 23.2 holds, but a different proof is needed.

**Proposition 23.20.** The composition \( \mathcal{A}^p(U; F) \xrightarrow{d} \mathcal{A}^{p+1}(U; F) \xrightarrow{d} \mathcal{A}^{p+2}(U; F) \) is identically zero for every \( p \geq 0 \); that is

\[
d \circ d = 0,
\]

which is an abbreviation for \( d^{p+1} \circ d^p = 0 \).
To generalize Proposition 23.2, we use Proposition 23.19 with the product $\cdot$ and the wedge product $\wedge$ induced by the bilinear form $\Phi$ given by scalar multiplication in $F$; that is, $\Phi(\lambda, f) = \lambda f$, for all $\lambda \in \mathbb{R}$ and all $f \in F$.

**Proposition 23.21.** For every $p$ form $\omega \in \mathcal{A}^p(U; F)$ with $\omega = dx_{i_1} \wedge \cdots \wedge dx_{i_p} \cdot f$, we have

$$d\omega = dx_{i_1} \wedge \cdots \wedge dx_{i_p} \wedge_F df,$$

where $\wedge$ is the usual wedge product on real-valued forms and $\wedge_F$ is the wedge product associated with scalar multiplication in $F$.

More explicitly, for every $x \in U$, for all $u_1, \ldots, u_{p+1} \in \mathbb{R}^n$, we have

$$(d\omega_x)(u_1, \ldots, u_{p+1}) = \sum_{i=1}^{p+1} (-1)^{i-1}(dx_{i_1} \wedge \cdots \wedge dx_{i_p})_x(u_1, \ldots, \hat{u}_i, \ldots, u_{p+1})df_x(u_i).$$

If we use the fact that

$$df = \sum_{i=1}^n dx_i \cdot \frac{\partial f}{\partial x_i},$$

we see easily that

$$d\omega = \sum_{j=1}^n dx_{i_1} \wedge \cdots \wedge dx_{i_p} \wedge dx_j \cdot \frac{\partial f}{\partial x_j},$$

the direct generalization of the real-valued case, except that the “coefficients” are functions with values in $F$.

The pull-back of forms in $\mathcal{A}^r(V, F)$ is defined as before. Luckily, Proposition 23.6 holds (see Cartan [37], Section 2.8).

**Proposition 23.22.** Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be two open sets and let $\varphi: U \to V$ be a smooth map. Then

(i) $\varphi^*(\omega \wedge \eta) = \varphi^*\omega \wedge \varphi^*\eta$, for all $\omega \in \mathcal{A}^p(V; F)$ and all $\eta \in \mathcal{A}^q(V; F)$.

(ii) $\varphi^*(f) = f \circ \varphi$, for all $f \in \mathcal{A}^0(V; F)$.

(iii) $d\varphi^*(\omega) = \varphi^*(d\omega)$, for all $\omega \in \mathcal{A}^p(V; F)$; that is, the following diagram commutes for all $p \geq 0$:

$$\begin{array}{ccc}
\mathcal{A}^p(V; F) & \xrightarrow{\varphi^*} & \mathcal{A}^p(U; F) \\
\downarrow d & & \downarrow d \\
\mathcal{A}^{p+1}(V; F) & \xrightarrow{\varphi^*} & \mathcal{A}^{p+1}(U; F).
\end{array}$$
Let us now consider the special case where $F$ has finite dimension $m$. Pick any basis $(f_1, \ldots, f_m)$ of $F$. Then, as every differential $p$-form $\omega \in \mathcal{A}^p(U; F)$ can be written uniquely as

$$\omega(x) = \sum_i e^*_{i_1} \wedge \cdots \wedge e^*_{i_p} \cdot f_I(x), \quad x \in U,$$

where each $f_I: U \to F$ is a smooth function on $U$, by expressing the $f_I$ over the basis $(f_1, \ldots, f_m)$, we see that $\omega$ can be written uniquely as

$$\omega = \sum_{i=1}^m \omega_i \cdot f_i,$$

where $\omega_1, \ldots, \omega_m$ are smooth real-valued differential forms in $\mathcal{A}^p(U; \mathbb{R})$, and we view $f_i$ as the constant map with value $f_i$ from $U$ to $F$. Then, as

$$\omega'(u) = \sum_{i=1}^m (\omega'_i)(u) f_i,$$

for all $u \in \mathbb{R}^n$, we see that

$$d\omega = \sum_{i=1}^m d\omega_i \cdot f_i.$$

Actually, because $d\omega$ is defined independently of bases, the $f_i$ do not need to be linearly independent; any choice of vectors and forms such that

$$\omega = \sum_{i=1}^k \omega_i \cdot f_i$$

will do.

Given a bilinear map $\Phi: F \times G \to H$, a simple calculation shows that for all $\omega \in \mathcal{A}^p(U; F)$ and all $\eta \in \mathcal{A}^q(U; G)$, we have

$$\omega \wedge \Phi \eta = \sum_{i=1}^m \sum_{j=1}^{m'} \omega_i \wedge \eta_j \cdot \Phi(f_i, g_j),$$

with $\omega = \sum_{i=1}^m \omega_i \cdot f_i$ and $\eta = \sum_{j=1}^{m'} \eta_j \cdot g_j$, where $(f_1, \ldots, f_m)$ is a basis of $F$ and $(g_1, \ldots, g_{m'})$ is a basis of $G$. From this and Proposition 23.3, it follows that Proposition 23.3 holds for finite-dimensional spaces.

**Proposition 23.23.** If $F, G, H$ are finite dimensional and $\Phi: F \times G \to H$ is a bilinear map, then for all $\omega \in \mathcal{A}^p(U; F)$ and all $\eta \in \mathcal{A}^q(U; G)$,

$$d(\omega \wedge \Phi \eta) = d\omega \wedge \Phi \eta + (-1)^p \omega \wedge \Phi d\eta.$$
On the negative side, in general, Proposition 23.1 still fails.

A special case of interest is the case where \( F = G = H = \mathfrak{g} \) is a Lie algebra, and \( \Phi(a, b) = [a, b] \) is the Lie bracket of \( \mathfrak{g} \). In this case, using a basis \((f_1, \ldots, f_r)\) of \( \mathfrak{g} \), if we write \( \omega = \sum_i \alpha_i f_i \) and \( \eta = \sum_j \beta_j f_j \), we have

\[
[\omega, \eta] = \sum_{i,j} \alpha_i \wedge \beta_j [f_i, f_j],
\]

where for simplicity of notation we dropped the subscript \( \Phi \) on \([\omega, \eta]\) and the multiplication sign \( \cdot \).

Let us figure out what \([\omega, \omega]\) is for a one-form \( \omega \in \mathcal{A}^1(U, \mathfrak{g}) \). By definition,

\[
[\omega, \omega] = \sum_{i,j} \omega_i \wedge \omega_j [f_i, f_j],
\]

so

\[
[\omega, \omega](u, v) = \sum_{i,j} (\omega_i(u) \wedge \omega_j(v))[f_i, f_j]
\]

\[
= \sum_{i,j} (\omega_i(u)\omega_j(v) - \omega_i(v)\omega_j(u))[f_i, f_j]
\]

\[
= \sum_{i,j} \omega_i(u)\omega_j(v)[f_i, f_j] - \sum_{i,j} \omega_i(v)\omega_j(u)[f_i, f_j]
\]

\[
= \left[ \sum_i \omega_i(u)f_i, \sum_j \omega_j(v)f_j \right] - \left[ \sum_i \omega_i(v)f_i, \sum_j \omega_j(u)f_j \right]
\]

\[
= [\omega(u), \omega(v)] - [\omega(v), \omega(u)]
\]

\[
= 2[\omega(u), \omega(v)].
\]

Therefore,

\[
[\omega, \omega](u, v) = 2[\omega(u), \omega(v)].
\]

Note that in general, \([\omega, \omega] \neq 0\), because \( \omega \) is vector valued. Of course, for real-valued forms, \([\omega, \omega] = 0\). Using the Jacobi identity of the Lie algebra, we easily find that

\[
[[\omega, \omega], \omega] = 0.
\]

The generalization of vector-valued differential forms to manifolds is no problem, except that some results involving the wedge product fail for the same reason that they fail in the case of forms on open subsets of \( \mathbb{R}^n \).

Given a smooth manifold \( M \) of dimension \( n \) and a vector space \( F \), the set \( \mathcal{A}^k(M; F) \) of \( \textit{differential} \ k-\textit{forms on} \ M \ \textit{with values in} \ F \) is the set of maps \( p \mapsto \omega_p \) with
\[ \omega_p \in \left( \bigwedge^k T^*_p M \right) \otimes F \cong \text{Alt}^k(T_p M; F), \] which vary smoothly in \( p \in M \). This means that the map

\[ p \mapsto \omega_p(X_1(p), \ldots, X_k(p)) \]

is smooth for all vector fields \( X_1, \ldots, X_k \in \mathfrak{X}(M) \). It can be shown (see Section 28.4) that

\[ \mathcal{A}^k(M; F) \cong \mathcal{A}^k(M) \otimes_{C^\infty(M)} C^\infty(M; F) \cong \text{Alt}^k_{C^\infty(M)}(\mathfrak{X}(M); C^\infty(M; F)), \]

which reduces to Proposition 23.12 when \( F = \mathbb{R} \).

The reader may want to carry out the verification that the theory generalizes to manifolds on her/his own. One result that will be used in the next section is that if \( \omega \in \mathcal{A}^1(M; F) \) is a vector valued one-form, then for any two vector fields \( X, Y \) on \( M \), we have

\[ d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y]). \]

In the next section, we consider some properties of differential forms on Lie groups.

## 23.5 Differential Forms on Lie Groups and Maurer-Cartan Forms

Given a Lie group \( G \), we saw in Section 15.2 that the set of left-invariant vector fields on \( G \) is isomorphic to the Lie algebra \( \mathfrak{g} = T_1 G \) of \( G \) (where 1 denotes the identity element of \( G \)). Recall that a vector field \( X \) on \( G \) is left-invariant iff

\[ d(L_a)_b(X_b) = X_{L_a b} = X_{ab}, \]

for all \( a, b \in G \). In particular, for \( b = 1 \), we get

\[ X_a = d(L_a)_1(X_1). \]

which shows that \( X \) is completely determined by its value at 1. The map \( X \mapsto X(1) \) is an isomorphism between left-invariant vector fields on \( G \) and \( \mathfrak{g} \).

The above suggests looking at left-invariant differential forms on \( G \). We will see that the set of left-invariant one-forms on \( G \) is isomorphic to \( \mathfrak{g}^* \), the dual of \( \mathfrak{g} \) as a vector space.

**Definition 23.16.** Given a Lie group \( G \), a differential form \( \omega \in \mathcal{A}^k(G) \) is left-invariant iff

\[ L_a^*\omega = \omega, \quad \text{for all } a \in G, \]

where \( L_a^*\omega \) is the pull-back of \( \omega \) by \( L_a \) (left multiplication by \( a \)). The left-invariant one-forms \( \omega \in \mathcal{A}^1(G) \) are also called *Maurer-Cartan forms*. 
For a one-form $\omega \in \mathcal{A}^1(G)$ left-invariance means that

$$(L_a^*\omega)_g(u) = \omega_{L_0g}(d(L_a)_g u) = \omega_{ag}(d(L_a)_g u) = \omega_g(u),$$

for all $a, g \in G$ and all $u \in T_gG$. For $a = g^{-1}$, we get

$$\omega_g(u) = \omega_1(d(L_{g^{-1}})_g u) = \omega_1(d(L_{g^{-1}})_g u),$$

which shows that $\omega_g$ is completely determined by $\omega_1$ (the value of $\omega_g$ at $g = 1$).

We claim that the map $\omega \mapsto \omega_1$ is an isomorphism between the set of left-invariant one-forms on $G$ and $g^*$. 

**Proof.** First, for any linear form $\alpha \in g^*$, the one-form $\alpha^L$ given by

$$\alpha^L_g(u) = \alpha(d(L_{g^{-1}})_g u)$$

is left-invariant, because

$$\begin{align*}
(L_h^*\alpha^L)_g(u) &= \alpha_{hg}(d(L_h)_g u) \\
&= \alpha(d(L_{hg}^{-1}d(L_h)_g u)) \\
&= \alpha(d(L_{hg}^{-1} \circ L_h)_g u) \\
&= \alpha(d(L_{g^{-1}})_g(u)) = \alpha^L_g(u).
\end{align*}$$

Second, we saw that for every one-form $\omega \in \mathcal{A}^1(G)$,

$$\omega_g(u) = \omega_1(d(L_{g^{-1}})_g u),$$

so $\omega_1 \in g^*$ is the unique element such that $\omega = \omega_1^L$, which shows that the map $\alpha \mapsto \alpha^L$ is an isomorphism whose inverse is the map $\omega \mapsto \omega_1$. 

Now, since every left-invariant vector field is of the form $X = u^L$ for some unique $u \in g$, where $u^L$ is the vector field given by $u^L(a) = d(L_a)_1 u$, and since

$$\omega_{ag}(d(L_a)_g u) = \omega_g(u),$$

for $g = 1$, we get $\omega_a(d(L_a)_1 u) = \omega_1(u)$; that is

$$\omega_a(X) = \omega_1(u), \quad a \in G,$$

which shows that $\omega(X)$ is a constant function on $G$. It follows that for every vector field $Y$ (not necessarily left-invariant),

$$Y[\omega(X)] = 0.$$

Recall that as a special case of Proposition 23.13 (generalized to vector-valued forms) we have

$$d\omega(X,Y) = X[\omega(Y)] - Y[\omega(X)] - \omega([X,Y]).$$
Consequently, for all left-invariant vector fields \( X, Y \) on \( G \), for every left-invariant one-form \( \omega \), we have

\[
d\omega(X, Y) = -\omega([X, Y]).
\]

If we identify the set of left-invariant vector fields on \( G \) with \( \mathfrak{g} \) and the set of left-invariant one-forms on \( G \) with \( \mathfrak{g}^* \), we have

\[
d\omega(X, Y) = -\omega([X, Y]), \quad \omega \in \mathfrak{g}^*, \ X, Y \in \mathfrak{g}.
\]

We summarize these facts in the following proposition:

**Proposition 23.24.** Let \( G \) be any Lie group.

1. The set of left-invariant one-forms on \( G \) is isomorphic to \( \mathfrak{g}^* \), the dual of the Lie algebra \( \mathfrak{g} \) of \( G \), via the isomorphism \( \omega \mapsto \omega_1 \).

2. For every left-invariant one form \( \omega \) and every left-invariant vector field \( X \), the value of the function \( \omega(X) \) is constant and equal to \( \omega_1(X_1) \).

3. If we identify the set of left-invariant vector fields on \( G \) with \( \mathfrak{g} \) and the set of left-invariant one-forms on \( G \) with \( \mathfrak{g}^* \), then

\[
d\omega(X, Y) = -\omega([X, Y]), \quad \omega \in \mathfrak{g}^*, \ X, Y \in \mathfrak{g}.
\]

Pick any basis \( X_1, \ldots, X_r \) of the Lie algebra \( \mathfrak{g} \), and let \( \omega_1, \ldots, \omega_r \) be the dual basis of \( \mathfrak{g}^* \). Then, there are some constants \( c_{ij}^k \) such that

\[
[X_i, X_j] = \sum_{k=1}^r c_{ij}^k X_k.
\]

The constants \( c_{ij}^k \) are called the structure constants of the Lie algebra \( \mathfrak{g} \). Observe that \( c_{ji}^k = -c_{ij}^k \).

As \( \omega_i([X_p, X_q]) = c_{pq}^i \) and \( d\omega_i(X, Y) = -\omega_i([X, Y]) \), we have

\[
\sum_{j,k} c_{jk}^i \omega_j \wedge \omega_k(X_p, X_q) = \sum_{j,k} c_{jk}^i (\omega_j(X_p)\omega_k(X_q) - \omega_j(X_q)\omega_k(X_p))
\]

\[
= \sum_{j,k} c_{jk}^i \omega_j(X_p)\omega_k(X_q) - \sum_{j,k} c_{jk}^i \omega_j(X_q)\omega_k(X_p)
\]

\[
= \sum_{j,k} c_{jk}^i \omega_j(X_p)\omega_k(X_q) + \sum_{j,k} c_{kj}^i \omega_j(X_q)\omega_k(X_p)
\]

\[
= c_{p,q}^i + c_{q,p}^i = 2c_{p,q}^i,
\]

so we get the equations

\[
d\omega_i = -\frac{1}{2} \sum_{j,k} c_{jk}^i \omega_j \wedge \omega_k,
\]
known as the *Maurer-Cartan equations*.

These equations can be neatly described if we use differential forms valued in \( g \). Let \( \omega_{MC} \) be the one-form given by

\[
(\omega_{MC})_g(u) = d(L_g^{-1})_g u, \quad g \in G, \; u \in T_g G.
\]

What \( \omega_{MC} \) does is to “bring back” a vector \( v \in T_g G \) to \( g = T_1 G \). The same computation that showed that \( \alpha^L \) is left-invariant if \( \alpha \in g \) shows that \( \omega_{MC} \) is left-invariant, and obviously \( (\omega_{MC})_1 = \text{id} \).

**Definition 23.17.** Given any Lie group \( G \), the *Maurer-Cartan form on \( G \)* is the \( g \)-valued differential 1-form \( \omega_{MC} \in \mathcal{A}^1(G, g) \) given by

\[
(\omega_{MC})_g = d(L_g^{-1})_g, \quad g \in G.
\]

The same argument that we used to prove property (2) of Proposition 23.24 shows that for every left-invariant one-form \( \omega \in \mathcal{A}^1(G, g) \) and every vector field \( X \in \mathfrak{X}(G) \), the value of the function \( \omega(X) \) is constant and equal to \( \omega_1(X_1) \). In particular, this holds for the Maurer-Cartan form \( \omega_{MC} \). As in Section 23.4, the Lie bracket on \( g \) induces a multiplication

\[
[-, -]: \mathcal{A}^p(G; g) \times \mathcal{A}^q(G; g) \to \mathcal{A}^{p+q}(G; g)
\]

given by

\[
[\omega, \eta] = \sum_{ij} \alpha_i \wedge \beta_j \cdot [f_i, f_j],
\]

where \((f_1, \ldots, f_r)\) is a basis of \( g \) and where \( \omega = \sum_i \alpha_i \cdot f_i \) and \( \eta = \sum_j \beta_j \cdot f_j \). Using the same proof, we obtain the equation

\[
[\omega, \omega](X, Y) = 2[\omega(X), \omega(Y)].
\]

Recall that for every \( g \in G \), conjugation by \( g \) is the map given by \( a \mapsto gag^{-1} \); that is, \( a \mapsto (L_g \circ R_{g^{-1}})a \), and the adjoint map \( \text{Ad}(g): g \to g \) associated with \( g \) is the derivative of \( \text{Ad}_g = L_g \circ R_{g^{-1}} \) at 1; that is, we have

\[
\text{Ad}(g)(u) = d(\text{Ad}_g)_1(u), \quad u \in g.
\]

Furthermore, it is obvious that \( L_g \) and \( R_h \) commute.

**Proposition 23.25.** Given any Lie group \( G \), for all \( g \in G \), the Maurer-Cartan form \( \omega_{MC} \) has the following properties:

1. \( (\omega_{MC})_1 = \text{id}_g \).
2. For all \( g \in G \),

\[
R_g^* \omega_{MC} = \text{Ad}(g^{-1}) \circ \omega_{MC}.
\]
(3) The 2-form $d\omega_{MC} \in \mathcal{A}^2(G, \mathfrak{g})$ satisfies the Maurer-Cartan equation

$$d\omega_{MC} = -\frac{1}{2}[\omega_{MC}, \omega_{MC}].$$

Proof. Property (1) is obvious.

(2) For simplicity of notation, if we write $\omega = \omega_{MC}$, then

$$\begin{align*}
(R^*_g \omega)_h &= \omega_h \circ d(R_g)_h \\
&= d(L^{-1}_h \omega_h) \circ d(R_g)_h \\
&= d(L^{-1}_h \circ R_g)_h \\
&= d((L_h \circ L_g)^{-1} \circ R_g)_h \\
&= d(L^{-1}_g \circ L^{-1}_h \circ R_g)_h \\
&= d(L_{g^{-1}} \circ R_g)_1 \circ d(L^{-1}_h)_h \\
&= \text{Ad}(g^{-1}) \circ \omega_h,
\end{align*}$$

as claimed.

(3) We can easily express $\omega_{MC}$ in terms of a basis of $\mathfrak{g}$. If $X_1, \ldots, X_r$ is a basis of $\mathfrak{g}$ and $\omega_1, \ldots, \omega_r$ is the dual basis, then by Proposition 23.24 (2) and part (1) of Proposition 23.25, we have $\omega_{MC}(X_i) = (\omega_{MC})_1(X_i) = X_i$, for $i = 1, \ldots, r$, so $\omega_{MC}$ is given by

$$\omega_{MC} = \omega_1 \cdot X_1 + \cdots + \omega_r \cdot X_r,$$

under the usual identification of left-invariant vector fields (resp. left-invariant one forms) with elements of $\mathfrak{g}$ (resp. elements of $\mathfrak{g}^*$). Then, we have

$$d\omega_{MC} = d\omega_1 \cdot X_1 + \cdots + d\omega_r \cdot X_r.$$

We will use the Maurer-Cartan equations

$$d\omega_i = -\frac{1}{2} \sum_{j,k} c^{ij}_{jk} \omega_j \wedge \omega_k$$

to obtain the desired equation. Using the fact that the $c^{ij}_{jk}$ are skew-symmetric in $j, k$, for all $u, v \in \mathfrak{g}$, we have

$$[\omega_{MC}, \omega_{MC}](u, v) = \left[ \sum_j \omega_j(u) \cdot X_j, \sum_k \omega_j(v) \cdot X_k \right]$$

$$= \sum_{i,j,k} \omega_j(u) \omega_k(v) c^{ij}_{jk} \cdot X_i$$

$$= \sum_{i,j,k} c^{ij}_{jk} (\omega_j \wedge \omega_k)(u, v) \cdot X_i$$

$$= -2 \sum_i d\omega_i(u, v) \cdot X_i$$

$$= -2d\omega_{MC}(u, v),$$
23.5. DIFFERENTIAL FORMS ON LIE GROUPS

\[ d\omega_{MC} = -\frac{1}{2}[\omega_{MC},\omega_{MC}], \]

as claimed. \qed

In the case of a matrix group \( G \subseteq \text{GL}(n, \mathbb{R}) \), it is easy to see that the Maurer-Cartan form is given explicitly by

\[ \omega_{MC}(v) = g^{-1}v, \quad v \in T_g G, \; g \in G. \]

Since \( T_g G \) is isomorphic to \( gg \), we have

\[ \omega_{MC}(gv) = v, \quad v \in g. \]

The above expression suggests that, with some abuse of notation, \( \omega_{MC} \) may be denoted by \( g^{-1}dg \), where \( g = (g_{ij}) \) and where \( dg \) is an abbreviation for the \( n \times n \) matrix \( (dg_{ij}) \). Thus, \( \omega_{MC} \) is a kind of logarithmic derivative of the identity. For \( n = 2 \), if we write

\[ g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \]

we get

\[ \omega_{MC} = \frac{1}{\alpha \delta - \beta \gamma} \begin{pmatrix} \delta d\alpha - \beta d\gamma & \delta d\beta - \beta d\delta \\ -\gamma d\alpha + \alpha d\gamma & -\gamma d\beta + \alpha d\delta \end{pmatrix}. \]

Remarks:

1. The quantity, \( d\omega_{MC} + \frac{1}{2}[\omega_{MC},\omega_{MC}] \) is the curvature of the connection \( \omega_{MC} \) on \( G \). The Maurer-Cartan equation says that the curvature of the Maurer-Cartan connection is zero. We also say that \( \omega_{MC} \) is a flat connection.

2. As \( d\omega_{MC} = -\frac{1}{2}[\omega_{MC},\omega_{MC}] \), we get

\[ d[\omega_{MC},\omega_{MC}] = 0, \]

which yields

\[ [[\omega_{MC},\omega_{MC}],\omega_{MC}] = 0. \]

It is easy to show that the above expresses the Jacobi identity in \( g \).

3. As in the case of real-valued one-forms, for every left-invariant one-form \( \omega \in \mathcal{A}^1(G, g) \), we have

\[ \omega_g(u) = \omega_1(d(L_g^{-1})_g)u = \omega_1((\omega_{MC})_g)u, \]

for all \( g \in G \) and all \( u \in T_g G \), and where \( \omega_1: g \to g \) is a linear map. Consequently, there is a bijection between the set of left-invariant one-forms in \( \mathcal{A}^1(G, g) \) and \( \text{Hom}(g, g) \).
The Maurer-Cartan form can be used to define the *Darboux derivative* of a map \( f: M \to G \), where \( M \) is a manifold and \( G \) is a Lie group. The Darboux derivative of \( f \) is the \( g \)-valued one-form \( \omega_f \in \mathcal{A}^1(M, g) \) on \( M \) given by

\[
\omega_f = f^*\omega_{MC}.
\]

Then, it can be shown that when \( M \) is connected, if \( f_1 \) and \( f_2 \) have the same Darboux derivative \( \omega_{f_1} = \omega_{f_2} \), then \( f_2 = L_g \circ f_1 \), for some \( g \in G \). Elie Cartan also characterized which \( g \)-valued one-forms on \( M \) are Darboux derivatives \((d\omega + \frac{1}{2}[\omega, \omega] = 0 \text{ must hold})\). For more on Darboux derivatives, see Sharpe [162] (Chapter 3) and Malliavin [120] (Chapter III).
Chapter 24

Integration on Manifolds

24.1 Orientation of Manifolds

Although the notion of orientation of a manifold is quite intuitive it is technically rather subtle. We restrict our discussion to smooth manifolds (the notion of orientation can also be defined for topological manifolds, but more work is involved).

Intuitively, a manifold $M$ is orientable if it is possible to give a consistent orientation to its tangent space $T_pM$ at every point $p \in M$. So, if we go around a closed curve starting at $p \in M$, when we come back to $p$, the orientation of $T_pM$ should be the same as when we started. For example, if we travel on a Möbius strip (a manifold with boundary) dragging a coin with us, we will come back to our point of departure with the coin flipped. Try it; see Figure 24.3 for an illustration.

To be rigorous, we have to say what it means to orient $T_pM$ (a vector space) and what consistency of orientation means. We begin by quickly reviewing the notion of orientation of a vector space. Let $E$ be a vector space of dimension $n$. If $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are two bases of $E$, a basic and crucial fact of linear algebra says that there is a unique linear map $g$ mapping each $u_i$ to the corresponding $v_i$ (i.e., $g(u_i) = v_i$, $i = 1, \ldots, n$). Then look at the determinant $\det(g)$ of this map. We know that $\det(g) = \det(P)$, where $P$ is the matrix whose $j$-th columns consist of the coordinates of $v_j$ over the basis $u_1, \ldots, u_n$. Either $\det(g)$ is negative, or it is positive. Thus, we define an equivalence relation on bases by saying that two bases have the same orientation iff the determinant of the linear map sending the first basis to the second has positive determinant. An orientation of $E$ is the choice of one of the two equivalence classes, which amounts to picking some basis as an orientation frame. For $E = \mathbb{R}$, an orientation is given by $e_1$ or $-e_1$. Such an orientation is visualized as either right or left translation from the origin. For $E = \mathbb{R}^2$, an orientation is given by $(e_1, e_2)$ or $(e_2, e_1)$, i.e. either counterclockwise or clockwise rotation about the origin. For $E = \mathbb{R}^3$, the orientation is represented by $(e_1, e_2, e_3)$ or $(e_2, e_1, e_3)$, namely the right-handed or left handed orientation of the $i, j, k$ axis system. See Figure 24.1.
The above definition is perfectly fine, but it turns out that it is more convenient, in the long term, to use a definition of orientation in terms of differential forms and the exterior algebra $\wedge^n E^*$. This approach is especially useful when defining the notion of integration on a manifold. We observe that two bases $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ have the same orientation iff

$$\omega(u_1, \ldots, u_n) \text{ and } \omega(v_1, \ldots, v_n)$$

have the same sign for all $\omega \in \wedge^n E^* - \{0\}$ (where 0 denotes the zero n-form). As $\wedge^n E^*$ is one-dimensional, picking an orientation of $E$ is equivalent to picking a generator (a one-element basis) $\omega$ of $\wedge^n E^*$, and to say that $u_1, \ldots, u_n$ has positive orientation iff $\omega(u_1, \ldots, u_n) > 0$.

Given an orientation (say, given by $\omega \in \wedge^n E^*$) of $E$, a linear map $f: E \to E$ is orientation preserving iff $\omega(f(u_1), \ldots, f(u_n)) > 0$ whenever $\omega(u_1, \ldots, u_n) > 0$ (or equivalently, iff $\det(f) > 0$).

To define the orientation of an n-dimensional manifold $M$ we use charts. Given any $p \in M$, for any chart $(U, \varphi)$ at $p$, the tangent map $d\varphi_p^{-1}: \mathbb{R}^n \to T_p M$ makes sense. If $(e_1, \ldots, e_n)$ is the standard basis of $\mathbb{R}^n$, as it gives an orientation to $\mathbb{R}^n$, we can orient $T_p M$ by giving it the orientation induced by the basis $d\varphi_p^{-1}(e_1), \ldots, d\varphi_p^{-1}(e_n)$. The consistency of orientations of the $T_p M$’s is given by the overlapping of charts. See Figure 24.2.
We require that the Jacobian determinants of all \( \varphi_j \circ \varphi_i^{-1} \) have the same sign whenever \((U_i, \varphi_i)\) and \((U_j, \varphi_j)\) are any two overlapping charts. Thus, we are led to the definition below. All definitions and results stated in the rest of this section apply to manifolds with or without boundary.

**Definition 24.1.** Given a smooth manifold \( M \) of dimension \( n \), an orientation atlas of \( M \) is any atlas so that the transition maps \( \varphi_j^i = \varphi_j \circ \varphi_i^{-1} \) (from \( \varphi_i(U_i \cap U_j) \) to \( \varphi_j(U_i \cap U_j) \)) all have a positive Jacobian determinant for every point in \( \varphi_i(U_i \cap U_j) \). A manifold is orientable iff its has some orientation atlas.

We should mention that not every manifold orientable. The open Mobius strip, i.e. the Mobius strip with circle boundary removed, is not orientable, as demonstrated in Figure 24.3.

Definition 24.1 can be hard to check in practice and there is an equivalent criterion is terms of \( n \)-forms which is often more convenient. The idea is that a manifold of dimension \( n \) is orientable iff there is a map \( p \mapsto \omega_p \), assigning to every point \( p \in M \) a nonzero \( n \)-form \( \omega_p \in \bigwedge^n T^*_p M \), so that this map is smooth.
Figure 24.3: The Mobius strip does not have a consistent orientation. The frame starting at 1 is reversed when traveling around the loop to 1′.

**Definition 24.2.** If $M$ is an $n$-dimensional manifold, recall that a smooth section $\omega \in \Gamma(M, \wedge^n T^* M)$ is called a (smooth) $n$-form. An $n$-form $\omega$ is a nowhere-vanishing $n$-form on $M$ or volume form on $M$ iff $\omega_p$ is a nonzero form for every $p \in M$. This is equivalent to saying that $\omega_p(u_1, \ldots, u_n) \neq 0$, for all $p \in M$ and all bases $u_1, \ldots, u_n$, of $T_p M$.

The determinant function $(u_1, \ldots, u_n) \mapsto \det(u_1, \ldots, u_n)$ where the $u_i$ are expressed over the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$, is a volume form on $\mathbb{R}^n$. We will denote this volume form by $\omega_{\mathbb{R}^n} = dx_1 \wedge \cdots \wedge dx_n$. Observe the justification for the term volume form: the quantity $\det(u_1, \ldots, u_n)$ is indeed the (signed) volume of the parallelepiped

$$\{\lambda_1 u_1 + \cdots + \lambda_n u_n | 0 \leq \lambda_i \leq 1, 1 \leq i \leq n\}.$$ 

A volume form on the sphere $S^n \subseteq \mathbb{R}^{n+1}$ is obtained as follows:

$$\omega_{S^n}(u_1, \ldots, u_n) = \det(p, u_1, \ldots, u_n),$$

where $p \in S^n$ and $u_1, \ldots, u_n \in T_p S^n$. As the $u_i$ are orthogonal to $p$, this is indeed a volume form.

Observe that if $f$ is a smooth function on $M$ and $\omega$ is any $n$-form, then $f\omega$ is also an $n$-form.

One checks immediately that $h^* \omega$ is indeed an $n$-form on $M$. More interesting is the following Proposition:

**Proposition 24.1.** (a) If $h: M \rightarrow N$ is a local diffeomorphism of manifolds, where $\dim M = \dim N = n$, and $\omega \in A^n(N)$ is a volume form on $N$, then $h^* \omega$ is a volume form on $M$. (b) Assume $M$ has a volume form $\omega$. For every $n$-form $\eta \in A^n(M)$, there is a unique smooth function $f \in C^\infty(M)$ so that $\eta = f \omega$. If $\eta$ is a volume form, then $f(p) \neq 0$ for all $p \in M$. 
Proof. (a) By definition,
\[ h^*\omega_p(u_1, \ldots, u_n) = \omega_{h(p)}(dh_p(u_1), \ldots, dh_p(u_n)), \]
for all \( p \in M \) and all \( u_1, \ldots, u_n \in T_pM \). As \( h \) is a local diffeomorphism, \( d_ph \) is a bijection for every \( p \). Thus, if \( u_1, \ldots, u_n \) is a basis, then so is \( dh_p(u_1), \ldots, dh_p(u_n) \), and as \( \omega \) is nonzero at every point for every basis, \( h^*\omega_p(u_1, \ldots, u_n) \neq 0 \).

(b) Pick any \( p \in M \) and let \( (U, \varphi) \) be any chart at \( p \). As \( \varphi \) is a diffeomorphism, by (a), we see that \( \varphi^{-1*}\omega \) is a volume form on \( \varphi(U) \). But then, it is easy to see that \( \varphi^{-1*}\eta = g\varphi^{-1*}\omega \), for some unique smooth function \( g \) on \( \varphi(U) \), and so \( \eta = f_U\omega \), for some unique smooth function \( f_U \) on \( U \). For any two overlapping charts \( (U_i, \varphi_i) \) and \( (U_j, \varphi_j) \), for every \( p \in U_i \cap U_j \), for every basis \( u_1, \ldots, u_n \) of \( T_pM \), we have
\[ \eta_p(u_1, \ldots, u_n) = f_i(p)\omega_p(u_1, \ldots, u_n) = f_j(p)\omega_p(u_1, \ldots, u_n), \]
and as \( \omega_p(u_1, \ldots, u_n) \neq 0 \), we deduce that \( f_i \) and \( f_j \) agree on \( U_i \cap U_j \). But, then the \( f_i \)'s patch on the overlaps of the cover \( \{U_i\} \) of \( M \), and so there is a smooth function \( f \) defined on the whole of \( M \) and such that \( f \mid U_i = f_i \). As the \( f_i \)'s are unique, so is \( f \). If \( \eta \) is a volume form, then \( \eta_p \) does not vanish for all \( p \in M \), and since \( \omega_p \) is also a volume form, \( \omega_p \) does not vanish for all \( p \in M \), so \( f(p) \neq 0 \) for all \( p \in M \). \( \square \)

Remark: If \( h_1 \) and \( h_2 \) are smooth maps of manifolds, it is easy to prove that
\[ (h_2 \circ h_1)^* = h_1^* \circ h_2^*, \]
and that for any smooth map \( h: M \to N \),
\[ h^*(f\omega) = (f \circ h)h^*\omega, \]
where \( f \) is any smooth function on \( N \) and \( \omega \) is any \( n \)-form on \( N \).

The connection between Definition 24.1 and volume forms is given by the following important theorem whose proof contains a wonderful use of partitions of unity.

Theorem 24.2. A smooth manifold (Hausdorff and second-countable) is orientable iff it possesses a volume form.

Proof. First assume that a volume form \( \omega \) exists on \( M \), and say \( n = \dim M \). For any atlas \( \{(U_i, \varphi_i)\}_i \) of \( M \), by Proposition 24.1, each \( n \)-form \( \varphi_i^{-1*}\omega \) is a volume form on \( \varphi_i(U_i) \subseteq \mathbb{R}^n \), and
\[ \varphi_i^{-1*}\omega = f_i\omega_{\mathbb{R}^n}, \]
for some smooth function \( f_i \) never zero on \( \varphi_i(U_i) \), where \( \omega_{\mathbb{R}^n} \) is the volume form on \( \mathbb{R}^n \). By composing \( \varphi_i \) with an orientation-reversing linear map if necessary, we may assume that for this new atlas, \( f_i > 0 \) on \( \varphi_i(U_i) \). We claim that the family \( (U_i, \varphi_i)_i \) is an orientation atlas.
This is because, on any (nonempty) overlap \( U_i \cap U_j \), as \( \omega = \varphi_j^* (f_j \omega_{\mathbb{R}^n}) \) and 
\( (\varphi_j \circ \varphi_i^{-1})^* = (\varphi_i^{-1})^* \circ \varphi_j^* \), we have
\[
(\varphi_j \circ \varphi_i^{-1})^* (f_j \omega_{\mathbb{R}^n}) = (\varphi_i^{-1})^* \circ \varphi_j^* (f_j \omega_{\mathbb{R}^n}) = (\varphi_i^{-1})^* \omega = f_i \omega_{\mathbb{R}^n},
\]
and by the definition of pull-backs, we see that for every \( x \in \varphi_i(U_i \cap U_j) \), if we let \( y = \varphi_j \circ \varphi_i^{-1}(x) \), then
\[
\begin{align*}
  f_i(x)(\omega_{\mathbb{R}^n})_x(e_1, \ldots, e_n) &= (\varphi_j \circ \varphi_i^{-1})^* (f_j \omega_{\mathbb{R}^n})(e_1, \ldots, e_n) \\
  &= f_j(y)(\omega_{\mathbb{R}^n})_y(d(\varphi_j \circ \varphi_i^{-1})_x(e_1), \ldots, d(\varphi_j \circ \varphi_i^{-1})_x(e_n)) \\
  &= f_j(y)J((\varphi_j \circ \varphi_i^{-1})_x)(\omega_{\mathbb{R}^n})_y(e_1, \ldots, e_n),
\end{align*}
\]
where \( e_1, \ldots, e_n \) is the standard basis of \( \mathbb{R}^n \) and \( J((\varphi_j \circ \varphi_i^{-1})_x) \) is the Jacobian determinant of \( \varphi_j \circ \varphi_i^{-1} \) at \( x \). As both \( f_j(y) > 0 \) and \( f_i(x) > 0 \), we have \( J((\varphi_j \circ \varphi_i^{-1})_x) > 0 \), as desired.

Conversely, assume that \( J((\varphi_j \circ \varphi_i^{-1})_x) > 0 \), for all \( x \in \varphi_i(U_i \cap U_j) \), whenever \( U_i \cap U_j \neq \emptyset \). We need to make a volume form on \( M \). For each \( U_i \), let
\[
\omega_i = \varphi_i^* \omega_{\mathbb{R}^n},
\]
where \( \omega_{\mathbb{R}^n} \) is the volume form on \( \mathbb{R}^n \). As \( \varphi_i \) is a diffeomorphism, by Proposition 24.1, we see that \( \omega_i \) is a volume form on \( U_i \). Then, if we apply Theorem 9.4, we can find a partition of unity \( \{ f_i \} \) subordinate to the cover \( \{ U_i \} \), with the same index set. Let,
\[
\omega = \sum_i f_i \omega_i.
\]
We claim that \( \omega \) is a volume form on \( M \).

It is clear that \( \omega \) is an \( n \)-form on \( M \). Now, since every \( p \in M \) belongs to some \( U_i \), check that on \( \varphi_i(U_i) \), we have
\[
\varphi_i^{-1}^* \omega = \sum_{j \in \text{finite set}} \varphi_i^{-1}^* (f_j \omega_j) = \sum_{j \in \text{finite set}} \varphi_i^{-1}^* (f_j \varphi_j^* \omega_{\mathbb{R}^n}) = \sum_{j \in \text{finite set}} (f_j \circ \varphi_i^{-1})((\varphi_i^{-1})^* \circ \varphi_j^*) \omega_{\mathbb{R}^n} = \sum_{j \in \text{finite set}} (f_j \circ \varphi_i^{-1})(\varphi_j \circ \varphi_i^{-1})^* \omega_{\mathbb{R}^n} = \left( \sum_{j \in \text{finite set}} (f_j \circ \varphi_i^{-1})J(\varphi_j \circ \varphi_i^{-1}) \right) \omega_{\mathbb{R}^n},
\]
and this sum is strictly positive because the Jacobian determinants are positive, and as \( \sum_j f_j = 1 \) and \( f_j \geq 0 \), some term must be strictly positive. Therefore, \( \varphi_i^{-1}^* \omega \) is a volume form on \( \varphi_i(U_i) \), so \( \varphi_i^* \varphi_i^{-1}^* \omega = \omega \) is a volume form on \( U_i \). As this holds for all \( U_i \), we conclude that \( \omega \) is a volume form on \( M \). \( \square \)
Since we showed that there is a volume form on the sphere $S^n$, by Theorem 24.2, the sphere $S^n$ is orientable. It can be shown that the projective spaces $\mathbb{RP}^n$ are non-orientable iff $n$ is even, and thus orientable iff $n$ is odd. In particular, $\mathbb{RP}^2$ is not orientable. Also, even though $M$ may not be orientable, its tangent bundle $T(M)$ is always orientable! (Prove it). It is also easy to show that if $f: \mathbb{R}^{n+1} \to \mathbb{R}$ is a smooth submersion, then $M = f^{-1}(0)$ is a smooth orientable manifold. Another nice fact is that every Lie group is orientable.

By Proposition 24.1 (b), given any two volume forms $\omega_1$ and $\omega_2$ on a manifold $M$, there is a function $f: M \to \mathbb{R}$ never 0 on $M$, such that $\omega_2 = f \omega_1$. This fact suggests the following definition:

**Definition 24.3.** Given an orientable manifold $M$, two volume forms $\omega_1$ and $\omega_2$ on $M$ are equivalent iff $\omega_2 = f \omega_1$ for some smooth function $f: M \to \mathbb{R}$, such that $f(p) > 0$ for all $p \in M$. An orientation of $M$ is the choice of some equivalence class of volume forms on $M$, and an oriented manifold is a manifold together with a choice of orientation. If $M$ is a manifold oriented by the volume form $\omega$, for every $p \in M$, a basis $(b_1, \ldots, b_n)$ of $T_pM$ is positively oriented iff $\omega_p(b_1, \ldots, b_n) > 0$, else it is negatively oriented (where $n = \dim(M)$).

If $M$ is an orientable manifold, for any two volume forms $\omega_1$ and $\omega_2$ on $M$, as $\omega_2 = f \omega_1$ for some function $f$ on $M$ which is never zero, $f$ has a constant sign on every connected component of $M$. Consequently, a connected orientable manifold has two orientations.

We will also need the notion of orientation-preserving diffeomorphism.

**Definition 24.4.** Let $h: M \to N$ be a diffeomorphism of oriented manifolds $M$ and $N$, of dimension $n$, and say the orientation on $M$ is given by the volume form $\omega_1$ while the orientation on $N$ is given by the volume form $\omega_2$. We say that $h$ is orientation preserving iff $h^* \omega_2$ determines the same orientation of $M$ as $\omega_1$.

Using Definition 24.4, we can define the notion of a positive atlas.

**Definition 24.5.** If $M$ is a manifold oriented by the volume form $\omega$, an atlas for $M$ is positive iff for every chart $(U, \varphi)$, the diffeomorphism $\varphi: U \to \varphi(U)$ is orientation preserving, where $U$ has the orientation induced by $M$ and $\varphi(U) \subseteq \mathbb{R}^n$ has the orientation induced by the standard orientation on $\mathbb{R}^n$ (with $\dim(M) = n$).

The proof of Theorem 24.2 shows

**Proposition 24.3.** If a manifold $M$ has an orientation atlas, then there is a uniquely determined orientation on $M$ such that this atlas is positive.
24.2 Volume Forms on Riemannian Manifolds and Lie Groups

Recall from Section 10.2 that a smooth manifold \( M \) is a Riemannian manifold iff the vector bundle \( TM \) has a Euclidean metric. This means that there is a family \( \langle \cdot, \cdot \rangle_p \) of inner products on the tangent spaces \( T_pM \), such that \( \langle \cdot, \cdot \rangle_p \) depends smoothly on \( p \), which can be expressed by saying that the maps

\[
x \mapsto \langle d\varphi^{-1}_x(e_i), d\varphi^{-1}_x(e_j) \rangle_{\varphi^{-1}(x)}, \quad x \in \varphi(U), \; 1 \leq i, j \leq n
\]

are smooth, for every chart \((U, \varphi)\) of \( M \), where \((e_1, \ldots, e_n)\) is the canonical basis of \( \mathbb{R}^n \). We let

\[
g_{ij}(x) = \langle d\varphi^{-1}_x(e_i), d\varphi^{-1}_x(e_j) \rangle_{\varphi^{-1}(x)},
\]

and we say that the \( n \times n \) matrix \((g_{ij}(x))\) is the local expression of the Riemannian metric on \( M \) at \( x \) in the coordinate patch \((U, \varphi)\).

If a Riemannian manifold \( M \) is orientable, then there is a volume form on \( M \) with some special properties.

**Proposition 24.4.** Let \( M \) be a Riemannian manifold with \( \dim(M) = n \). If \( M \) is orientable, then there is a uniquely determined volume form \( \text{Vol}_M \) on \( M \) with the following property: For every \( p \in M \), for every positively oriented orthonormal basis \((b_1, \ldots, b_n)\) of \( T_pM \), we have

\[
(\text{Vol}_M)_p(b_1, \ldots, b_n) = 1.
\]

Furthermore, if the above equation holds then in every orientation preserving local chart \((U, \varphi)\), we have

\[
((\varphi^{-1})^*\text{Vol}_M)_q = \sqrt{\det(g_{ij}(q))} \, dx_1 \wedge \cdots \wedge dx_n, \quad q \in \varphi(U).
\]

**Proof.** Say the orientation of \( M \) is given by \( \omega \in \mathcal{A}^n(M) \). For any two positively oriented orthonormal bases \((b_1, \ldots, b_n)\) and \((b'_1, \ldots, b'_n)\) in \( T_pM \), by expressing the second basis over the first, there is an orthogonal matrix \( C = (c_{ij}) \) so that

\[
b'_i = \sum_{j=1}^n c_{ij} b_j.
\]

We have

\[
\omega_p(b'_1, \ldots, b'_n) = \det(C) \omega_p(b_1, \ldots, b_n),
\]

and as these bases are positively oriented, we conclude that \( \det(C) = 1 \) (as \( C \) is orthogonal, \( \det(C) = \pm 1 \)). As a consequence, we have a well-defined function \( \rho: M \to \mathbb{R} \) with \( \rho(p) > 0 \) for all \( p \in M \), such that

\[
\rho(p) = \omega_p(b_1, \ldots, b_n)
\]
for every positively oriented orthonormal basis \((b_1, \ldots, b_n)\) of \(T_pM\). If we can show that \(\rho\) is smooth, then \((\Vol(M)_p = \frac{1}{\rho(p)} \omega_p\) is the required volume form.

Let \((U, \varphi)\) be a positively oriented chart and consider the vector fields \(X_j\) on \(\varphi(U)\) given by
\[
X_j(q) = d\varphi_q^{-1}(e_j), \quad q \in \varphi(U), \quad 1 \leq j \leq n.
\]
Then, \((X_1(q), \ldots, X_n(q))\) is a positively oriented basis of \(T_{\varphi^{-1}(q)}\). If we apply Gram-Schmidt orthogonalization, we get an upper triangular matrix \(A(q) = (a_{ij}(q))\) of smooth functions on \(\varphi(U)\) with \(a_{ii}(q) > 0\), such that
\[
b_i(q) = \sum_{j=1}^{n} a_{ij}(q) X_j(q), \quad 1 \leq i \leq n,
\]
and \((b_1(q), \ldots, b_n(q))\) is a positively oriented orthonormal basis of \(T_{\varphi^{-1}(q)}\). We have
\[
\rho(\varphi^{-1}(q)) = \omega_{\varphi^{-1}(q)}(b_1(q), \ldots, b_n(q)) = \det(A(q)) \omega_{\varphi^{-1}(q)}(X_1(q), \ldots, X_n(q)) = \det(A(q))(\varphi^{-1})^* \omega_q(e_1, \ldots, e_n),
\]
which shows that \(\rho\) is smooth.

If we repeat the end of the proof with \(\omega = \Vol_M\), then \(\rho \equiv 1\) on \(M\), and the above formula yield
\[
((\varphi^{-1})^* \Vol_M)_q = (\det(A(q)))^{-1} dx_1 \wedge \cdots \wedge dx_n.
\]
If we compute \(\langle b_i(q), b_k(q) \rangle_{\varphi^{-1}(q)}\), we get
\[
\delta_{ik} = \langle b_i(q), b_k(q) \rangle_{\varphi^{-1}(q)} = \sum_{j=1}^{n} \sum_{l=1}^{n} a_{ij}(q) g_{jl}(q) a_{kl}(q),
\]
and so \(I = A(q)G(q)A(q)^\top\), where \(G(q) = (g_{jl}(q))\). Thus, \((\det(A(q)))^2 \det(G(q)) = 1\), and since \(\det(A(q)) = \prod_i a_{ii}(q) > 0\), we conclude that
\[
(\det(A(q)))^{-1} = \sqrt{\det(g_{ij}(q))},
\]
which proves the second formula.

We saw in Section 24.1 that a volume form \(\omega_{S^n}\) on the sphere \(S^n \subseteq \mathbb{R}^{n+1}\) is given by
\[
(\omega_{S^n})_p(u_1, \ldots, u_n) = \det(p, u_1, \ldots, u_n),
\]
where \(p \in S^n\) and \(u_1, \ldots, u_n \in T_pS^n\). To be more precise, we consider the \(n\)-form \(\omega_{\mathbb{R}^n} \in A^n(\mathbb{R}^{n+1})\) given by the above formula. As
\[
(\omega_{\mathbb{R}^n})_p(e_1, \ldots, e_i, \ldots, e_{n+1}) = (-1)^{i-1} p_i,
\]
where \( p = (p_1, \ldots, p_{n+1}) \), we have
\[
(\omega_{\mathbb{R}^n})_p = \sum_{i=1}^{n+1} (-1)^{i-1} p_i \, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{n+1}.
\]

Let \( i: S^n \to \mathbb{R}^{n+1} \) be the inclusion map. For every \( p \in S^n \) and every basis \((u_1, \ldots, u_n)\) of \( T_p S^n \), the \((n + 1)\)-tuple \((p, u_1, \ldots, u_n)\) is a basis of \( \mathbb{R}^{n+1} \), and so \((\omega_{\mathbb{R}^n})_p \neq 0\). Hence, \( \omega_{\mathbb{R}^n} \restriction S^n = i^* \omega_{\mathbb{R}^n} \) is a volume form on \( S^n \). If we give \( S^n \) the Riemannian structure induced by \( \mathbb{R}^{n+1} \), then the discussion above shows that
\[
\text{Vol}_{S^n} = \omega_{\mathbb{R}^n} \restriction S^n.
\]

To obtain another representation for \( \text{Vol}_{S^n} \), let \( r: \mathbb{R}^{n+1} - \{0\} \to S^n \) be the map given by
\[
r(x) = \frac{x}{\|x\|},
\]
and set
\[
\omega = r^* \text{Vol}_{S^n},
\]
a closed \( n \)-form on \( \mathbb{R}^{n+1} - \{0\} \). Clearly,
\[
\omega \restriction S^n = \text{Vol}_{S^n}.
\]

Furthermore
\[
\omega_x (u_1, \ldots, u_n) = (\omega_{\mathbb{R}^n})_{r(x)} (dr_x (u_1), \ldots, dr_x (u_n))
= \|x\|^{-1} \det (x, dr_x (u_1), \ldots, dr_x (u_n)).
\]
We leave it as an exercise to prove that \( \omega \) is given by
\[
\omega_x = \frac{1}{\|x\|^n} \sum_{i=1}^{n+1} (-1)^{i-1} x_i \, dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_{n+1}.
\]

The procedure used to construct \( \text{Vol}_{S^n} \) can be generalized to any \( n \)-dimensional orientable manifold embedded in \( \mathbb{R}^m \). Let \( U \) be an open subset of \( \mathbb{R}^n \) and \( \psi: U \to M \subseteq \mathbb{R}^m \) be an orientation-preserving parametrization. Assume that \( x_1, x_2, \ldots, x_m \) are the coordinates of \( \mathbb{R}^m \) (the ambient coordinates of \( M \)) and that \( u_1, u_2, \ldots, u_n \) are the coordinates of \( U \) (the local coordinates of \( M \)). Let \( x = \psi(u) \) be a point in \( M \). Edwards [66] (Theorem 5.6) shows that
\[
\text{Vol}_M = \sum_{\substack{(i_1, i_2, \ldots, i_n) \\
1 \leq i_1 < i_2 < \cdots < i_n \leq m}} n_{i_1,i_2,\ldots,i_n} \, dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_n},
\]
where
\[ n_{i_1,i_2,\ldots,i_n}(x) = \frac{1}{D} \frac{\partial (\psi_{i_1}, \psi_{i_2}, \ldots, \psi_{i_n})}{\partial (u_1, u_2, \ldots, u_n)}, \quad D = |\det (J^T(\psi)(u)J(\psi)(u))|^\frac{1}{2} \]
and \( \frac{\partial (\psi_{i_1}, \psi_{i_2}, \ldots, \psi_{i_n})}{\partial (u_1, u_2, \ldots, u_n)} \) is the determinant of the \( n \times n \) matrix obtained by selecting rows \( i_1 \) through \( i_n \) of \( d\psi_u \).

If \( M \) is a smooth orientable manifold of dimension \( m - 1 \), Edwards's formula for \( \text{Vol}_M \) reduces to
\[ \text{Vol}_M = \sum_{i=1}^{m} (-1)^{i-1} n_i \, dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_m, \quad (****) \]
where \( n_i = n_i(x) \) is the \( i \)th component of the unit normal vector \( N(x) \) on \( M \) given by
\[ n_i(x) = \frac{(-1)^{i-1} \partial (\psi_1, \ldots, \hat{\psi_i}, \ldots, \psi_m)}{D} \frac{\partial (u_1, u_2, \ldots, u_{m-1})}{\partial (u_1, u_2, \ldots, u_{m-1})}. \]
In particular, if \( M = S^n \) embedded in \( \mathbb{R}^{n+1} \), for \( p \in S^n \), \( N(p) = (p_1, p_2, \ldots, p_{n+1}) \) and \( (****) \) becomes \((*)\).

For a particular example of \((*)\), let \( M = S^2 \) and \( \psi : U \to S^2 \) where
\[ x = \sin \theta \cos \varphi, \quad y = \sin \theta \sin \varphi, \quad z = \cos \theta. \]
and \( U = \{ (\theta, \varphi) : 0 < \theta < \pi, \ 0 < \varphi < 2\pi \} \subset \mathbb{R}^2 \). See Figure 23.1. Clearly
\[ J(\psi)(\theta, \varphi) = \begin{pmatrix} \cos \theta \cos \varphi & -\sin \theta \sin \varphi \\ \cos \theta \sin \varphi & \sin \theta \cos \varphi \\ -\sin \theta & 0 \end{pmatrix}, \]
which in turn implies
\[ D = |\det (J^T(\psi)(\theta, \varphi)J(\psi)(\theta, \varphi))|^\frac{1}{2} = \left[ \det \begin{pmatrix} 1 & 0 \\ 0 & \sin^2 \theta \end{pmatrix} \right]^\frac{1}{2} = \sin \theta. \]
Then
\[ \text{Vol}_{S^2} = n_{1,2} \, dx \wedge dy + n_{1,3} \, dx \wedge dz + n_{2,3} \, dy \wedge dz, \]
where
\[ \begin{align*}
n_{1,2} &= \frac{1}{\sin \theta} \frac{\partial (x, y)}{\partial (\theta, \varphi)} = \frac{1}{\sin \theta} \det \begin{pmatrix} \cos \theta \cos \varphi & -\sin \theta \sin \varphi \\ \cos \theta \sin \varphi & \sin \theta \cos \varphi \\ -\sin \theta & 0 \end{pmatrix} = \frac{\cos \theta \sin \theta}{\sin \theta} \\
&= \cos \theta = z \\
n_{1,3} &= \frac{1}{\sin \theta} \frac{\partial (x, z)}{\partial (\theta, \varphi)} = \frac{1}{\sin \theta} \det \begin{pmatrix} \cos \theta \cos \varphi & -\sin \theta \sin \varphi \\ \cos \theta \sin \varphi & \sin \theta \cos \varphi \\ -\sin \theta & 0 \end{pmatrix} = -\frac{\sin^2 \theta \sin \varphi}{\sin \theta} \\
&= -\sin \theta \sin \varphi = -y \\
n_{2,3} &= \frac{1}{\sin \theta} \frac{\partial (y, z)}{\partial (\theta, \varphi)} = \frac{1}{\sin \theta} \det \begin{pmatrix} \cos \theta \sin \varphi & \sin \theta \cos \varphi \\ \cos \theta \sin \varphi & \sin \theta \cos \varphi \\ -\sin \theta & 0 \end{pmatrix} = \frac{\sin^2 \theta \cos \varphi}{\sin \theta} \\
&= \sin \theta \cos \varphi = x. \end{align*} \]

24.2. VOLUME FORMS ON RIEMANNIAN MANIFOLDS AND LIE GROUPS 733
Thus
\[ \text{Vol}_{S^2} = n_{1,2} \, dx \wedge dy + n_{1,3} \, dx \wedge dz + n_{2,3} \, dy \wedge dz = z \, dx \wedge dy - y \, dx \wedge dz + x \, dy \wedge dz, \]
which agrees with \((*)\) when \(n = 2\).

We mention that the orientation of \(S^n\) provides a way of oriented projective spaces of even dimension. We know that there is a map \(\pi: S^n \to \mathbb{RP}^n\) such that \(\pi^{-1}([p])\) consists of two antipodal points for every \([p] \in \mathbb{RP}^n\). It can be shown that there is a volume form on \(\mathbb{RP}^n\) iff \(n\) is even, in which case
\[ \pi^*(\text{Vol}_{\mathbb{RP}^n}) = \text{Vol}_{S^n}. \]
Thus, \(\mathbb{RP}^n\) is orientable iff \(n\) is even.

We end this section with an important result regarding orientability of Lie groups. Let \(G\) be a Lie group of dimension \(n\). For any basis \((\omega_1, \ldots, \omega_n)\) of the dual \(g^*\) of the Lie algebra \(g\) of \(G\), we have the left-invariant one-forms defined by the \(\omega_i\), also denoted \(\omega_i\), and obviously \((\omega_1, \ldots, \omega_n)\) is a frame for \(TG\). Therefore, \(\omega = \omega_1 \wedge \cdots \wedge \omega_n\) is an \(n\)-form on \(G\) that is never zero; that is, a volume form. Since pull-back commutes with \(\wedge\), the \(n\)-form \(\omega\) is left-invariant. We summarize this as

**Proposition 24.5.** Every Lie group \(G\) possesses a left-invariant volume form. Therefore, every Lie group is orientable.

### 24.3 Integration in \(\mathbb{R}^n\)

As we said in Section 23.1, one of the *raison d’être* for differential forms is that they are the objects that can be integrated on manifolds. We will be integrating differential forms that are at least continuous (in most cases, smooth) and with compact support. In the case of forms \(\omega\) on \(\mathbb{R}^n\), this means that the closure of the set \(\{x \in \mathbb{R}^n \mid \omega_x \neq 0\}\) is compact. Similarly, for a form \(\omega \in \mathcal{A}^*(M)\) where \(M\) is a manifold, the support \(\text{supp}_M(\omega)\) of \(\omega\) is the closure of the set \(\{p \in M \mid \omega_p \neq 0\}\). We let \(\mathcal{A}^*_c(M)\) denote the set of differential forms with compact support on \(M\). If \(M\) is a smooth manifold of dimension \(n\), our ultimate goal is to define a linear operator
\[ \int_M: \mathcal{A}^*_c(M) \longrightarrow \mathbb{R} \]
which generalizes in a natural way the usual integral on \(\mathbb{R}^n\).

In this section we assume that \(M = \mathbb{R}^n\) or \(M = U\) for some open subset \(U\) of \(\mathbb{R}^n\). Now every \(n\)-form (with compact support) on \(\mathbb{R}^n\) is given by
\[ \omega_x = f(x) \, dx_1 \wedge \cdots \wedge dx_n, \]
where \( f \) is a smooth function with compact support. Thus, we set

\[
\int_{\mathbb{R}^n} \omega = \int_{\mathbb{R}^n} f(x) dx_1 \cdots dx_n,
\]

where the expression on the right-hand side is the usual Riemann integral of \( f \) on \( \mathbb{R}^n \). For the reader who would like to review the definition of the Riemann integral, we suggest Sections 23.1 to 23.3 of [170] and Sections 4.1 to 4.3 of [66]. Actually we will need to integrate smooth forms \( \omega \in A^n_c(U) \) with compact support defined on some open subset \( U \subseteq \mathbb{R}^n \) (with \( \text{supp}(\omega) \subseteq U \)). However, this is no problem since we still have

\[
\omega_x = f(x) dx_1 \wedge \cdots \wedge dx_n,
\]

where \( f: U \to \mathbb{R} \) is a smooth function with compact support contained in \( U \), and \( f \) can be smoothly extended to \( \mathbb{R}^n \) by setting it to 0 on \( \mathbb{R}^n - \text{supp}(f) \). We write \( \int_U \omega \) for this integral and make the following definition.

**Definition 24.6.** Let \( U \) be an open subset of \( \mathbb{R}^n \) and let \( A^n(U) \) denote the set of smooth \( n \)-forms with compact support contained in \( U \). In other words \( \omega \in A^n(U) \) if and only if \( \omega_x = f(x) dx_1 \wedge \cdots \wedge dx_n \) for some smooth function \( f: U \to \mathbb{R} \) and the closure of \( \{ x \in \mathbb{R}^n \mid \omega_x \neq 0 \} \) is a compact set of \( \mathbb{R}^n \) contained in \( U \). For \( \omega \in A^n(U) \), the expression \( \int_U \omega \) is defined as

\[
\int_U \omega = \int_U f(x) dx_1 \cdots dx_n, \tag{*}
\]

where the right side of (*) is interpreted as the Riemann integral.

In Definition 24.6, the \( n \)-form must be represented as \( dx_1 \wedge \cdots \wedge dx_n \). This is not a problem since Proposition 23.1 that we may switch order within the wedge product by adjusting the functional coefficient with the appropriate negative signs. For example, if \( \omega_x = f(x) dx_1 \wedge dx_3 \wedge dx_2 \), Definition 24.6 implies that

\[
\int_U \omega = \int_U -f(x) dx_1 \wedge dx_3 \wedge dx_2 = -\int_U f(x) dx_1 dx_2 dx_3.
\]

For this reason, \( \int_U \omega \) is often called a “signed” integral.

It is crucial for the generalization of the integral to manifolds to see what the change of variable formula looks like in terms of differential forms.

**Proposition 24.6.** Let \( \varphi: U \to V \) be a diffeomorphism between two open subsets of \( \mathbb{R}^n \). If the Jacobian determinant \( J(\varphi)(x) \) has a constant sign \( \delta = \pm 1 \) on \( U \), then for every \( \omega \in A^n_c(V) \), we have

\[
\int_U \varphi^* \omega = \delta \int_V \omega.
\]
Proof. We know that $\omega$ can be written as
\[
\omega_x = f(x) \, dx_1 \wedge \cdots \wedge dx_n, \quad x \in V,
\]
where $f: V \to \mathbb{R}$ has compact support. From the example after Proposition 23.6, we have
\[
(\varphi^*\omega)_y = f(\varphi(y)) J(\varphi)_y \, dy_1 \wedge \cdots \wedge dy_n = \delta f(\varphi(y)) |J(\varphi)_y| \, dy_1 \wedge \cdots \wedge dy_n.
\]
On the other hand, the change of variable formula (using $\varphi$) is
\[
\int_{\varphi(U)} f(x) \, dx_1 \cdots dx_n = \int_U f(\varphi(y)) |J(\varphi)_y| \, dy_1 \cdots dy_n,
\]
so the formula follows. \qed

We will promote the integral on open subsets of $\mathbb{R}^n$ to manifolds using partitions of unity.

### 24.4 Integration on Manifolds

**Definition 24.7.** Let $M$ be an oriented manifold of dimension $n$. We say $\omega$ is a smooth $n$-form on $M$ with compact support if the closure of $\{ p \in M \mid \omega_p \neq 0 \}$ is compact in $M$. We denote $\{ p \in M \mid \omega_p \neq 0 \}$ by $\text{supp}(\omega)$. The set of smooth $n$-forms on $M$ with compact support is denoted $\omega \in \mathcal{A}_c^n(M)$ while $\omega \in \mathcal{A}_c^*(M)$ is the set of all smooth differential forms on $M$ with compact support.

Intuitively, for any $n$-form $\omega \in \mathcal{A}_c^n(M)$ on a smooth $n$-dimensional oriented manifold $M$, the integral $\int_M \omega$ is computed by patching together the integrals on small-enough open subsets covering $M$ using a partition of unity. If $(U, \varphi)$ is a chart such that $\text{supp}(\omega) \subseteq U$, then the form $(\varphi^{-1})^*\omega$ is an $n$-form on $\mathbb{R}^n$, and the integral $\int_{\varphi(U)} (\varphi^{-1})^*\omega$ makes sense. The orientability of $M$ is needed to ensure that the above integrals have a consistent value on overlapping charts.

**Proposition 24.7.** Let $M$ be a smooth oriented manifold of dimension $n$. There exists a unique linear operator
\[
\int_M : \mathcal{A}_c^n(M) \rightarrow \mathbb{R}
\]
with the following property: For any $\omega \in \mathcal{A}_c^n(M)$, if $\text{supp}(\omega) \subseteq U$, where $(U, \varphi)$ is a positively oriented chart, then
\[
\int_M \omega = \int_{\varphi(U)} (\varphi^{-1})^*\omega. \quad (\dagger)
\]
Proof. First, assume that $\text{supp}(\omega) \subseteq U$, where $(U, \varphi)$ is a positively oriented chart. Then, we wish to set
\[ \int_M \omega = \int_{\varphi(U)} (\varphi^{-1})^* \omega. \]
However, we need to prove that the above expression does not depend on the choice of the chart. Let $(V, \psi)$ be another chart such that $\text{supp}(\omega) \subseteq V$, so that $\text{supp}(\omega) \subseteq U \cap V$. The map $\theta = \psi \circ \varphi^{-1}$ is a diffeomorphism from $W = \varphi(U \cap V)$ to $W' = \psi(U \cap V)$, and by hypothesis, its Jacobian determinant is positive on $W$. Since $\text{supp}(\varphi(U))((\varphi^{-1})^* \omega) \subseteq W$, $\text{supp}(\psi(V))((\psi^{-1})^* \omega) \subseteq W'$, and $\theta^* (\psi^{-1})^* \omega = (\varphi^{-1})^* \circ \psi^* \circ (\psi^{-1})^* \omega = (\varphi^{-1})^* \omega$, Proposition 24.6 yields
\[ \int_{W'} (\psi^{-1})^* \omega = \int_W \theta^* (\psi^{-1})^* \omega = \int_W (\varphi^{-1})^* \omega, \]
as claimed.

In the general case, using Theorem 9.4, for every open cover of $M$ by positively oriented charts $(U_i, \varphi_i)$, we have a partition of unity $(\rho_i)_{i \in I}$ subordinate to this cover. Recall that $\text{supp}(\rho_i) \subseteq U_i$, $i \in I$.

Thus, $\rho_i \omega$ is an $n$-form whose support is a subset of $U_i$. Furthermore, as $\sum_i \rho_i = 1$,
\[ \omega = \sum_i \rho_i \omega. \]
Define
\[ I(\omega) = \sum_i \int_{U_i} \rho_i \omega, \]
where each term in the sum is defined by
\[ \int_{U_i} \rho_i \omega = \int_{\varphi_i(U_i)} (\varphi_i^{-1})^* \rho_i \omega, \]
where $(U_i, \varphi_i)$ is the chart associated with $i \in I$.

It remains to show that $I(\omega)$ does not depend on the choice of open cover and on the choice of partition of unity. Let $(V_j, \psi_j)$ be another open cover by positively oriented charts, and let $(\theta_j)_{j \in J}$ be a partition of unity subordinate to the open cover $(V_j)$. Note that
\[ \int_{U_i} \rho_i \theta_j \omega = \int_{V_j} \rho_i \theta_j \omega, \]
since supp(\(\rho_i \theta_j \omega\)) \(\subseteq U_i \cap V_j\), and as \(\sum_i \rho_i = 1\) and \(\sum_j \theta_j = 1\), we have
\[
\sum_i \int_{U_i} \rho_i \omega = \sum_{i,j} \int_{U_i} \rho_i \theta_j \omega = \sum_{i,j} \int_{V_j} \rho_i \theta_j \omega = \sum_j \int_{V_j} \theta_j \omega,
\]
proving that \(I(\omega)\) is indeed independent of the open cover and of the partition of unity. The uniqueness assertion is easily proved using a partition of unity. \(\square\)

Since the integral at \((\dagger)\) is well-defined we are able to make the following definition.

**Definition 24.8.** Let \(M\) be a smooth oriented manifold of dimension \(n\). For \(\omega \in A_c^n(M)\), if \((U, \varphi)\) is a positively oriented chart, and \(\text{supp}(\omega) \subseteq U\), we define \(\int_M \omega\) by
\[
\int_M \omega = \int_{\varphi(U)} (\varphi^{-1})^* \omega.
\]

Given an embedded manifold \(M\) in \(\mathbb{R}^n\), Definition 24.8 shows that integration of a form over a manifold reduces, after a change of variables, to an appropriate Riemann integral over the parameter space. We will demonstrate the meaning of this sentence by explicitly calculating \(\int_{S^2} \text{Vol}_{S^2}\). In Section 24.2 we described a parametrization of \(S^2\) by \(\psi: U \to S^2\) where
\[
x = \sin \theta \cos \varphi, \quad y = \sin \theta \sin \varphi, \quad z = \cos \theta,
\]
and \(U = \{ (\theta, \varphi) : 0 < \theta < \pi, \ 0 < \varphi < 2\pi \} \subseteq \mathbb{R}^2\). See Figure 23.1. We then showed that
\[
\text{Vol}_{S^2} = z \, dx \wedge dy - y \, dx \wedge dz + x \, dy \wedge dz.
\]
To calculate \(\int_{S^2} \text{Vol}_{S^2}\), we first use \((\ast)\) of Section 23.2 to calculate
\[
\psi^*(z \, dx \wedge dy) = \cos \theta (d(\sin \theta \cos \varphi) \wedge d(\sin \theta \sin \varphi))
\]
\[
= \cos \theta ((\cos \theta \cos \varphi \, d\theta - \sin \theta \sin \varphi \, d\varphi) \wedge (\cos \theta \sin \varphi \, d\theta + \sin \theta \cos \varphi \, d\varphi))
\]
\[
= \cos \theta (\cos^2 \varphi \cos \theta \sin \theta + \sin^2 \varphi \sin \theta \cos \theta) \, d\theta \wedge d\varphi
\]
\[
= \cos^2 \theta \sin \theta \, d\theta \wedge d\varphi
\]
\[
\psi^*(-y \, dx \wedge dz) = -\sin \theta \sin \varphi (d(\sin \theta \cos \varphi) \wedge d(\cos \theta))
\]
\[
= -\sin \theta \sin \varphi ((\cos \theta \cos \varphi \, d\theta - \sin \theta \sin \varphi \, d\varphi) \wedge -\sin \theta \, d\theta)
\]
\[
= \sin^3 \theta \sin^2 \varphi \, d\theta \wedge d\varphi
\]
\[
\psi^*(x \, dy \wedge dz) = \sin \theta \cos \varphi (d(\sin \theta \sin \varphi) \wedge d(\cos \theta))
\]
\[
= \sin \theta \cos \varphi ((\cos \theta \sin \varphi \, d\theta + \sin \theta \cos \varphi \, d\varphi) \wedge -\sin \theta \, d\theta)
\]
\[
= \sin^3 \theta \cos^2 \varphi \, d\theta \wedge d\varphi
\].
24.4. INTEGRATION ON MANIFOLDS

Then
\[
\varphi^*(\text{Vol}_{S^2}) = (\cos^2 \theta \sin \theta + \sin^3 \theta \sin^2 \varphi + \sin^3 \theta \cos^2 \varphi) d\theta \wedge d\varphi \\
= (\cos^2 \theta \sin \theta + \sin^3 \theta) d\theta \wedge d\varphi = \sin \theta (\cos^2 \theta + \sin^2 \theta) d\theta \wedge d\varphi \\
= \sin \theta d\theta \wedge d\varphi,
\]
and Line (†) implies that
\[
\int_{S^2} (\text{Vol}_{S^2}) = \int_0^{2\pi} \int_0^\pi \varphi^*(\text{Vol}_{S^2}) = \int_0^{2\pi} \int_0^\pi \sin \varphi, d\theta d\varphi = 2\pi [-\cos \varphi]_0^\pi = 4\pi.
\]
Observe that 4\pi is indeed the surface area of S^2, a result we should have expected since we were integrating the volume form.

The integral of Definition 24.8 has the following properties:

**Proposition 24.8.** Let M be an oriented manifold of dimension n. The following properties hold:

1. If M is connected, then for every n-form \( \omega \in \mathcal{A}^n_c(M) \), the sign of \( \int_M \omega \) changes when the orientation of M is reversed.

2. For every n-form \( \omega \in \mathcal{A}^n_c(M) \), if \( \text{supp}(\omega) \subseteq W \) for some open subset W of M, then
\[
\int_M \omega = \int_W \omega,
\]
where W is given the orientation induced by M.

3. If \( \varphi: M \to N \) is an orientation-preserving diffeomorphism, then for every \( \omega \in \mathcal{A}^n_c(N) \), we have
\[
\int_N \omega = \int_M \varphi^* \omega.
\]

**Proof.** Use a partition of unity to reduce to the case where \( \text{supp}(\omega) \) is contained in the domain of a chart, and then use Proposition 24.6 and (†) from Proposition 24.7.

It is also possible to define integration on non-orientable manifolds using densities.

**Definition 24.9.** Given a vector space V of dimension \( n \geq 1 \), a density on V is a function \( \mu: V^n \to \mathbb{R} \) such that for every linear map \( f: V \to V \), we have
\[
\mu(f(v_1), \ldots, f(v_n)) = |\det(f)| \mu(v_1, \ldots, v_n)
\]
for all \( v_1, \ldots, v_n \in V \).
If \((v_1, \ldots, v_n)\) are linearly dependent, then for any basis \((e_1, \ldots, e_n)\) of \(V\) there is a unique linear map \(f\) such that \(f(e_i) = v_i\) for \(i = 1, \ldots, n\), and since \((v_1, \ldots, v_n)\) are linearly dependent, \(f\) is singular so \(\det(f) = 0\), which implies that
\[
\mu(v_1, \ldots, v_n) = |\det(f)|\mu(e_1, \ldots, e_n) = 0
\]
for any linearly dependent vectors \(v_1, \ldots, v_n \in V\).

In view of this fact, a density is sometimes defined as a function \(\mu \colon \bigwedge^n V \to \mathbb{R}\) such that for every automorphism \(f \in \text{GL}(V)\),
\[
\mu(f(v_1) \wedge \cdots \wedge f(v_n)) = |\det(f)|\mu(v_1 \wedge \cdots \wedge v_n)
\]
(\(\dagger\)) for all \(v_1 \wedge \cdots \wedge v_n \in V\) (with \(\mu(0) = 0\)). For any nonzero \(v_1 \wedge \cdots \wedge v_n, w_1 \wedge \cdots \wedge w_n \in \bigwedge^n V\), because
\[
w_1 \wedge \cdots \wedge w_n = \det(P)v_1 \wedge \cdots \wedge v_n
\]
where \(P\) is the matrix whose \(j\)th column consists of the coefficients of \(w_j\) over the basis \((v_1, \ldots, v_n)\), it is not hard to show that condition (\(\dagger\)) is equivalent to
\[
\mu(cw) = |c|\mu(w), \quad w \in \bigwedge^n V, \ c \in \mathbb{R}.
\]

Densities are not multilinear, but it is not hard to show that for any fixed \(n\), they form a vector space of dimension 1 which is spanned by the absolute value \(|\omega|\) of any nonzero \(n\)-form \(\omega \in \bigwedge^n V^*\). Let \(\text{den}(V)\) be the set of all densities on \(V\). We have the following proposition from Lee [117] (Chapter 14, Proposition 14.26).

**Proposition 24.9.** Let \(V\) be any vector space of dimension \(n \geq 1\). The following properties hold:

(a) The set \(\text{den}(V)\) is a vector space.

(b) For any two densities \(\mu_1, \mu_2 \in \text{den}(V)\) and for any basis \((e_1, \ldots, e_n)\) of \(V\), if \(\mu_1(e_1, \ldots, e_n) = \mu_2(e_1, \ldots, e_n)\), then \(\mu_1 = \mu_2\).

(c) For any \(n\)-form \(\omega \in \bigwedge^n V^*\), the function \(|\omega|\) given by
\[
|\omega|(v_1, \ldots, v_n) = |\omega(v_1, \ldots, v_n)|
\]
is a density.

(d) The vector space \(\text{den}(V)\) is a one-dimensional space spanned by \(|\omega|\) for any nonzero \(\omega \in \bigwedge^n V^*\).
Proof. (a) That $\text{den}(V)$ is a vector space is immediate from the definition.

(b) Pick any $n$ vectors $(v_1, \ldots, v_n) \in V^n$. Since $(e_1, \ldots, e_n)$ is a basis of $V$, there is a unique linear map $f: V \rightarrow V$ such that $f(e_i) = v_i$ for $i = 1, \ldots, n$, and since by hypothesis $\mu_1(e_1, \ldots, e_n) = \mu_2(e_1, \ldots, e_n)$, we have

$$\begin{align*}
\mu_1(v_1, \ldots, v_n) &= \mu_1(f(e_1), \ldots, f(e_n)) \\
&= |\det(f)| \mu_1(e_1, \ldots, e_n) \\
&= |\det(f)| \mu_2(e_1, \ldots, e_n) \\
&= \mu_2(f(e_1), \ldots, f(e_n)) \\
&= \mu_2(v_1, \ldots, v_n),
\end{align*}$$

which proves that $\mu_1 = \mu_2$.

(c) If $\omega \in \bigwedge^n V^*$, then

$$\begin{align*}
|\omega|(f(v_1), \ldots, f(v_n)) &= |\omega(f(v_1), \ldots, f(v_n))| \\
&= |\det(f)\omega(v_1, \ldots, v_n)| \\
&= |\det(f)| |\omega|(v_1, \ldots, v_n),
\end{align*}$$

which shows that $|\omega|$ is a density.

(d) Let $(e_1, \ldots, e_n)$ be any basis of $V$, and let $\omega \in \bigwedge^n V^*$ be any nonzero $n$-form. For any density $\mu$, we need to show that $\mu = c|\omega|$ for some $c \in \mathbb{R}$. Let

$$\begin{align*}
a &= |\omega|(e_1, \ldots, e_n) = |\omega(e_1, \ldots, e_n)| \\
b &= \mu(e_1, \ldots, e_n).
\end{align*}$$

Since $\omega \neq 0$ and $(e_1, \ldots, e_n)$ is a basis, $\omega(e_1, \ldots, e_n) \neq 0$ so $a \neq 0$, and by (c) $(b/a)|\omega|$ is a density. Since

$$\begin{align*}
(b/a)|\omega|(e_1, \ldots, e_n) &= b = \mu(e_1, \ldots, e_n),
\end{align*}$$

by (b) $\mu = (b/a)|\omega|$, as desired. 

If we denote the vector space of densities on $V$ by $\text{den}(V)$, then given a manifold $M$, we can form the density bundle $\text{den}(M)$ whose underlying set is the disjoint union of the vector spaces $\text{den}(T_pM)$ for all $p \in M$. This set can be made into a smooth bundle, and a density on $M$ is a smooth section of the density bundle. The main property of densities is that every smooth manifold admits a global smooth (positive) density, without any orientability assumptions. Then, it is possible to carry out the theory of integration on manifolds using densities instead of volume forms, as we did in this section. This development can be found in Lee [117] (Chapter 14), but we have no need for this extra generality.
It turns out that orientations can be defined as certain functions satisfying a variant of the condition used in Definition 24.9, and this definition clarifies the relationship between volume forms and densities. The sign function is defined such that for any \( \lambda \in \mathbb{R} \),

\[
\text{sign}(\lambda) = \begin{cases} 
+1 & \text{if } \lambda > 0 \\
-1 & \text{if } \lambda < 0 \\
0 & \text{if } \lambda = 0 
\end{cases}
\]

**Definition 24.10.** Given a vector space \( V \) of dimension \( n \geq 1 \), an orientation on \( V \) is a function \( o: V^n \to \mathbb{R} \) such that for every linear map \( f: V \to V \), we have

\[
o(f(v_1), \ldots, f(v_n)) = \text{sign}(\text{det}(f))o(v_1, \ldots, v_n)
\]

for all \( v_1, \ldots, v_n \in V \).

If \((v_1, \ldots, v_n)\) are linearly dependent, then for any basis \((e_1, \ldots, e_n)\) of \( V \) there is a unique linear map \( f \) such that \( f(e_i) = v_i \) for \( i = 1, \ldots, n \), and since \((v_1, \ldots, v_n)\) are linearly dependent, \( f \) is singular so \( \text{det}(f) = 0 \), which implies that

\[
o(v_1, \ldots, v_n) = \text{sign}(\text{det}(f))o(e_1, \ldots, e_n) = 0
\]

for any linearly dependent vectors \( v_1, \ldots, v_n \in V \).

For any two bases \((u_1, \ldots, u_n)\) and \((v_1, \ldots, v_n)\), there is a unique linear map \( f \) such that \( f(u_i) = v_i \) for \( i = 1, \ldots, n \), and \( o(u_1, \ldots, u_n) = o(v_1, \ldots, v_n) \) if \( \text{det}(f) > 0 \), which is indeed the condition for \((u_1, \ldots, u_n)\) and \((v_1, \ldots, v_n)\) to have the same orientation. There are exactly two orientations \( o \) such that \( |o(u_1, \ldots, u_n)| = 1 \).

Let \( \text{Or}(V) \) be the set of all orientations on \( V \). We have the following proposition.

**Proposition 24.10.** Let \( V \) be any vector space of dimension \( n \geq 1 \). The following properties hold:

(a) The set \( \text{Or}(V) \) is a vector space.

(b) For any two orientations \( o_1, o_2 \in \text{Or}(V) \) and for any basis \((e_1, \ldots, e_n)\) of \( V \), if \( o_1(e_1, \ldots, e_n) = o_2(e_1, \ldots, e_n) \), then \( o_1 = o_2 \).

(c) For any nonzero \( n \)-form \( \omega \in \wedge^n V^* \), the function \( o(\omega) \) given by \( o(\omega)(v_1, \ldots, v_n) = 0 \) if \((v_1, \ldots, v_n)\) are linearly dependent and

\[
o(\omega)(v_1, \ldots, v_n) = \frac{\omega(v_1, \ldots, v_n)}{|\omega(v_1, \ldots, v_n)|}
\]

if \((v_1, \ldots, v_n)\) are linearly independent is an orientation.

(d) The vector space \( \text{Or}(V) \) is a one-dimensional space, and it is spanned by \( o(\omega) \) for any nonzero \( \omega \in \wedge^n V^* \).
Proof. (a) That $\text{Or}(V)$ is a vector space is immediate from the definition.

(b) Pick any $n$ vectors $(v_1, \ldots, v_n) \in V^n$. Since $(e_1, \ldots, e_n)$ is a basis of $V$, there is a unique linear map $f: V \to V$ such that $f(e_i) = v_i$ for $i = 1, \ldots, n$, and since by hypothesis $o_1(e_1, \ldots, e_n) = o_2(e_1, \ldots, e_n)$, we have

$$
\begin{align*}
o_1(v_1, \ldots, v_n) &= o_1(f(e_1), \ldots, f(e_n)) \\
&= \text{sign}(|\text{det}(f)|) o_1(e_1, \ldots, e_n) \\
&= \text{sign}(|\text{det}(f)|) o_2(e_1, \ldots, e_n) \\
&= o_2(f(e_1), \ldots, f(e_n)) \\
&= o_2(v_1, \ldots, v_n),
\end{align*}
$$

which proves that $o_1 = o_2$.

(c) Let $\omega \in \bigwedge^n V^*$ be any nonzero form. If $(v_1, \ldots, v_n)$ are linearly independent, then we know that

$$
\begin{align*}
\omega(f(v_1), \ldots, f(v_n)) &= \text{det}(f) \omega(v_1, \ldots, v_n) \\
|\omega|(f(v_1), \ldots, f(v_n)) &= |\text{det}(f)||\omega|(v_1, \ldots, v_n).
\end{align*}
$$

We know that $\text{det}(f) = 0$ iff $f$ is singular, but then $(f(v_1), \ldots, f(v_n))$ are linearly dependent so

$$
o_\omega(f(v_1), \ldots, f(v_n)) = 0 = \text{sign}(|\text{det}(f)|) o_\omega(v_1, \ldots, v_n).
$$

If $\text{det}(f) \neq 0$, then

$$
\begin{align*}
o_\omega(f(v_1), \ldots, f(v_n)) &= \frac{\omega(f(v_1), \ldots, f(v_n))}{|\omega|(f(v_1), \ldots, f(v_n))} \\
&= \frac{\text{det}(f)}{|\text{det}(f)|} \frac{\omega(f(v_1), \ldots, f(v_n))}{|\omega|(f(v_1), \ldots, f(v_n))} \\
&= \text{sign}(|\text{det}(f)|) o_\omega(v_1, \ldots, v_n),
\end{align*}
$$

which shows that $o_\omega$ is an orientation.

(d) Let $(e_1, \ldots, e_n)$ be any basis of $V$, and let $\omega \in \bigwedge^n V^*$ be any nonzero $n$-form. For any orientation $o$, we need to show that $o = co(\omega)$ for some $c \in \mathbb{R}$. Let

$$
\begin{align*}
a &= \omega(o)(e_1, \ldots, e_n) \\
b &= o(e_1, \ldots, e_n).
\end{align*}
$$

Since $\omega \neq 0$ and $(e_1, \ldots, e_n)$ is a basis, $\omega(e_1, \ldots, e_n) \neq 0$ so $a \neq 0$, and by (c) $(b/a)o(\omega)$ is an orientation. Since

$$
(b/a)o(\omega) = b = o(e_1, \ldots, e_n),
$$

by (b) $o = (b/a)o(\omega)$, as desired.
Part (c) of Proposition 24.10 implies that for every nonzero $n$-form $\omega \in \bigwedge^n V^*$, there exists some density $|\omega|$ and some orientation $o(\omega)$ such that
\[ o(\omega)|\omega| = \omega. \]
This shows that orientations are just normalized volume forms that take exactly two values $c$ and $-c$ on linearly independent vectors (with $c > 0$), whereas densities are absolute values of volume forms. We have the following results showing the relationship between the spaces $\bigwedge^n V^*$, $\text{Or}(V)$, and $\text{den}(V)$.

**Proposition 24.11.** Let $V$ be any vector space of dimension $n \geq 1$. For any nonzero $n$-form $\omega \in \bigwedge^n V^*$, the bilinear map $\Phi: \text{Or}(V) \times \text{den}(V) \to \bigwedge^n V^*$ given by
\[ \Phi(\alpha o(\omega), \beta |\omega|) = \alpha \beta \omega, \quad \alpha, \beta \in \mathbb{R} \]
induces an isomorphism $\text{Or}(V) \otimes \text{den}(V) \cong \bigwedge^n V^*$.

**Proof.** The spaces $\bigwedge^n V^*$, $\text{Or}(V)$, and $\text{den}(V)$ are all one-dimensional, and if $\omega \neq 0$, then $\omega$ is a basis of $\bigwedge^n V^*$ and Propositions 24.9 and 24.10 show that $o(\omega)$ is a basis of $\text{Or}(V)$ and $|\omega|$ is a basis of $\text{den}(V)$, so the map $\Phi$ defines a bilinear map from $\text{Or}(V) \times \text{den}(V)$ to $\bigwedge^n V^*$. Therefore, by the universal mapping property, we obtain a linear map $\Phi_{\otimes}: \text{Or}(V) \otimes \text{den}(V) \to \bigwedge^n V^*$. Since $\omega \neq 0$, we have
\[ o(\omega)|\omega| = \omega, \]
which shows that $\Phi$ is surjective, and thus $\Phi_{\otimes}$ is surjective. Since all the spaces involved are one-dimensional, $\text{Or}(V) \otimes \text{den}(V)$ is also one-dimensional, so $\Phi_{\otimes}$ is bijective. \[ \Box \]

Given a manifold $M$, we can form the orientation bundle $\text{Or}(M)$ whose underlying set is the disjoint union of the vector spaces $\text{Or}(T_pM)$ for all $p \in M$. This set can be made into a smooth bundle, and an orientation of $M$ is a smooth global section of the orientation bundle. Then, it can be shown that there is a bundle isomorphism
\[ \text{Or}(M) \otimes \text{den}(M) \cong \bigwedge^n T^*M. \]
and since $\text{den}(M)$ always has global sections, we see that there is a global volume form iff $\text{Or}(M)$ has a global section iff $M$ is orientable.

The theory of integration developed so far deals with domains that are not general enough. Indeed, for many applications, we need to integrate over domains with boundaries.
24.5 Manifolds With Boundary

Up to now, we have defined manifolds locally diffeomorphic to an open subset of $\mathbb{R}^m$. This excludes many natural spaces such as a closed disk, whose boundary is a circle, a closed ball $B(1)$, whose boundary is the sphere $S^{m-1}$, a compact cylinder $S^1 \times [0, 1]$, whose boundary consist of two circles, a Möbius strip, etc. These spaces fail to be manifolds because they have a boundary; that is, neighborhoods of points on their boundaries are not diffeomorphic to open sets in $\mathbb{R}^m$. Perhaps the simplest example is the (closed) upper half space

$$ \mathbb{H}^m = \{(x_1, \ldots, x_m) \in \mathbb{R}^m \mid x_m \geq 0\}. $$

Under the natural embedding $\mathbb{R}^{m-1} \cong \mathbb{R}^{m-1} \times \{0\} \hookrightarrow \mathbb{R}^m$, the subset $\partial \mathbb{H}^m$ of $\mathbb{H}^m$ defined by

$$ \partial \mathbb{H}^m = \{x \in \mathbb{H}^m \mid x_m = 0\} $$

is isomorphic to $\mathbb{R}^{m-1}$, and is called the boundary of $\mathbb{H}^m$. We also define the interior of $\mathbb{H}^m$ as

$$ \text{Int}(\mathbb{H}^m) = \mathbb{H}^m - \partial \mathbb{H}^m. $$

Now, if $U$ and $V$ are open subsets of $\mathbb{H}^m$, where $\mathbb{H}^m \subseteq \mathbb{R}^m$ has the subset topology, and if $f: U \rightarrow V$ is a continuous function, we need to explain what we mean by $f$ being smooth. We say that $f: U \rightarrow V$ as above is smooth if it has an extension $\tilde{f}: \tilde{U} \rightarrow \tilde{V}$, where $\tilde{U}$ and $\tilde{V}$ are open subsets of $\mathbb{R}^m$ with $U \subseteq \tilde{U}$ and $V \subseteq \tilde{V}$, and with $\tilde{f}$ a smooth function. We say that $f$ is a (smooth) diffeomorphism iff $f^{-1}$ exists and if both $f$ and $f^{-1}$ are smooth, as just defined.

To define a manifold with boundary, we replace everywhere $\mathbb{R}$ by $\mathbb{H}$ in Definition 7.1 and Definition 7.2. So, for instance, given a topological space $M$, a chart is now pair $(U, \varphi)$, where $U$ is an open subset of $M$ and $\varphi: U \rightarrow \Omega$ is a homeomorphism onto an open subset $\Omega = \varphi(U)$ of $\mathbb{H}^{n_\varphi}$ (for some $n_\varphi \geq 1$), etc. Thus, we obtain

**Definition 24.11.** Given some integer $n \geq 1$ and given some $k$ such that $k$ is either an integer $k \geq 1$ or $k = \infty$, a $C^k$-manifold of dimension $n$ with boundary consists of a topological space $M$ together with an equivalence class $\mathcal{A}$ of $C^k$ $n$-atlases on $M$ (where the charts are now defined in terms of open subsets of $\mathbb{H}^n$). Any atlas $\mathcal{A}$ in the equivalence class $\mathcal{A}$ is called a differentiable structure of class $C^k$ (and dimension $n$) on $M$. We say that $M$ is modeled on $\mathbb{H}^n$. When $k = \infty$, we say that $M$ is a smooth manifold with boundary.

It remains to define what is the boundary of a manifold with boundary! By definition, the boundary $\partial M$ of a manifold (with boundary) $M$ is the set of all points $p \in M$, such that there is some chart $(U_\alpha, \varphi_\alpha)$, with $p \in U_\alpha$ and $\varphi_\alpha(p) \in \partial \mathbb{H}^n$. We also let $\text{Int}(M) = M - \partial M$ and call it the interior of $M$.

Do not confuse the boundary $\partial M$ and the interior $\text{Int}(M)$ of a manifold with boundary embedded in $\mathbb{R}^N$ with the topological notions of boundary and interior of $M$ as a topological space. In general, they are different. For example, if $M$ is the subset $[0, 1) \cup \{2\}$ of the real line, then its manifold boundary is $\partial M = \{0\}$, and its topological boundary is $\text{Bd}(M) = \{0, 1, 2\}$. 

---

**Definition 7.1.** A topological space $M$ is a manifold, when $M$ is locally homeomorphic to an open subset of $\mathbb{R}^n$. This notion allows for the boundary of $M$.

**Definition 7.2.** So, for instance, given a topological space $M$ such that $f$ is a (smooth) function. We say a manifold with boundary $M$ has the subset topology, and if $f: U \rightarrow V$ is a continuous function, we need to explain what we mean by $f$ being smooth. We say that $f: U \rightarrow V$ as above is smooth if it has an extension $\tilde{f}: \tilde{U} \rightarrow \tilde{V}$, where $\tilde{U}$ and $\tilde{V}$ are open subsets of $\mathbb{R}^m$ with $U \subseteq \tilde{U}$ and $V \subseteq \tilde{V}$, and with $\tilde{f}$ a smooth function. We say that $f$ is a (smooth) diffeomorphism iff $f^{-1}$ exists and if both $f$ and $f^{-1}$ are smooth, as just defined.

---

**Example 24.5.** The closed unit ball $B(1)$ in $\mathbb{R}^m$ is a manifold with boundary.

---

**Example 24.6.** The closed disk $D(1)$ in $\mathbb{R}^2$ is a manifold with boundary.

---

**Example 24.7.** The closed ball $B(1)$ in $\mathbb{R}^3$ is a manifold with boundary.

---

**Example 24.8.** The closed upper half space $\mathbb{H}^m$ in $\mathbb{R}^m$ is a manifold with boundary.
Figure 24.4: A two dimensional manifold with red boundary

Note that manifolds as defined earlier (In Definition 7.3) are also manifolds with boundary: their boundary is just empty. We shall still reserve the word “manifold” for these, but for emphasis, we will sometimes call them “boundaryless.”

The definition of tangent spaces, tangent maps, etc., are easily extended to manifolds with boundary. The reader should note that if $M$ is a manifold with boundary of dimension $n$, the tangent space $T_p M$ is defined for all $p \in M$ and has dimension $n$, even for boundary points $p \in \partial M$. The only notion that requires more care is that of a submanifold. For more on this, see Hirsch [91], Chapter 1, Section 4. One should also beware that the product of two manifolds with boundary is generally not a manifold with boundary (consider the product $[0,1] \times [0,1]$ of two line segments). There is a generalization of the notion of a manifold with boundary called manifold with corners, and such manifolds are closed under products (see Hirsch [91], Chapter 1, Section 4, Exercise 12).

If $M$ is a manifold with boundary, we see that $\text{Int}(M)$ is a manifold without boundary. What about $\partial M$? Interestingly, the boundary $\partial M$ of a manifold with boundary $M$ of dimension $n$ is a manifold of dimension $n - 1$. For this we need the following proposition:

**Proposition 24.12.** If $M$ is a manifold with boundary of dimension $n$, for any $p \in \partial M$ on the boundary on $M$, for any chart $(U, \varphi)$ with $p \in M$, we have $\varphi(p) \in \partial \mathbb{H}^n$.

**Proof.** Since $p \in \partial M$, by definition, there is some chart $(V, \psi)$ with $p \in V$ and $\psi(p) \in \partial \mathbb{H}^n$. Let $(U, \varphi)$ be any other chart, with $p \in M$, and assume that $q = \varphi(p) \in \text{Int}(\mathbb{H}^n)$. The transition map $\psi \circ \varphi^{-1}: \varphi(U \cap V) \to \psi(U \cap V)$ is a diffeomorphism, and $q = \varphi(p) \in \text{Int}(\mathbb{H}^n)$. By the inverse function theorem, there is some open $W \subseteq \varphi(U \cap V) \cap \text{Int}(\mathbb{H}^n) \subseteq \mathbb{R}^n$, with
\( q \in W \), so that \( \psi \circ \varphi^{-1} \) maps \( W \) homeomorphically onto some subset \( \Omega \) open in \( \text{Int}(\mathbb{H}^n) \), with \( \psi(p) \in \Omega \), contradicting the hypothesis, \( \psi(p) \in \partial \mathbb{H}^n \).

Using Proposition 24.12, we immediately derive the fact that \( \partial M \) is a manifold of dimension \( n - 1 \). We obtain charts on \( \partial M \) by considering the charts \((U \cap \partial M, L \circ \varphi)\), where \((U, \varphi)\) is a chart on \( M \) such that \( U \cap \partial M = \varphi^{-1}(\partial \mathbb{H}^n) \neq \emptyset \) and \( L : \partial \mathbb{H}^n \to \mathbb{R}^{n-1} \) is the natural isomorphism (see see Hirsch [91], Chapter 1, Section 4).

### 24.6 Integration on Regular Domains and Stokes’ Theorem

Given a manifold \( M \), we define a class of subsets with boundaries that can be integrated on, and for which Stokes’ Theorem holds. In Warner [175] (Chapter 4), such subsets are called regular domains, and in Madsen and Tornehave [119] (Chapter 10), they are called domains with smooth boundary.

**Definition 24.12.** Let \( M \) be a smooth manifold of dimension \( n \). A subset \( N \subseteq M \) is called a domain with smooth boundary (or codimension zero submanifold with boundary) iff for every \( p \in M \), there is a chart \((U, \varphi)\) with \( p \in U \) such that

\[
\varphi(U \cap N) = \varphi(U) \cap \mathbb{H}^n, \tag{*}
\]

where \( \mathbb{H}^n \) is the closed upper-half space

\[
\mathbb{H}^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_n \geq 0\}.
\]

Note that \((*)\) is automatically satisfied when \( p \) is an interior or an exterior point of \( N \), since we can pick a chart such that \( \varphi(U) \) is contained in an open half space of \( \mathbb{R}^n \) defined by either \( x_n > 0 \) or \( x_n < 0 \). If \( p \) is a boundary point of \( N \), then \( \varphi(p) \) has its last coordinate equal to 0; see Figure 24.5.

If \( M \) is orientable, then any orientation of \( M \) induces an orientation of \( \partial N \), the boundary of \( N \). This follows from the following proposition:

**Proposition 24.13.** Let \( \varphi : \mathbb{H}^n \to \mathbb{H}^n \) be a diffeomorphism with everywhere positive Jacobian determinant. Then, \( \varphi \) induces a diffeomorphism \( \Phi : \partial \mathbb{H}^n \to \partial \mathbb{H}^n \), which viewed as a diffeomorphism of \( \mathbb{R}^{n-1} \), also has everywhere positive Jacobian determinant.

**Proof.** By the inverse function theorem, every interior point of \( \mathbb{H}^n \) is the image of an interior point, so \( \varphi \) maps the boundary to itself. If \( \varphi = (\varphi_1, \ldots, \varphi_n) \), then

\[
\Phi = (\varphi_1(x_1, \ldots, x_{n-1}, 0), \ldots, \varphi_{n-1}(x_1, \ldots, x_{n-1}, 0)),
\]

so \( \Phi \) is a diffeomorphism of \( \mathbb{R}^{n-1} \). The Jacobian determinant of \( \Phi \) is the product of the Jacobian determinants of the \( \varphi_i \)'s, which are all positive by assumption, so the Jacobian determinant of \( \Phi \) is also positive. Therefore, \( \Phi \) has everywhere positive Jacobian determinant.
Figure 24.5: The subset $N$, the peach region of the torus $M$, is a domain with smooth boundary.

since $\varphi_n(x_1, \ldots, x_{n-1}, 0) = 0$. It follows that $\frac{\partial \varphi_n}{\partial x_i}(x_1, \ldots, x_{n-1}, 0) = 0$ for $i = 1, \ldots, n - 1$, and as $\varphi$ maps $\mathbb{H}^n$ to itself,

$$\frac{\partial \varphi_n}{\partial x_n}(x_1, \ldots, x_{n-1}, 0) > 0.$$  

Now the Jacobian matrix of $\varphi$ at $q = \varphi(p) \in \partial \mathbb{H}^n$ is of the form

$$J(\varphi)(q) = \begin{pmatrix} d\Phi_q & * \\ \vdots & * \\ 0 & \cdots & 0 & \frac{\partial \varphi_n}{\partial x_n}(q) \end{pmatrix}$$  

and since $\frac{\partial \varphi_n}{\partial x_n}(q) > 0$ and by hypothesis $\det(J(\varphi)_q) > 0$, we have $\det(J(\Phi)_q) > 0$, as claimed.

In order to make Stokes' formula sign free, if $\mathbb{R}^n$ has the orientation given by $dx_1 \wedge \cdots \wedge dx_n$, then $\partial \mathbb{H}^n$ is given the orientation given by $(-1)^n dx_1 \wedge \cdots \wedge dx_{n-1}$ if $n \geq 2$, and $-1$ for $n = 1$. In particular $\partial \mathbb{H}^2$ is oriented by $e_1$ while $\partial \mathbb{H}^3$ is oriented by $-e_1 \wedge e_2 = e_2 \wedge e_1$. See Figure 24.6.
24.6. INTEGRATION ON REGULAR DOMAINS AND STOKES’ THEOREM

\[ e_2 \wedge e_1 = e_1 \wedge e_2 \]

- e_2 \wedge e_1 = e_1 \wedge e_3 - e_2 - ^ e_1 = e_1 \wedge e_2 \wedge e_3

Figure 24.6: The boundary orientations of \( \partial \mathbb{H}^2 \) and \( \partial \mathbb{H}^3 \).

**Definition 24.13.** Given any domain with smooth boundary \( N \subseteq M \), a tangent vector \( w \in T_p M \) at a boundary point \( p \in \partial N \) is **outward directed** iff there is a chart \( (U, \varphi) \) with \( p \in U \), \( \varphi(U \cap N) = \varphi(U) \cap \mathbb{H}^n \), and \( d\varphi_p(w) \) has a negative \( n^{\text{th}} \) coordinate \( pr_n(d\varphi_p(w)) \); see Figure 24.7.

Let \((V, \psi)\) be another chart with \( p \in V \). The transition map

\[ \theta = \psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V) \]

induces a map

\[ \varphi(U \cap V) \cap \mathbb{H}^n \to \psi(U \cap V) \cap \mathbb{H}^n \]

which restricts to a diffeomorphism

\[ \Theta : \varphi(U \cap V) \cap \partial \mathbb{H}^n \to \psi(U \cap V) \cap \partial \mathbb{H}^n. \]

The proof of Proposition 24.13 shows that the Jacobian matrix of \( d\theta_q \) at \( q = \varphi(p) \in \partial \mathbb{H}^n \) is of the form

\[ J(\theta)(q) = \begin{pmatrix} J(\Theta) & * \\ \vdots & J(\Theta) \\ * & \frac{\partial \Theta_n}{\partial x_n}(q) \\ 0 & \cdots & 0 \end{pmatrix} \]
Figure 24.7: An example of an outward directed tangent vector to \( N \). Notice this red tangent vector points away from \( N \).

with \( \theta = (\theta_1, \ldots, \theta_n) \), and that \( \frac{\partial \theta_n}{\partial x_n}(q) > 0 \). As \( d\psi_p = d(\psi \circ \varphi^{-1})_q \circ d\varphi_p \), we see that for any \( w \in T_pM \) with \( pr_n(d\varphi_p(w)) < 0 \), we also have \( pr_n(d\psi_p(w)) < 0 \). Therefore, the negativity condition of Definition 24.13 does not depend on the chart at \( p \). The following proposition is then easy to show:

**Proposition 24.14.** Let \( N \subseteq M \) be a domain with smooth boundary, where \( M \) is a smooth manifold of dimension \( n \).

(1) The boundary \( \partial N \) of \( N \) is a smooth manifold of dimension \( n - 1 \).

(2) Assume \( M \) is oriented. If \( n \geq 2 \), there is an induced orientation on \( \partial N \) determined as follows: For every \( p \in \partial N \), if \( v_1 \in T_pM \) is an outward directed tangent vector, then a basis \( (v_2, \ldots, v_n) \) for \( T_p\partial N \) is positively oriented iff the basis \( (v_1, v_2, \ldots, v_n) \) for \( T_pM \) is positively oriented. When \( n = 1 \), every \( p \in \partial N \) has the orientation \(+1\) iff for every outward directed tangent vector \( v_1 \in T_pM \), the vector \( v_1 \) is a positively oriented basis of \( T_pM \).

Part (2) of Proposition 24.14 is summarized as “outward pointing vector first”. When \( M \) is an \( n \)-dimensional embedded manifold in \( \mathbb{R}^m \) with an orientation preserving parametrization \( \psi : U \to \mathbb{R}^m \), for any point \( p = \psi(q) \in \partial N \), let \( v_1 \) be a tangent vector pointing away from \( N \). This means \( d\psi_q(-e_n) = v_1 \). To complete the basis of \( T_pM \) in a manner consistent with the
positive orientation of $U$ given by $dx_1 \wedge \cdots \wedge dx_n$, we choose an ordered basis $(v_2, \cdots, v_n)$ of $T_p\partial N$ such that $d\psi_q((-1)^n e_1) = v_2$ and $d\psi_q(e_i) = v_{i+1}$ whenever $2 \leq i \leq n - 1$. Intuitively, $\psi$ maps the positive orientation of $U$ to a positive orientation of $T_pM$ with the condition that the first vector in the orientation frame of $T_pM$ points away from $N$. See Figure 24.8.

![Figure 24.8: The orientation of $T_p\partial N$ consistent with the positive orientation of $\mathbb{R}^2$](image)

If $M$ is oriented, then for every $n$-form $\omega \in \mathcal{A}^n_c(M)$, the integral $\int_N \omega$ is well-defined. More precisely, Proposition 24.7 can be generalized to domains with a smooth boundary. This can be shown in various ways. The most natural way to proceed is to prove an extension of Proposition 24.6 using a slight generalization of the change of variable formula:

**Proposition 24.15.** Let $\varphi: U \to V$ be a diffeomorphism between two open subsets of $\mathbb{R}^n$, and assume that $\varphi$ maps $U \cap \mathbb{H}^n$ to $V \cap \mathbb{H}^n$. Then, for every smooth function $f: V \to \mathbb{R}$ with compact support,

$$\int_{V \cap \mathbb{H}^n} f(x)dx_1 \cdots dx_n = \int_{U \cap \mathbb{H}^n} f(\varphi(y)) |J(\varphi)_y| dy_1 \cdots dy_n.$$
One alternative way to define \( \int_N \omega \) involves covering \( N \) with special kinds of open subsets arising from regular simplices (see Warner [175], Chapter 4).

Another alternative way to proceed is to apply techniques of measure theory. In Madsen and Tornehave [119] it is argued that integration theory goes through for continuous \( n \)-forms with compact support. If \( \sigma \) is a volume form on \( M \), then for every continuous function with compact support \( f \), the map

\[
f \mapsto I_\sigma(f) = \int_M f \sigma
\]

is a linear positive operator (which means that \( I(f) \geq 0 \) for \( f \geq 0 \)). By Riesz’ representation theorem, \( I_\sigma \) determines a positive Borel measure \( \mu_\sigma \) which satisfies

\[
\int_M f d\mu_\sigma = \int_M f \sigma
\]

for all continuous functions \( f \) with compact support. Then we can set

\[
\int_N \omega = \int_M 1_N \omega,
\]

where \( 1_N \) is the function with value 1 on \( N \) and 0 outside \( N \).

We now have all the ingredient to state and prove Stokes’s formula. Our proof is based on the proof found in Section 23.5 of Tu [170]. Alternative proofs can be found in many places (for example, Warner [175] (Chapter 4), Bott and Tu [24] (Chapter 1), and Madsen and Tornehave [119] (Chapter 10).

**Theorem 24.16.** (Stokes’ Theorem) Let \( N \subseteq M \) be a domain with smooth boundary, where \( M \) is a smooth oriented manifold of dimension \( n \), give \( \partial N \) the orientation induced by \( M \), and let \( i: \partial N \to M \) be the inclusion map. For every differential form with compact support \( \omega \in \mathcal{A}^{n-1}_c(M) \), we have

\[
\int_{\partial N} i^* \omega = \int_N d\omega.
\]

In particular, if \( N = M \) is a smooth oriented manifold with boundary, then

\[
\int_{\partial M} i^* \omega = \int_M d\omega, \tag{***}
\]

and if \( M \) is a smooth oriented manifold without boundary, then

\[
\int_M d\omega = 0.
\]

Of course, \( i^* \omega \) is the restriction of \( \omega \) to \( \partial N \), and for simplicity of notation \( i^* \omega \) is usually written \( \omega \), and Stokes’ formula is written

\[
\int_{\partial N} \omega = \int_N d\omega.
\]
Proof based on Tu [170]. We select a covering \( \{(U_i, \varphi_i)\}_{i \in I} \) of \( M \) such that \( \varphi_\alpha(U_\alpha \cap N) = \varphi_i(U_i) \cap \mathbb{H}^n \) is diffeomorphic to either \( \mathbb{R}^n \) or \( \mathbb{H}^n \) via an orientation preserving diffeomorphism. Let \((\rho_i)_{i \in I}\) be a partition of unity subordinate to this cover. An adaptation of the proof of Proposition 24.7 shows that \( \rho_i \omega \) is an \((n-1)\)-form on \( M \) with compact support in \( U_i \).

Assume that Stokes’ theorem is true for \( \mathbb{R}^n \) and \( \mathbb{H}^n \). Then Stokes’ theorem will hold for all \( U_i \) which are diffeomorphic to either \( \mathbb{R}^n \) or \( \mathbb{H}^n \). Observe that the paragraph preceding Proposition 24.14 implies that \( \partial N \cap U_i = \partial U_i \). Since \( \sum_i \rho_i = 1 \), we have

\[
\int_{\partial N} \omega = \sum_i \int_{\partial N} \rho_i \omega, \\
= \sum_i \int_{\partial U_i} \rho_i \omega, \quad \text{since } \sum_i \rho_i \omega \text{ is finite} \\
= \sum_i \int_{U_i} d(\rho_i \omega), \quad \text{since supp}(\rho_i \omega) \subset U_i \\
= \sum_i \int_{U_i} d(\rho_i \omega), \quad \text{by assumption that Stokes is true for } U_i \\
= \int_N d\left( \sum_i \rho_i \omega \right) = \int_N d\omega.
\]

Thus it remains to prove Stokes’ theorem for \( \mathbb{R}^n \) and \( \mathbb{H}^n \). Since \( \omega \) is now assumed to be an \((n-1)\)-form on \( \mathbb{R}^n \) or \( \mathbb{H}^n \) with compact support,

\[
\omega = \sum_{i=1}^n f_i \, dx_1 \wedge \cdots \wedge \hat{dx}_i \wedge \cdots \wedge dx_n,
\]

where each \( f_i \) is a smooth function with compact support in \( \mathbb{R}^n \) or \( \mathbb{H}^n \). By using the \( \mathbb{R} \)-linearity of the exterior derivative and the integral operator we may assume that \( \omega \) has only one term, namely

\[
\omega = f \, dx_1 \wedge \cdots \wedge \hat{dx}_i \wedge \cdots \wedge dx_n,
\]

and

\[
d\omega = \sum_{j=1}^n \frac{\partial f}{\partial x_j} \, dx_j \wedge dx_1 \wedge \cdots \wedge \hat{dx}_i \wedge \cdots \wedge dx_n \\
= (-1)^{i-1} \frac{\partial f}{\partial x_i} \, dx_1 \wedge \cdots \wedge dx_i \wedge \cdots \wedge dx_n.
\]
where \( f \) is smooth function on \( \mathbb{R}^n \) such that \( \text{supp}(f) \) is contained in the interior of the \( n \)-cube \([-a,a]^n\) for some fixed \( a > 0 \).

To verify Stokes’ theorem for \( \mathbb{R}^n \), we evaluate \( \int_{\mathbb{R}^n} d\omega \) as an iterated integral via Fubini’s theorem. (See Edwards [66], Theorem 4.1.) In particular, we find that

\[
\int_{\mathbb{R}^n} d\omega = \int_{\mathbb{R}^n} (-1)^{i-1} \frac{\partial f}{\partial x_i} \, dx_1 \cdots dx_i \cdots dx_n
\]

\[
= (-1)^{i-1} \int_{\mathbb{R}^n} \left( \int_{-\infty}^{\infty} \frac{\partial f}{\partial x_i} \, dx_i \right) \, dx_1 \cdots \hat{dx}_i \cdots dx_n
\]

\[
= (-1)^{i-1} \int_{\mathbb{R}^n} \left( \int_{-a}^{a} \frac{\partial f}{\partial x_i} \, dx_i \right) \, dx_1 \cdots \hat{dx}_i \cdots dx_n
\]

\[
= (-1)^{i-1} \int_{\mathbb{R}^n} 0 \, dx_1 \cdots \hat{dx}_i \cdots dx_n \quad \text{since \( \text{supp}(f) \subset [-a,a]^n \)}
\]

\[
= 0 = \int_{\emptyset} \omega = \int_{\partial \mathbb{R}^n} \omega.
\]

The verification of Stokes’ theorem for \( \mathbb{H}^n \) involves the analysis of two cases. For the first case assume \( i \neq n \). Since \( \partial \mathbb{H}^n \) is given by \( x_n = 0 \), then \( dx_n \equiv 0 \) on \( \partial \mathbb{H}^n \). An application of Fubini’s theorem shows that

\[
\int_{\mathbb{H}^n} d\omega = \int_{\mathbb{H}^n} (-1)^{i-1} \frac{\partial f}{\partial x_i} \, dx_1 \cdots dx_i \cdots dx_n
\]

\[\quad = (-1)^{i-1} \int_{\mathbb{H}^n} \left( \int_{-\infty}^{\infty} \frac{\partial f}{\partial x_i} \, dx_i \right) \, dx_1 \cdots \hat{dx}_i \cdots dx_n
\]

\[\quad = (-1)^{i-1} \int_{\mathbb{H}^n} \left( \int_{-a}^{a} \frac{\partial f}{\partial x_i} \, dx_i \right) \, dx_1 \cdots \hat{dx}_i \cdots dx_n
\]

\[\quad = (-1)^{i-1} \int_{\mathbb{H}^n} 0 \, dx_1 \cdots \hat{dx}_i \cdots dx_n \quad \text{since \( \text{supp}(f) \subset [-a,a]^n \)}
\]

\[\quad = 0 = \int_{\partial \mathbb{H}^n} f \, dx_1 \cdots \hat{dx}_i \cdots dx_n, \quad \text{since \( dx_n \equiv 0 \) on \( \partial \mathbb{H}^n \).}
\]

It remains to analyze the case \( i = n \). Fubini’s theorem implies

\[
\int_{\mathbb{H}^n} d\omega = \int_{\mathbb{H}^n} (-1)^{i-1} \frac{\partial f}{\partial x_n} \, dx_1 \cdots dx_n
\]

\[\quad = (-1)^{i-1} \int_{\mathbb{H}^n} \left( \int_{0}^{\infty} \frac{\partial f}{\partial x_n} \, dx_n \right) \, dx_1 \cdots dx_{n-1}
\]

\[\quad = (-1)^{i-1} \int_{\mathbb{H}^n} \left( \int_{0}^{a} \frac{\partial f}{\partial x_n} \, dx_n \right) \, dx_1 \cdots dx_{n-1}
\]

\[\quad = (-1)^i \int_{\mathbb{H}^n} f(x^1, \cdots, x^{n-1}, 0) \, dx_1 \cdots dx_{n-1}, \quad \text{since \( \text{supp}(f) \subset [-a,a]^n \)}
\]

\[\quad = \int_{\partial \mathbb{H}^n} \omega.
\]
where the last equality follows from the fact that \((-1)^n\mathbb{R}^{n-1}\) is the induced boundary orientation of \(\partial\mathbb{R}^n\).

Stokes’ theorem, as presented in Theorem 24.16, unifies the integral theorems of vector calculus since the classical integral theorems of vector calculus are particular examples of (***\(^{\ast}\)) when \(M\) is an \(n\)-dimensional manifold embedded in \(\mathbb{R}^3\). If \(n = 3\), \(\omega \in \mathcal{A}^2(M)\), and (***\(^{\ast}\)) becomes the divergence theorem. Given a smooth \(F: \mathbb{R}^3 \to \mathbb{R}^3\), recall that the divergence of \(F\) is the smooth real-valued function \(\operatorname{div} F: \mathbb{R}^3 \to \mathbb{R}\) where

\[
\operatorname{div} F = \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} + \frac{\partial F_3}{\partial x_3},
\]

and \((x_1, x_2, x_3)\) are the standard coordinates of \(\mathbb{R}^3\) (often represented as \((x, y, z)\)). The divergence theorem is as follows:

**Proposition 24.17.** (Divergence Theorem) Let \(F: \mathbb{R}^3 \to \mathbb{R}^3\) be a smooth vector field defined on a neighborhood of \(M\), a compact oriented smooth 3-dimensional manifold with boundary. Then

\[
\int_M \operatorname{div} F \operatorname{Vol}_M = \int_{\partial M} F \cdot N \operatorname{Vol}_{\partial M},
\]

where \(N(x) = (n_1(x), n_2(x), n_3(x))\) is the unit outer normal vector field on \(\partial M\),

\[
\operatorname{Vol}_{\partial M} = \sum_{i=1}^{3} (-1)^{i-1} n_i \, dx_1 \wedge \overset{\sim}{dx}_i \wedge dx_3,
\]

and \(\partial M\) is positively oriented as the boundary of \(M \subseteq \mathbb{R}^3\).

In calculus books (1) is often written as

\[
\int \int \int \operatorname{div} F \, dx \, dy \, dz = \int F \cdot N \, dS,
\]

where \(dS\) is the surface area differential. In particular if \(\partial M\) is parametrized by \(\varphi(x, y) = (x, y, f(x, y))\),

\[
\int F \cdot N \, dS = \int \int F \cdot \left(-\frac{\partial f}{\partial x}, -\frac{\partial f}{\partial y}, 1\right) \, dx \, dy.
\]

The verification of (3) is an application of Equation (\(\ast\)) from Section 24.2. In particular

\[
J(\varphi)(x, y) = \begin{pmatrix}
1 & 0 \\
0 & 1 \\
\frac{\partial f}{\partial x} & \frac{\partial f}{\partial y}
\end{pmatrix},
\]
which in turn implies

\[ D = \det \left[ J(\varphi)^\top (x,y) J(\varphi) (x,y) \right]^{\frac{1}{2}} = \sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}. \]

Hence

\[
\begin{align*}
    n_{1,2} &= \frac{\det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}{D} = \frac{1}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}}, \\
    n_{1,3} &= \frac{\det \begin{pmatrix} \frac{\partial f}{\partial x} & 0 \\ \frac{\partial f}{\partial y} & 1 \end{pmatrix}}{D} = \frac{\frac{\partial f}{\partial y}}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}}, \\
    n_{2,3} &= \frac{\det \begin{pmatrix} 0 & \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} & 1 \end{pmatrix}}{D} = \frac{-\frac{\partial f}{\partial x}}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}}.
\end{align*}
\]

and

\[
    dS = n_{1,2} \, dx \wedge dy + n_{1,3} \, dx \wedge dz + n_{2,3} \, dy \wedge dz
\]

\[
= \frac{dx \wedge dy}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}} + \frac{\frac{\partial f}{\partial y} \, dx \wedge dz}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}} + \frac{-\frac{\partial f}{\partial x} \, dy \wedge dz}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}}.
\]

Since \( z = f(x,y) \),

\[
    \varphi^*(dS) = \frac{dx \wedge dy}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}} + \frac{\frac{\partial f}{\partial y} \, dx \wedge (\frac{\partial f}{\partial x} \, dx + \frac{\partial f}{\partial y} \, dy)}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}} + \frac{-\frac{\partial f}{\partial x} \, dy \wedge (\frac{\partial f}{\partial x} \, dx + \frac{\partial f}{\partial y} \, dy)}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}}
\]

\[
= \sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2} \, dx \wedge dy.
\]

Furthermore,

\[
    N = \frac{\frac{\partial \varphi}{\partial x} \times \frac{\partial \varphi}{\partial y}}{\left\| \frac{\partial \varphi}{\partial x} \times \frac{\partial \varphi}{\partial y} \right\|} = \frac{1}{\sqrt{1 + \left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}} \left( -\frac{\partial f}{\partial x}, -\frac{\partial f}{\partial y}, 1 \right).\]
Substituting the expressions for $N$ and $\varphi^*(dS)$ into $\int F \cdot N \, dS$ give the right side of (3).

If $n = 2$, $\omega \in A^1_c(M)$, and (*** ) becomes the classical Stokes’ theorem. Given a smooth $F : \mathbb{R}^3 \to \mathbb{R}^3$, recall that the curl of $F$ is the smooth function $\text{curl} F : \mathbb{R}^3 \to \mathbb{R}^3$

$$\text{curl} F = \left( \frac{\partial F_3}{\partial x_2} - \frac{\partial F_2}{\partial x_3}, \frac{\partial F_1}{\partial x_3} - \frac{\partial F_3}{\partial x_1}, \frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2} \right).$$

The classical Stokes’ theorem is as follows:

**Proposition 24.18.** Let $M$ be an oriented compact 2-dimensional manifold with boundary locally parametrized in $\mathbb{R}^3$ by the orientation-preserving local diffeomorphism $\psi : U \to \mathbb{R}^3$ such that $\psi(u,v) = (x_1, x_2, x_3) \in M$. Define

$$N = \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \ \left/ \left\| \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \right\| \right.$$

to be the smooth outward unit normal vector field on $M$. Let $n$ be the outward directed tangent vector field on $\partial M$. Let $T = N \times n$. Given $F : \mathbb{R}^3 \to \mathbb{R}^3$, a smooth vector field defined on a open subset of $\mathbb{R}^3$ containing $M$,

$$\int_M \text{curl} F \cdot N \, \text{Vol}_M = \int_{\partial M} F \cdot T \, \text{Vol}_{\partial M},$$

(4)

where $\text{Vol}_M$ is defined as in $\text{Vol}_{\partial M}$ of Proposition 24.17 and $\text{Vol}_{\partial M} = ds$, the line integral form.

If $M$ is parametrized by $\varphi(x,y) = (x, y, f(x, y))$, we have shown that the left side of (4) may be written as

$$\int_M \text{curl} F \cdot N \, \text{Vol}_M = \int \text{curl} F \cdot N \, dS = \int \int \text{curl} F \cdot \left( -\frac{\partial f}{\partial x}, -\frac{\partial f}{\partial y}, 1 \right) \, dx \, dy.$$

Many calculus books represent the right side of (4) as

$$\int_{\partial M} F \cdot T \, ds = \int F \cdot dr,$$

(5)

where $dr = (dx, dy, dz)$. Once again the verification of (5) is an application of Equation (**) from Section 24.2. Let $\psi(x) = (x, y(x), z(x))$. Then $J(\psi)(x) = (1, y_x, z_x)^\top$, where $y_x = \frac{dy}{dx}$ and $z_x = \frac{dz}{dx}$. Then

$$D = \det \left[ J(\psi)^\top(x)J(\psi)(x) \right]^{\frac{1}{2}} = \sqrt{1 + y_x^2 + z_x^2},$$

$$ds = \frac{dx + y_x \, dy + z_x \, dz}{\sqrt{1 + y_x^2 + z_x^2}},$$
\[ \psi^*ds = \sqrt{1 + y_x^2 + z_x^2} \, dx. \]

Furthermore
\[ T = \frac{J(\psi)(x)}{\sqrt{1 + y_x^2 + z_x^2}} = \frac{(1, y_x, z_x)^\top}{\sqrt{1 + y_x^2 + z_x^2}}. \]

Substituting the expressions for \( T \) and \( \psi^*ds \) into the left side of (5) gives
\[ \int_{\partial M} F \cdot T \, ds = \int F \cdot \left( \frac{dy}{dx} \frac{dz}{dx} \right) dx = \int F \cdot (dx, dy, dz) = \int F \cdot dr. \]

Thus the classical form of Stokes’ theorem often appears as
\[ \int \int \text{curl} F \cdot \left( -\frac{\partial f}{\partial x}, -\frac{\partial f}{\partial y}, 1 \right) \, dx \, dy = \int F \cdot \left( 1, \frac{dy}{dx}, \frac{dz}{dx} \right) \, dx = \int F \cdot dr, \]

where \( M \) is parametrized via \( \varphi(x, y) = (x, y, f(x, y)) \).

The orientation frame \((n, T, N)\) given in Proposition 24.18 provides the standard orientation of \( \mathbb{R}^3 \) given by \((e_1, e_2, e_3)\) and is visualized as follows. Pick a preferred side of the surface. This choice is represented by \( N \). At each boundary point, draw the outward pointing tangent vector \( n \) which is locally perpendicular (in the tangent plane) to the boundary curve. To determine \( T \), pretend you are a bug on the side of the surface selected by \( N \). You must walk along the boundary curve in the direction that keeps the boundary of the surface your right. Then \( T = N \times n \) and \((n, T, N)\) is oriented via the right-hand rule in the same manner as \((e_1, e_2, e_3)\); see Figure 24.9.

![Figure 24.9: The orientation frame \((n, T, N)\) for the bell shaped surface \( M \). Notice the bug must walk along the boundary in a counter clockwise direction.](image-url)
For those readers who wish to learn more about the connections between the classical integration theorems of vector calculus and Stokes’ theorem, we refer them to Edwards [66] (Chapter 5, Section 7).

The version of Stokes’ theorem that we have presented applies to domains with smooth boundaries, but there are many situations where it is necessary to deal with domains with singularities, for example corners (as a cube, a tetrahedron, etc.). Manifolds with corners form a nice class of manifolds that allow such a generalization of Stokes’ theorem.

To model corners, we adapt the idea that we used when we defined charts of manifolds with boundaries but instead of using the closed half space $H^m$, we use the closed convex cone $R_+^m = \{(x_1, \ldots, x_m) \in \mathbb{R}^m \mid x_1 \geq 0, \ldots, x_m \geq 0\}$.

The boundary $\partial R_+^m$ of $R_+^m$ is the space $\partial R_+^m = \{(x_1, \ldots, x_m) \in \mathbb{R}^m \mid x_1 \geq 0, \ldots, x_m \geq 0, x_i = 0 \text{ for some } i\}$, which can also be written as $\partial R_+^m = H_1 \cup \cdots \cup H_m$,

with $H_i = \{(x_1, \ldots, x_m) \in \mathbb{R}_+^m \mid x_i = 0\}$.

The set of corner points of $R_+^m$ is the subset $\{(x_1, \ldots, x_m) \in \mathbb{R}_+^m \mid \exists i \exists j (i \neq j), x_i = 0 \text{ and } x_j = 0\}$.

Equivalently, the set of corner points is the union of all intersections $H_{i_1} \cap \cdots \cap H_{i_k}$ for all finite subsets $\{i_1, \ldots, i_k\}$ of $\{1, \ldots, m\}$ with $k \geq 2$. See Figure 24.10.

For a topological space $M$, a chart with corners is a pair $(U, \varphi)$ where $U$ is some open subset of $M$ and $\varphi$ is a homeomorphism of $U$ onto some open subset of $\mathbb{R}_+^m$ (with the subspace
topology of $\mathbb{R}^m$). Compatible charts, atlases, equivalent atlases are defined as usual, and a smooth manifold with corners is a topological space together with an equivalence class of atlases of charts with corners.

A point $p \in M$ is a corner point if there is a chart $(U, \varphi)$ with $p \in U$ such that $\varphi(p)$ is a corner point of $\mathbb{R}^m$. It is not hard to show that this definition does not depend on the chart $(U, \varphi)$ with $p \in U$. See Figure 24.11

![Figure 24.11](image)

Figure 24.11: The three types of charts on $M$, a manifold with corners. Note that $p_2$ is a corner point of $M$.

Now, in general, the boundary of a smooth manifold with corners is not a smooth manifold with corners. For example, $\partial \mathbb{R}^m_+$ is not a smooth manifold with corners, but it is the union of smooth manifolds with corners, since $\partial \mathbb{R}^m_+ = H_1 \cup \cdots \cup H_m$, and each $H_i$ is a smooth manifold with corners. We can use this fact to define $\int_{\partial M} \omega$ where $\omega$ is an $(n-1)$-form whose support in contained in the domain of a chart with corners $(U, \varphi)$ by setting

$$
\int_{\partial M} \omega = \sum_{i=1}^{m} \int_{H_i} (\varphi^{-1})^* \omega,
$$

where each $H_i$ is given a suitable orientation. Then, it is not hard to prove a version of Stokes’ theorem for manifolds with corners. For a detailed exposition, see Lee [117], Chapter
14. An even more general class of manifolds with singularities (in $\mathbb{R}^N$) for which Stokes’ theorem is valid is discussed in Lang [114] (Chapter XVII. §3).

24.7 Integration on Riemannian Manifolds and Lie Groups

We saw in Section 24.2 that every orientable Riemannian manifold has a uniquely defined volume form $\text{Vol}_M$ (see Proposition 24.4).

**Definition 24.14.** Given any smooth real-valued function $f$ with compact support on $M$, we define the integral of $f$ over $M$ by

$$\int_M f = \int_M f \text{Vol}_M.$$  

Actually it is possible to define the integral $\int_M f$ using densities even if $M$ is not orientable, but we do not need this extra generality. If $M$ is compact, then $\int_M 1_M = \int_M \text{Vol}_M$ is the volume of $M$ (where $1_M$ is the constant function with value 1).

If $M$ and $N$ are Riemannian manifolds, then we have the following version of Proposition 24.8 (3):

**Proposition 24.19.** If $M$ and $N$ are oriented Riemannian manifolds and if $\varphi: M \to N$ is an orientation preserving diffeomorphism, then for every function $f \in C^\infty(N)$ with compact support, we have

$$\int_N f \text{Vol}_N = \int_M f \circ \varphi | \det(d\varphi)| \text{Vol}_M,$$

where $f \circ \varphi | \det(d\varphi)$ denotes the function $p \mapsto f(\varphi(p)) | \det(d\varphi_p)|$, with $d\varphi_p: T_pM \to T_{\varphi(p)}N$. In particular, if $\varphi$ is an orientation preserving isometry (see Definition 10.4), then

$$\int_N f \text{Vol}_N = \int_M f \circ \varphi \text{Vol}_M.$$

We often denote $\int_M f \text{Vol}_M$ by $\int_M f(t) dt$.

If $f: M \to \mathbb{C}$ is a smooth complex valued function then we can write $f = u + iv$ for two real-valued functions $u: M \to \mathbb{R}$ and $v: M \to \mathbb{R}$ with $u(p) = \Re(f(p))$ and $v(p) = \Im(f(p))$ for all $p \in M$. Then, if $f$ has compact support so do $u$ and $v$, and we define $\int_M f \text{Vol}_M$ by

$$\int_M f \text{Vol}_M = \int_M u \text{Vol}_M + i \int_M v \text{Vol}_M.$$

If $G$ is a Lie group, we know from Section 24.2 that $G$ is always orientable and that $G$ possesses left-invariant volume forms. Since $\dim(\bigwedge^n g^*) = 1$ if $\dim(G) = n$, and since
every left-invariant volume form is determined by its value at the identity, the space of left-invariant volume forms on $G$ has dimension 1. If we pick some left-invariant volume form $\omega$ defining the orientation of $G$, then every other left-invariant volume form is proportional to $\omega$.

Given any smooth real-valued function $f$ with compact support on $G$, we define the integral of $f$ over $G$ (w.r.t. $\omega$) by

$$\int_G f = \int_G f \omega.$$ 

This integral depends on $\omega$, but since $\omega$ is defined up to some positive constant, so is the integral. When $G$ is compact, we usually pick $\omega$ so that

$$\int_G \omega = 1.$$ 

If $f: G \to \mathbb{C}$ is a smooth complex valued-function then we can write $f = u + iv$ for two real-valued functions $u: G \to \mathbb{R}$ and $v: G \to \mathbb{R}$ as before and we define

$$\int_G f \omega = \int_G u \omega + i \int_M v \omega.$$ 

For every $g \in G$, as $\omega$ is left-invariant, $L_g^* \omega = \omega$, so $L_g^*$ is an orientation-preserving diffeomorphism, and by Proposition 24.8 (3),

$$\int_G f \omega = \int_G L_g^*(f \omega),$$ 

so using Proposition 23.9, we get

$$\int_G f = \int_G f \omega = \int_G L_g^*(f \omega) = \int_G L_g^* f L_g^* \omega = \int_G L_g^* f \omega = \int_G (f \circ L_g) \omega = \int_G f \circ L_g.$$ 

Thus we proved the following proposition.

**Proposition 24.20.** Given any left-invariant volume form $\omega$ on a Lie group $G$, for any smooth function $f$ with compact support, we have

$$\int_G f = \int_G f \circ L_g,$$

a property called left-invariance.

It is then natural to ask when our integral is right-invariant; that is, when

$$\int_G f = \int_G f \circ R_g.$$
Observe that $R_g^*\omega$ is left-invariant, since

$$L_h^*R_g^*\omega = R_g^*L_h^*\omega = R_g^*\omega.$$ 

It follows that $R_g^*\omega$ is some constant multiple of $\omega$, and so there is a function $\overline{\Delta}: G \to \mathbb{R}$ such that

$$R_g^*\omega = \overline{\Delta}(g)\omega.$$

One can check that $\overline{\Delta}$ is smooth, and we let

$$\Delta(g) = |\overline{\Delta}(g)|.$$

Clearly,

$$\Delta(gh) = \Delta(g)\Delta(h),$$

so $\Delta$ is a homomorphism of $G$ into $\mathbb{R}_+$. The function $\Delta$ is called the modular function of $G$.

**Proposition 24.21.** Given any left-invariant volume form $\omega$ on a Lie group $G$, for any smooth function $f$ with compact support, we have

$$\int_G f\omega = \Delta(g) \int_G (f \circ R_g)\omega.$$

**Proof.** By Proposition 24.8 (3), as $R_g^*$ is an orientation-preserving diffeomorphism,

$$\int_G f\omega = \int_G R_g^*(f\omega) = \int_G R_g^*f R_g^*\omega = \int_G (f \circ R_g)\Delta(g)\omega,$$

or equivalently,

$$\int_G f\omega = \Delta(g) \int_G (f \circ R_g)\omega,$$

which is the desired formula. \qed

Consequently, our integral is right-invariant iff $\Delta \equiv 1$ on $G$. Thus, our integral is not always right-invariant. When it is, i.e. when $\Delta \equiv 1$ on $G$, we say that $G$ is unimodular. This happens in particular when $G$ is compact, since in this case,

$$1 = \int_G \omega = \int_G 1_G\omega = \int_G \Delta(g)\omega = \Delta(g) \int_G \omega = \Delta(g),$$

for all $g \in G$. Therefore, for a compact Lie group $G$, our integral is both left and right invariant. We say that our integral is bi-invariant.

As a matter of notation, the integral $\int_G f = \int_G f\omega$ is often written $\int_G f(g)dg$. Then left-invariance can be expressed as

$$\int_G f(g)dg = \int_G f(hg)dg,$$
and right-invariance as
\[ \int_G f(g) dg = \int_G f(gh) dg, \]
for all \( h \in G \).

If \( \omega \) is left-invariant, then it can be shown (see Dieudonné [55], Chapter XIV, Section 3) that
\[ \int_G f(g^{-1}) \Delta(g^{-1}) dg = \int_G f(g) dg. \]
Consequently, if \( G \) is unimodular, then
\[ \int_G f(g^{-1}) dg = \int_G f(g) dg. \]

If \( \omega_l \) is any left-invariant volume form on \( G \) and if \( \omega_r \) is any right-invariant volume form on \( G \), then
\[ \omega_r(g) = c \Delta(g^{-1}) \omega_l(g), \]
for some constant \( c \neq 0 \).

Indeed, define the form \( \omega \) by \( \omega(g) = \Delta(g^{-1}) \omega_l(g) \). By Proposition 23.9 we have
\[
(R_h^* \omega)_h = \Delta((gh)^{-1})(R_h^* \omega_l)_h \\
= \Delta(h^{-1}) \Delta(g^{-1}) \Delta(h)(\omega_l)_h \\
= \Delta(g^{-1})(\omega_l)_h,
\]
which shows that \( \omega \) is right-invariant, and thus \( \omega_r(g) = c \Delta(g^{-1}) \omega_l(g) \), as claimed (since \( \Delta(g^{-1}) = \pm \Delta(g^{-1}) \)).

Actually it is not difficult to prove that
\[ \Delta(g) = |\det(\text{Ad}(g^{-1}))|. \]

For this recall that \( \text{Ad}(g) = d(L_g \circ R_{g^{-1}})_1 \). For any left-invariant \( n \)-form \( \omega \in \bigwedge^n g^* \), we claim that
\[ (R_g^* \omega)_h = \det(\text{Ad}(g^{-1})) \omega_h, \]
which shows that $\Delta(g) = |\det(\text{Ad}(g^{-1}))|$. Indeed, for all $v_1, \ldots, v_n \in T_h G$, we have

$$(R_g^*\omega)(v_1, \ldots, v_n)$$

$$= \omega_{hg}(d(R_g)h(v_1), \ldots, d(R_g)h(v_n))$$

$$= \omega_{hg}(d(L_g \circ L_{g^{-1}} \circ R_g \circ L_h \circ L_{h^{-1}})h(v_1), \ldots, d(L_g \circ L_{g^{-1}} \circ R_g \circ L_h \circ L_{h^{-1}})h(v_n))$$

$$= \omega_{hg}(d(L_h \circ L_g \circ L_{g^{-1}} \circ R_g \circ L_{h^{-1}})h(v_1), \ldots, d(L_h \circ L_g \circ L_{g^{-1}} \circ R_g \circ L_{h^{-1}})h(v_n))$$

$$= \omega_{hg}(d(L_h g_1)(\text{Ad}(g^{-1}))(d(L_{h^{-1}})h(v_1)), \ldots, d(L_h g_1)(\text{Ad}(g^{-1}))(d(L_{h^{-1}})h(v_n)))$$

$$= \omega_1(\text{Ad}(g^{-1}))(d(L_{h^{-1}})h(v_1)), \ldots, \text{Ad}(g^{-1}))(d(L_{h^{-1}})h(v_n))$$

$$= \det(\text{Ad}(g^{-1}))\omega_1(d(L_{h^{-1}})h(v_1), \ldots, d(L_{h^{-1}})h(v_n))$$

$$= \det(\text{Ad}(g^{-1}))\omega_1(d(L_{h^{-1}})h(v_1), \ldots, d(L_{h^{-1}})h(v_n))$$

$$= \det(\text{Ad}(g^{-1})) \omega(h(v_1, \ldots, v_n),$$

where we used the left-invariance of $\omega$ twice.

In general, if $G$ is not unimodular then $\omega_l \neq \omega_r$. A simple example provided by Vinroot [174] is the group $G$ of direct affine transformations of the real line, which can be viewed as the group of matrices of the form

$$g = \begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix}, \quad x, y, \in \mathbb{R}, \ x > 0.$$

Let $A = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \in G$ and define $T: G \to G$ as

$$T(g) = Ag = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} ax & ay + b \\ 0 & 1 \end{pmatrix}.$$ 

Since $G$ is homeomorphic to $\mathbb{R}^+ \times \mathbb{R}$, $T(g)$ is also represented by $T(x, y) = (ax, ay + b)$. Then the Jacobian matrix of $T$ is given by

$$J(T)(x, y) = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix},$$

which implies that $\det(J(T)(x, y)) = a^2$. Let $F: G \to \mathbb{R}^+$ be a smooth function on $G$ with compact support. Furthermore assume that $\Theta(x, y) = F(x, y)x^{-2}$ is also smooth on $G$ with compact support. Since $\Theta \circ T(x, y) = \Theta(ax, ay + b) = F(ax, ay + b)(ax)^{-2}$, Proposition 24.19 implies that

$$\int_G F(x, y)x^{-2}dx dy = \int_G \Theta(x, y) \circ T \det(J(T)(x, y)) dx dy$$

$$= \int_G F(ax, ay + b)(ax)^{-2}a^2 dx dy = \int_G F \circ T \cdot x^{-2} dx dy.$$
In summary we have shown for $g = \begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix}$, we have

$$\int_G F(Ag)x^{-2} dx \, dy = \int_G F(g)x^{-2} dx \, dy$$

which implies that the left-invariant volume form on $G$ is

$$\omega_l = \frac{dx \, dy}{x^2}.$$

To define a right-invariant volume form on $G$, define $S: G \to G$ as

$$S(g) = gA = \begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} ax & bx + y \\ 0 & 1 \end{pmatrix},$$

which is represented by $S(x, y) = (ax, bx + y)$. Then the Jacobian matrix of $S$ is

$$J(S)(x, y) = \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix},$$

and $\det(J(S)(x, y)) = a$. Using $F(x, y)$ as above and $\Theta(x, y) = F(x, y)x^{-1}$, we find that

$$\int_G F(x, y)x^{-1} dx \, dy = \int_G \Theta(x, y) \circ S \mid \det(J(S)(x, y)) \mid dx \, dy$$

$$= \int_G F(ax, ay + b)(ax)^{-1} a \, dx \, dy = \int_G F \circ Sx^{-1} dx \, dy,$$

which implies that

$$\omega_r = \frac{dx \, dy}{x}.$$

Note that $\Delta(g) = |x^{-1}|$.

**Remark:** By the Riesz’ representation theorem, $\omega$ defines a positive measure $\mu_\omega$ which satisfies

$$\int_G f \, d\mu_\omega = \int_G f \omega.$$

Using what we have shown, this measure is left-invariant. Such measures are called *left Haar measures*, and similarly we have *right Haar measures*.

It can be shown that every two left Haar measures on a Lie group are proportional (see Knapp, [106], Chapter VIII). Given a left Haar measure $\mu$, the function $\Delta$ such that

$$\mu(R_g h) = \Delta(g) \mu(h)$$


for all \( g, h \in G \) is the modular function of \( G \). However, beware that some authors, including Knapp, use \( \Delta(g^{-1}) \) instead of \( \Delta(g) \). As above, we have

\[
\Delta(g) = |\det(\text{Ad}(g^{-1}))|.
\]

Beware that authors who use \( \Delta(g^{-1}) \) instead of \( \Delta(g) \) give a formula where \( \text{Ad}(g) \) appears instead of \( \text{Ad}(g^{-1}) \). Again, \( G \) is unimodular iff \( \Delta \equiv 1 \).

It can be shown that compact, semisimple, reductive, and nilpotent Lie groups are unimodular (for instance, see Knapp, [106], Chapter VIII). On such groups, left Haar measures are also right Haar measures (and vice versa). In this case, we can speak of Haar measures on \( G \). For more details on Haar measures on locally compact groups and Lie groups, we refer the reader to Folland [67] (Chapter 2), Helgason [87] (Chapter 1), and Dieudonné [55] (Chapter XIV).
Chapter 25

Distributions and the Frobenius Theorem

25.1 Tangential Distributions, Involutive Distributions

Given any smooth manifold $M$ (of dimension $n$), for any smooth vector field $X$ on $M$, we know from Section 8.3 that for every point $p \in M$, there is a unique maximal integral curve through $p$. Furthermore, any two distinct integral curves do not intersect each other, and the union of all the integral curves is $M$ itself. A nonvanishing vector field $X$ can be viewed as the smooth assignment of a one-dimensional vector space to every point of $M$, namely $p \mapsto \mathbb{R}X_p \subseteq T_pM$, where $\mathbb{R}X_p$ denotes the line spanned by $X_p$. Thus, it is natural to consider the more general situation where we fix some integer $r$, with $1 \leq r \leq n$, and we have an assignment $p \mapsto D(p) \subseteq T_pM$, where $D(p)$ is some $r$-dimensional subspace of $T_pM$ such that $D(p)$ “varies smoothly” with $p \in M$. Is there a notion of integral manifold for such assignments? Do they always exist?

It is indeed possible to generalize the notion of integral curve and to define integral manifolds, but unlike the situation for vector fields ($r = 1$), not every assignment $D$ as above possess an integral manifold. However, there is a necessary and sufficient condition for the existence of integral manifolds given by the Frobenius Theorem. This theorem has several equivalent formulations. First we will present a formulation in terms of vector fields. Then we show that there are advantages in reformulating the notion of involutivity in terms of differential ideals, and we state a differential form version of the Frobenius Theorem. The above versions of the Frobenius Theorem are “local.” We will briefly discuss the notion of foliation and state a global version of the Frobenius Theorem.

Since Frobenius’ Theorem is a standard result of differential geometry, we will omit most proofs, and instead refer the reader to the literature. A complete treatment of Frobenius’ Theorem can be found in Warner [175], Morita [133], and Lee [117].

Our first task is to define precisely what we mean by a smooth assignment $p \mapsto D(p) \subseteq T_pM$, where $D(p)$ is an $r$-dimensional subspace. Recall that the definition of immersed
submanifold is given by Definition 7.17.

**Definition 25.1.** Let $M$ be a smooth manifold of dimension $n$. For any integer $r$, with $1 \leq r \leq n$, an $r$-dimensional **tangential distribution** (for short, a **distribution**) is a map $D: M \to TM$, such that

(a) $D(p) \subseteq T_p M$ is an $r$-dimensional subspace for all $p \in M$.

(b) For every $p \in M$, there is some open subset $U$ with $p \in U$, and $r$ smooth vector fields $X_1, \ldots, X_r$ defined on $U$, such that $(X_1(q), \ldots, X_r(q))$ is a basis of $D(q)$ for all $q \in U$.

We say that $D$ is **locally spanned** by $X_1, \ldots, X_r$.

An immersed submanifold $N$ of $M$ is an **integral manifold** of $D$ iff $D(p) = T_p N$ for all $p \in N$.

We say that $D$ is **completely integrable** iff there exists an integral manifold of $D$ through every point of $M$.

We also write $D_p$ for $D(p)$.

**Remarks:**

1. An $r$-dimensional distribution $D$ is just a smooth subbundle of $TM$.

2. An integral manifold is only an immersed submanifold, not necessarily an embedded submanifold.

3. Some authors (such as Lee) reserve the locution “completely integrable” to a seemingly strongly condition (See Lee [117], Chapter 19, page 500). This condition is in fact equivalent to “our” definition (which seems the most commonly adopted).

4. Morita [133] uses a stronger notion of integral manifold. Namely, an integral manifold is actually an embedded manifold. Most of the results including Frobenius Theorem still hold, but maximal integral manifolds are immersed but not embedded manifolds, and this is why most authors prefer to use the weaker definition (immersed manifolds).

Here is an example of a distribution which does not have any integral manifolds: This is the two-dimensional distribution in $\mathbb{R}^3$ spanned by the vector fields

$$X = \frac{\partial}{\partial x} + y \frac{\partial}{\partial z}, \quad Y = \frac{\partial}{\partial y}.$$ 

To show why this distribution is not integrable, we will need an involutivity condition. Here is the definition.

**Definition 25.2.** Let $M$ be a smooth manifold of dimension $n$ and let $D$ be an $r$-dimensional distribution on $M$. For any smooth vector field $X$, we say that $X$ **belongs to** $D$ (or **lies in** $D$) iff $X_p \in D_p$ for all $p \in M$. We say that $D$ is **involutive** iff for any two smooth vector fields $X, Y$ on $M$, if $X$ and $Y$ belong to $D$, then $[X, Y]$ also belongs to $D$. 

Proposition 25.1. Let $M$ be a smooth manifold of dimension $n$. If an $r$-dimensional distribution $D$ is completely integrable, then $D$ is involutive.

Proof. A proof can be found in in Warner [175] (Chapter 1), and Lee [117] (Proposition 19.3). These proofs use Proposition 8.5. Another proof is given in Morita [133] (Section 2.3), but beware that Morita defines an integral manifold to be an embedded manifold.

In the example before Definition 25.1, we have

$$[X, Y] = -\frac{\partial}{\partial z},$$

so this distribution is not involutive. Therefore, by Proposition 25.1, this distribution is not completely integrable.

25.2 Frobenius Theorem

Frobenuis’ Theorem asserts that the converse of Proposition 25.1 holds. Although we do not intend to prove it in full, we would like to explain the main idea of the proof of Frobenius’ Theorem. It turns out that the involutivity condition of two vector fields is equivalent to the commutativity of their corresponding flows, and this is the crucial fact used in the proof.

Given a manifold, $M$, we say that two vector fields $X$ and $Y$ are mutually commutative iff $[X, Y] = 0$.

For example, on $\mathbb{R}^2$, the vector fields $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are commutative, but $\frac{\partial}{\partial x}$ and $x\frac{\partial}{\partial y}$ are not.

Recall from Definition 8.9 that we denote by $\Phi^X$ the (global) flow of the vector field $X$. For every $p \in M$, the map $t \mapsto \Phi^X(t, p) = \gamma_p(t)$ is the maximal integral curve through $p$. We also write $\Phi_t(p)$ for $\Phi^X(t, p)$ (dropping $X$). Recall that the map $p \mapsto \Phi_t(p)$ is a diffeomorphism on its domain (an open subset of $M$). For the next proposition, given two vector fields $X$ and $Y$, we write $\Phi$ for the flow associated with $X$ and $\Psi$ for the flow associated with $Y$.

Proposition 25.2. Given a manifold $M$, for any two smooth vector fields $X$ and $Y$, the following conditions are equivalent:

1. $X$ and $Y$ are mutually commutative (i.e. $[X, Y] = 0$).
2. $Y$ is invariant under $\Phi_t$; that is, $(\Phi_t)_* Y = Y$, whenever the left-hand side is defined.
3. $X$ is invariant under $\Psi_s$; that is, $(\Psi_s)_* X = X$, whenever the left-hand side is defined.
4. The maps $\Phi_t$ and $\Psi_t$ are mutually commutative. This means that

$$\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t,$$

for all $s, t$ such that both sides are defined.
(5) $\mathcal{L}_X Y = [X, Y] = 0$.

(6) $\mathcal{L}_Y X = [Y, X] = 0$.

(In (5) $\mathcal{L}_X Y$ is the Lie derivative and similarly in (6).)

Proof. A proof can be found in Lee [117] (Chapter 18, Proposition 18.5) and in Morita [133] (Chapter 2, Proposition 2.18). For example, to prove the implication \( (2) \implies (4) \), we observe that if $\varphi$ is a diffeomorphism on some open subset $U$ of $M$, then the integral curves of $\varphi_* Y$ through a point $p \in M$ are of the form $\varphi \circ \gamma$, where $\gamma$ is the integral curve of $Y$ through $\varphi^{-1}(p)$. Consequently, the local one-parameter group generated by $\varphi_* Y$ is $\{ \varphi \circ \Psi_s \circ \varphi^{-1} \}$. If we apply this to $\varphi = \Phi_t$, as $(\Phi_t)_* Y = Y$, we get $\Phi_t \circ \Psi_s \circ \Phi_t^{-1} = \Psi_s$, and hence $\Phi_t \circ \Psi_s = \Psi_s \circ \Phi_t$. \qed

In order to state our first version of the Frobenius Theorem we make the following definition:

**Definition 25.3.** Let $M$ be a smooth manifold of dimension $n$. Given any smooth $r$-dimensional distribution $D$ on $M$, a chart $(U, \varphi)$ is flat for $D$ iff

$$\varphi(U) \cong U' \times U'' \subseteq \mathbb{R}^r \times \mathbb{R}^{n-r},$$

where $U'$ and $U''$ are connected open subsets such that for every $p \in U$, the distribution $D$ is spanned by the vector fields

$$\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_r}.$$

If $(U, \varphi)$ is flat for $D$, then each slice of $(U, \varphi)$

$$S_c = \{ q \in U \mid x_{r+1} = c_{r+1}, \ldots, x_n = c_n \},$$

is an integral manifold of $D$, where $x_i = pr_i \circ \varphi$ is the $i$th-coordinate function on $U$ and $c = (c_{r+1}, \ldots, c_n) \in \mathbb{R}^{n-r}$ is a fixed vector. See Figure 25.1.

**Theorem 25.3.** (Frobenius) Let $M$ be a smooth manifold of dimension $n$. A smooth $r$-dimensional distribution $D$ on $M$ is completely integrable iff it is involutive. Furthermore, for every $p \in U$, there is flat chart $(U, \varphi)$ for $D$ with $p \in U$ so that every slice of $(U, \varphi)$ is an integral manifold of $D$.

Proof. A proof of Theorem 25.3 can be found in Warner [175] (Chapter 1, Theorem 1.60), Lee [117] (Chapter 19, Theorem 19.10), and Morita [133] (Chapter 2, Theorem 2.17). Since we already have Proposition 25.1, it is only necessary to prove that if a distribution is involutive, then it is completely integrable. Here is a sketch of the proof, following Morita.

Pick any $p \in M$. As $D$ is a smooth distribution, we can find some chart $(U, \varphi)$ with $p \in U$, and some vector fields $Y_1, \ldots, Y_r$ so that $Y_1(q), \ldots, Y_r(q)$ are linearly independent and span $D_q$ for all $q \in U$. Locally, we can write

$$Y_i = \sum_{j=1}^n a_{ij} \frac{\partial}{\partial x_j}, \quad i = 1, \ldots, r.$$
25.2. **FROBENIUS THEOREM**

Since $Y_1, \ldots, Y_r$ are linearly independent, the $r \times n$ matrix $(a_{ij})$ has rank $r$, so by renumbering the coordinates if necessary, we may assume that the first $r$ columns are linearly independent in which case the $r \times r$ matrices

$$A(q) = (a_{ij}(q)) \quad q \in U$$

are invertible. Then the inverse matrices $B(q) = A^{-1}(q)$ define $r \times r$ functions $b_{ij}(q)$, and let

$$X_i = \sum_{j=1}^{r} b_{ij}Y_j, \quad j = 1, \ldots, r.$$ 

Now, in matrix form

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix} = (A \quad R) \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}$$

Figure 25.1: A flat chart for the solid ball $B^3$. Each slice in $\varphi(U)$ is parallel to the $xy$-plane and turns into a cap shape inside of $B^3$. 
for some \( r \times (n - r) \) matrix \( R \), and

\[
\begin{pmatrix}
X_1 \\
\vdots \\
X_r
\end{pmatrix}
= B
\begin{pmatrix}
Y_1 \\
\vdots \\
Y_r
\end{pmatrix},
\]

so we get

\[
\begin{pmatrix}
X_1 \\
\vdots \\
X_r
\end{pmatrix}
= (I \ B R)
\begin{pmatrix}
\frac{\partial}{\partial x_1} \\
\vdots \\
\frac{\partial}{\partial x_n}
\end{pmatrix},
\]

that is,

\[
X_i = \frac{\partial}{\partial x_i} + \sum_{j=r+1}^{n} c_{ij} \frac{\partial}{\partial x_j}, \quad i = 1, \ldots, r, \tag{\ast}
\]

where the \( c_{ij} \) are functions defined on \( U \). Obviously, \( X_1, \ldots, X_r \) are linearly independent
and they span \( D_q \) for all \( q \in U \). Since \( D \) is involutive, there are some functions \( f_k \) defined
on \( U \) so that

\[
[X_i, X_j] = \sum_{k=1}^{r} f_k X_k.
\]

On the other hand, by (\ast), each \( [X_i, X_j] \) is a linear combination of \( \frac{\partial}{\partial x_{r+1}}, \ldots, \frac{\partial}{\partial x_n} \). Using (\ast),
we obtain

\[
[X_i, X_j] = \sum_{k=1}^{r} f_k X_k = \sum_{k=1}^{r} f_k \frac{\partial}{\partial x_k} + \sum_{k=1}^{r} \sum_{j=r+1}^{n} f_k c_{kj} \frac{\partial}{\partial x_j},
\]

and since this is supposed to be a linear combination of \( \frac{\partial}{\partial x_{r+1}}, \ldots, \frac{\partial}{\partial x_n} \), we must have \( f_k = 0 \)
for \( k = 1, \ldots, r \), which shows that

\[
[X_i, X_j] = 0, \quad 1 \leq i, j \leq r;
\]

that is, the vector fields \( X_1, \ldots, X_r \) are mutually commutative.

Let \( \Phi^i_t \) be the local one-parameter group associated with \( X_i \). By Proposition 25.2 (4),
the \( \Phi^i_t \) commute; that is,

\[
\Phi^i_t \circ \Phi^j_s = \Phi^j_s \circ \Phi^i_t \quad 1 \leq i, j \leq r,
\]

whenever both sides are defined. We can pick a sufficiently small open subset \( V \) in \( \mathbb{R}^r \)
containing the origin and define the map \( \Phi: V \rightarrow U \) by

\[
\Phi(t_1, \ldots, t_r) = \Phi^1_{t_1} \circ \cdots \circ \Phi^r_{t_r}(p).
\]

Clearly, \( \Phi \) is smooth, and using the fact that each \( X_i \) is invariant under each \( \Phi^j_s \) for \( j \neq i \),
and

\[
d\Phi^i_p \left( \frac{\partial}{\partial t_i} \right) = X_i(p),
\]
25.2. FROBENIUS THEOREM

we get
\[ d\Phi_p \left( \frac{\partial}{\partial t_i} \right) = X_i(p). \]

As \( X_1, \ldots, X_r \) are linearly independent, we deduce that \( d\Phi_p : T_0\mathbb{R}^r \to T_p M \) is an injection, and thus we may assume by shrinking \( V \) if necessary that our map \( \Phi : V \to M \) is an embedding. But then, \( N = \Phi(V) \) is an immersed submanifold of \( M \), and it only remains to prove that \( N \) is an integral manifold of \( D \) through \( p \).

Obviously, \( T_p N = D_p \), so we just have to prove that \( T_q N = D_q N \) for all \( q \in N \). Now, for every \( q \in N \), we can write
\[ q = \Phi(t_1, \ldots, t_r) = \Phi_{t_1} \circ \cdots \circ \Phi_{t_r}(p), \]
for some \( (t_1, \ldots, t_r) \in V \). Since the \( \Phi_i \) commute for any \( i \), with \( 1 \leq i \leq r \), we can write
\[ q = \Phi_{t_i} \circ \Phi_{t_1} \circ \cdots \circ \Phi_{t_{i-1}} \circ \Phi_{t_{i+1}} \circ \cdots \circ \Phi_{t_r}(p). \]
If we fix all the \( t_j \) but \( t_i \) and vary \( t_i \) by a small amount, we obtain a curve in \( N \) through \( q \), and this is an orbit of \( \Phi_i \). Therefore, this curve is an integral curve of \( X_i \) through \( q \) whose velocity vector at \( q \) is equal to \( X_i(q) \), and so \( X_i(q) \in T_q N \). Since the above reasoning holds for all \( i \), we get \( T_q N = D_q \), as claimed. Therefore, \( N \) is an integral manifold of \( D \) through \( p \). \( \square \)

To best understand how the proof of Theorem 25.3 constructs the integral manifold \( N \), we provide the following example found in Chapter 19 of Lee [117]. Let \( D \subset T\mathbb{R}^3 \) be the distribution
\[ Y_1 := V = x \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + x(y+1) \frac{\partial}{\partial z} \]
\[ Y_2 := W = \frac{\partial}{\partial x} + y \frac{\partial}{\partial z}. \]
Given \( f \in C^\infty(\mathbb{R}^3) \), observe that
\[ [V, W](f) = V(W(f)) - W(V(f)) \]
\[ = \left( x \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + x(y+1) \frac{\partial}{\partial z} \right) \left( \frac{\partial f}{\partial x} + y \frac{\partial f}{\partial z} \right) \]
\[ - \left( \frac{\partial}{\partial x} + y \frac{\partial}{\partial z} \right) \left( x \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + x(y+1) \frac{\partial f}{\partial z} \right) \]
\[ = \frac{\partial f}{\partial z} - \frac{\partial f}{\partial x} - (y+1) \frac{\partial f}{\partial z} \]
\[ = - \frac{\partial f}{\partial x} - y \frac{\partial f}{\partial z} = -W(f). \]
Thus $D$ is involutive and Theorem 25.3 is applicable. Our goal is to find a flat chart around the origin. In order to construct this chart, we note that $\frac{\partial}{\partial z}$ is not in the span of $V$ and $W$ since if $\frac{\partial}{\partial z} = aV + bW$, then
\[
\frac{\partial}{\partial z} = (ax + b) \frac{\partial}{\partial x} + a \frac{\partial}{\partial y} + (a(x + 1) + by) \frac{\partial}{\partial z},
\]
which in turn implies $a = 0 = b$, a contradiction. This means we may rewrite a basis for $D$ in terms of Line (*) and find that
\[
X_1 := X = W = \frac{\partial}{\partial x} + y \frac{\partial}{\partial z},
\]
\[
X_2 := Y = V - xW = \frac{\partial}{\partial y} + x \frac{\partial}{\partial z}.
\]
The flow of $X$ is
\[
\alpha_u(x, y, z) := \Phi^1_u(x, y, z) = (x + u, y, z + uy),
\]
while the flow of $Y$ is
\[
\beta_v(x, y, z) := \Phi^2_v(x, y, z) = (x, y + v, z + vx).
\]
For a fixed point on the $z$-axis near the origin, say $(0, 0, w)$, we define $\Phi: \mathbb{R}^3 \to \mathbb{R}^3$ as a composition of the flows, namely
\[
\Phi(u, v)(0, 0, w) = \alpha_u \circ \beta_v (0, 0, w) = \alpha_u (0, v, w) = (u, v, w + uv).
\]
In other words $\Phi(u, v)(0, 0, w)$ provides the parameterization of $\mathbb{R}^3$ given by
\[
x = u, \quad y = v, \quad z = w + uv,
\]
and thus the flat chart is given by
\[
\Phi^{-1}(x, y, z) = (u, v, z - xy).
\]
By the paragraph immediately preceding Theorem 25.3, we conclude that the $N$, the integral manifolds of $D$, are given by the level sets of $w(x, y, z) = z - xy$.

In preparation for a global version of Frobenius Theorem in terms of foliations, we state the following Proposition proved in Lee [117] (Chapter 19, Proposition 19.12):

**Proposition 25.4.** Let $M$ be a smooth manifold of dimension $n$ and let $D$ be an involutive $r$-dimensional distribution on $M$. For every flat chart $(U, \varphi)$ for $D$, for every integral manifold $N$ of $D$, the set $N \cap U$ is a countable disjoint union of open parallel $k$-dimensional slices of $U$, each of which is open in $N$ and embedded in $M$.

We now describe an alternative method for describing involutivity in terms of differential forms.
25.3 Differential Ideals and Frobenius Theorem

First, we give a smoothness criterion for distributions in terms of one-forms.

**Proposition 25.5.** Let \( M \) be a smooth manifold of dimension \( n \) and let \( D \) be an assignment \( p \mapsto D_p \subseteq T_p M \) of some \( r \)-dimensional subspace of \( T_p M \), for all \( p \in M \). Then, \( D \) is a smooth distribution iff for every \( p \in U \), there is some open subset \( U \) with \( p \in U \), and some linearly independent one-forms \( \omega_1, \ldots, \omega_{n-r} \) defined on \( U \), so that

\[
D_q = \{ u \in T_q M \mid (\omega_1)_q(u) = \cdots = (\omega_{n-r})_q(u) = 0 \}, \quad \text{for all } q \in U.
\]

**Proof.** Proposition 25.5 is proved in Lee [117] (Chapter 19, Lemma 19.5). The idea is to either extend a set of linearly independent differential one-forms to a coframe and then consider the dual frame, or to extend some linearly independent vector fields to a frame and then take the dual basis. \( \square \)

Proposition 25.5 suggests the following definitions:

**Definition 25.4.** Let \( M \) be a smooth manifold of dimension \( n \) and let \( D \) be an \( r \)-dimensional distribution on \( M \).

1. Some linearly independent one-forms \( \omega_1, \ldots, \omega_{n-r} \) defined on some open subset \( U \subseteq M \) are called **local defining one-forms** for \( D \) if

\[
D_q = \{ u \in T_q M \mid (\omega_1)_q(u) = \cdots = (\omega_{n-r})_q(u) = 0 \}, \quad \text{for all } q \in U.
\]

2. We say that a \( k \)-form \( \omega \in \Lambda^k(M) \) **annihilates** \( D \) iff

\[
\omega_q(X_1(q), \ldots, X_r(q)) = 0,
\]

for all \( q \in M \) and for all vector fields \( X_1, \ldots, X_r \) belonging to \( D \). We write

\[
\mathcal{I}^k(D) = \{ \omega \in \Lambda^k(M) \mid \omega_q(X_1(q), \ldots, X_r(q)) = 0 \};
\]

for all \( q \in M \) and for all vector fields \( X_1, \ldots, X_r \) belonging to \( D \), and we let

\[
\mathcal{I}(D) = \bigoplus_{k=1}^n \mathcal{I}^k(D).
\]

Thus, \( \mathcal{I}(D) \) is the collection of differential forms that “vanish on \( D \).” In the classical terminology, a system of local defining one-forms as above is called a **system of Pfaffian equations**.

It turns out that \( \mathcal{I}(D) \) is not only a vector space, but also an ideal of \( \Lambda^\bullet(M) \).

A subspace \( \mathcal{I} \) of \( \Lambda^\bullet(M) \) is an **ideal** iff for every \( \omega \in \mathcal{I} \), we have \( \theta \wedge \omega \in \mathcal{I} \) for every \( \theta \in \Lambda^\bullet(M) \).
Proposition 25.6. Let $M$ be a smooth $n$-dimensional manifold and $D$ be an $r$-dimensional distribution. If $\mathcal{J}(D)$ is the space of forms annihilating $D$, then the following hold:

(a) $\mathcal{J}(D)$ is an ideal in $\mathcal{A}^*(M)$.

(b) $\mathcal{J}(D)$ is locally generated by $n - r$ linearly independent one-forms, which means: For every $p \in U$, there is some open subset $U \subseteq M$ with $p \in U$ and a set of linearly independent one-forms $\omega_1, \ldots, \omega_{n-r}$ defined on $U$, so that

(i) If $\omega \in \mathcal{A}^k(D)$, then $\omega \restriction U$ belongs to the ideal in $\mathcal{A}^*(U)$ generated by $\omega_1, \ldots, \omega_{n-r}$; that is,

$$\omega = \sum_{i=1}^{n-r} \theta_i \wedge \omega_i, \quad \text{on } U,$$

for some $(k-1)$-forms $\theta_i \in \mathcal{A}^{k-1}(U)$.

(ii) If $\omega \in \mathcal{A}^k(M)$ and if there is an open cover by subsets $U$ (as above) such that for every $U$ in the cover, $\omega \restriction U$ belongs to the ideal generated by $\omega_1, \ldots, \omega_{n-r}$, then $\omega \in \mathcal{J}(D)$.

(c) If $\mathcal{I} \subseteq \mathcal{A}^*(M)$ is an ideal locally generated by $n - r$ linearly independent one-forms, then there exists a unique smooth $r$-dimensional distribution $D$ for which $\mathcal{J} = \mathcal{J}(D)$.

Proof. Proposition 25.6 is proved in Warner (Chapter 2, Proposition 2.28); see also Morita [133] (Chapter 2, Lemma 2.19), and Lee [117] (Chapter 19, page 498-500).

In order to characterize involutive distributions, we need the notion of a differential ideal.

Definition 25.5. Let $M$ be a smooth manifold of dimension $n$. An ideal $\mathcal{I} \subseteq \mathcal{A}^*(M)$ is a differential ideal iff it is closed under exterior differentiation; that is,

$$d\omega \in \mathcal{I} \quad \text{whenever} \quad \omega \in \mathcal{I},$$

which we also express by $d\mathcal{I} \subseteq \mathcal{I}$.

Here is the differential ideal criterion for involutivity.

Proposition 25.7. Let $M$ be a smooth manifold of dimension $n$. A smooth $r$-dimensional distribution $D$ is involutive iff the ideal $\mathcal{J}(D)$ is a differential ideal.

Proof. Proposition 25.7 is proved in Warner [175] (Chapter 2, Proposition 2.30), Morita [133] (Chapter 2, Proposition 2.20), and Lee [117] (Chapter 19, Proposition 19.19).

Assume $D$ is involutive. Let $\omega \in \mathcal{A}^k(M)$ be any $k$ form on $M$ and let $X_0, \ldots, X_k$ be $k + 1$ smooth vector fields lying in $D$. Then, by Proposition 23.13 and the fact that $D$ is involutive, we deduce that $d\omega(X_0, \ldots, X_k) = 0$. Hence, $d\omega \in \mathcal{J}(D)$, which means that $\mathcal{J}(D)$ is a differential ideal.
For the converse, assume $\mathcal{I}(D)$ is a differential ideal. We know that for any one-form $\omega$,
\[ d\omega(X, Y) = X(\omega(Y)) - Y(\omega(X)) - \omega([X, Y]), \]
for any vector fields $X, Y$. Now, if $\omega_1, \ldots, \omega_{n-r}$ are linearly independent one-forms that define $D$ locally on $U$, using a bump function, we can extend $\omega_1, \ldots, \omega_{n-r}$ to $M$, and then using the above equation, for any vector fields $X, Y$ belonging to $D$, we get
\[ \omega_i([X, Y]) = X(\omega_i(Y)) - Y(\omega_i(X)) - d\omega_i(X, Y), \]
and since $\omega_i(X) = \omega_i(Y) = d\omega_i(X, Y) = 0$ because $\mathcal{I}(D)$ is a differential ideal and $\omega_i \in \mathcal{I}(D)$, we get $\omega_i([X, Y]) = 0$ for $i = 1, \ldots, n-r$, which means that $[X, Y]$ belongs to $D$.

Using Proposition 25.6, we can give a more concrete criterion: $D$ is involutive iff for every local defining one-forms $\omega_1, \ldots, \omega_{n-r}$ for $D$ (on some open subset, $U$), there are some one-forms $\omega_{ij} \in \mathcal{A}^1(U)$ so that
\[ d\omega_i = \sum_{j=1}^{n-r} \omega_{ij} \wedge \omega_j \quad (i = 1, \ldots, n-r). \]
The above conditions are often called the integrability conditions.

**Definition 25.6.** Let $M$ be a smooth manifold of dimension $n$. Given any ideal $\mathcal{I} \subseteq \mathcal{A}^\bullet(M)$, an immersed manifold $(M, \psi)$ of $M$ is an integral manifold of $\mathcal{I}$ iff $\psi^*\omega = 0$, for all $\omega \in \mathcal{I}$.

A connected integral manifold of the ideal $\mathcal{I}$ is maximal iff its image is not a proper subset of the image of any other connected integral manifold of $\mathcal{I}$.

Finally, here is the differential form version of the Frobenius Theorem.

**Theorem 25.8.** (Frobenius Theorem, Differential Ideal Version) Let $M$ be a smooth manifold of dimension $n$. If $\mathcal{I} \subseteq \mathcal{A}^\bullet(M)$ is a differential ideal locally generated by $n - r$ linearly independent one-forms, then for every $p \in M$, there exists a unique maximal, connected, integral manifold of $\mathcal{I}$ through $p$, and this integral manifold has dimension $r$.

**Proof.** Theorem 25.8 is proved in Warner [175]. This theorem follows immediately from Theorem 1.64 in Warner [175].

Another version of the Frobenius Theorem goes as follows; see Morita [133] (Chapter 2, Theorem 2.21).
**Theorem 25.9.** (Frobenius Theorem, Integrability Conditions Version) Let $M$ be a smooth manifold of dimension $n$. An $r$-dimensional distribution $D$ on $M$ is completely integrable iff for every local defining one-forms $\omega_1, \ldots, \omega_{n-r}$ for $D$ (on some open subset, $U$), there are some one-forms $\omega_{ij} \in \mathcal{A}^1(U)$ so that we have the integrability conditions

$$d\omega_i = \sum_{j=1}^{n-r} \omega_{ij} \wedge \omega_j \quad (i = 1, \ldots, n-r).$$

There are applications of Frobenius Theorem (in its various forms) to systems of partial differential equations, but we will not deal with this subject. The reader is advised to consult Lee [117], Chapter 19, and the references there.

### 25.4 A Glimpse at Foliations and a Global Version of Frobenius Theorem

All the maximal integral manifolds of an $r$-dimensional involutive distribution on a manifold $M$ yield a decomposition of $M$ with some nice properties, those of a foliation.

**Definition 25.7.** Let $M$ be a smooth manifold of dimension $n$. A family $\mathcal{F} = \{\mathcal{F}_\alpha\}_\alpha$ of subsets of $M$ is a $k$-dimensional foliation iff it is a family of pairwise disjoint, connected, immersed $k$-dimensional submanifolds of $M$ called the leaves of the foliation, whose union is $M$, and such that for every $p \in M$, there is a chart $(U, \varphi)$ with $p \in U$ called a flat chart for the foliation, and the following property holds:

$$\varphi(U) \cong U' \times U'' \subseteq \mathbb{R}^r \times \mathbb{R}^{n-r},$$

where $U'$ and $U''$ are some connected open subsets, and for every leaf $\mathcal{F}_\alpha$ of the foliation, if $\mathcal{F}_\alpha \cap U \neq \emptyset$, then $\mathcal{F}_\alpha \cap U$ is a countable union of $k$-dimensional slices given by

$$x_{r+1} = c_{r+1}, \ldots, x_n = c_n,$$

for some constants $c_{r+1}, \ldots, c_n \in \mathbb{R}$.

The structure of a foliation can be very complicated. For instance, the leaves can be dense in $M$. For example, there are spirals on a torus that form the leaves of a foliation (see Lee [117], Example 19.9). Foliations are in one-to-one correspondence with involutive distributions.

**Proposition 25.10.** Let $M$ be a smooth manifold of dimension $n$. For any foliation $\mathcal{F}$ on $M$, the family of tangent spaces to the leaves of $\mathcal{F}$ forms an involutive distribution on $M$.

The converse to the above proposition may be viewed as a global version of Frobenius Theorem.
Theorem 25.11. Let $M$ be a smooth manifold of dimension $n$. For every $r$-dimensional smooth, involutive distribution $D$ on $M$, the family of all maximal, connected, integral manifolds of $D$ forms a foliation of $M$.

Proof. The proof of Theorem 25.11 can be found in Lee [117] (Theorem 19.21).
CHAPTER 25. DISTRIBUTIONS AND THE FROBENIUS THEOREM
Chapter 26

Spherical Harmonics and Linear Representations of Lie Groups

26.1 Hilbert Spaces and Hilbert Sums

The material in this chapter assumes that the reader has some familiarity with the concepts of a Hilbert space and a Hilbert basis. We present this section to review these important concepts. Many of the proofs are omitted and are found in traditional sources such as Bourbaki [27], Dixmier [57], Lang [112, 113], and Rudin [148]. The special case of separable Hilbert spaces is treated very nicely in Deitmar [48].

We begin our review by recalling the definition of a Hermitian space. To do this we need to define the notion of a Hermitian form.

Definition 26.1. Given two vector spaces $E$ and $F$ over $\mathbb{C}$, a function $f: E \to F$ is semilinear if

\[
\begin{align*}
    f(u + v) &= f(u) + f(v) \\
    f(\lambda u) &= \overline{\lambda} u,
\end{align*}
\]

for all $u, v \in E$ and $\lambda \in \mathbb{C}$.

Definition 26.2. Given a complex vector space $E$, a function $\varphi: E \times E \to \mathbb{C}$ is a sesquilinear form if it is linear in its first argument and semilinear in its second argument, which means that

\[
\begin{align*}
    \varphi(u_1 + u_2, v) &= \varphi(u_1, v) + \varphi(u_2, v) \\
    \varphi(u, v_1 + v_2) &= \varphi(u, v_1) + \varphi(u, v_2) \\
    \varphi(\lambda u, v) &= \overline{\lambda} \varphi(u, v) \\
    \varphi(u, \lambda v) &= \overline{\lambda} \varphi(u, v),
\end{align*}
\]
for all \( u, v, u_1, u_2, v_1, v_2 \in E \) and \( \lambda \in \mathbb{C} \). A function \( \varphi: E \times E \to \mathbb{C} \) is a **Hermitian form** if it is sesquilinear and if
\[
\varphi(u, v) = \overline{\varphi(v, u)},
\]
for all \( u, v \in E \).

**Definition 26.3.** Given a complex vector space \( E \), a Hermitian form \( \varphi: E \times E \to \mathbb{C} \) is **positive definite** if \( \varphi(u, u) > 0 \) for all \( u \neq 0 \). A pair \( \langle E, \varphi \rangle \) where \( E \) is a complex vector space and \( \varphi \) is a Hermitian form on \( E \) is called a **Hermitian (or unitary)** space if \( \varphi \) is positive definite.

Given a Hermitian space \( \langle E, \varphi \rangle \), we can readily show that the function \( \| \cdot \|: E \to \mathbb{R} \) defined such that \( \| u \| = \varphi(u, u) \), is a norm on \( E \). Thus, \( E \) is a normed vector space. If \( E \) is also complete, then it is a very interesting space.

Recall that completeness has to do with the convergence of Cauchy sequences. A normed vector space \( \langle E, \| \cdot \| \rangle \) is automatically a metric space under the metric \( d \) defined such that \( d(u, v) = \| v - u \| \) (for the definition of a norm and of a metric space see Section 3.1). This leads us to the following definition:

**Definition 26.4.** Given a metric space \( E \) with metric \( d \), a sequence \( (a_n)_{n \geq 1} \) of elements \( a_n \in E \) is a **Cauchy sequence** iff for every \( \epsilon > 0 \), there is some \( N \geq 1 \) such that
\[
d(a_m, a_n) < \epsilon \quad \text{for all} \quad m, n \geq N.
\]
We say that \( E \) is **complete** iff every Cauchy sequence converges to a limit (which is unique, since a metric space is Hausdorff).

Every finite dimensional vector space over \( \mathbb{R} \) or \( \mathbb{C} \) is complete. One can show by induction that given any basis \((e_1, \ldots, e_n)\) of \( E \), the linear map \( h: \mathbb{C}^n \to E \) defined such that
\[
h((z_1, \ldots, z_n)) = z_1 e_1 + \cdots + z_n e_n
\]
is a homeomorphism (using the sup-norm on \( \mathbb{C}^n \)). One can also use the fact that any two norms on a finite dimensional vector space over \( \mathbb{R} \) or \( \mathbb{C} \) are equivalent (see Lang [113], Dixmier [57], or Schwartz [155]).

However, if \( E \) has infinite dimension, it may not be complete. When a Hermitian space is complete, a number of the properties that hold for finite dimensional Hermitian spaces also hold for infinite dimensional spaces. For example, any closed subspace has an orthogonal complement, and in particular, a finite dimensional subspace has an orthogonal complement. Hermitian spaces that are also complete play an important role in analysis. Since they were first studied by Hilbert, they are called Hilbert spaces.
Definition 26.5. A (complex) Hermitian space \( \langle E, \varphi \rangle \) which is a complete normed vector space under the norm \( \| \| \) induced by \( \varphi \) is called a Hilbert space. A real Euclidean space \( \langle E, \varphi \rangle \) which is complete under the norm \( \| \| \) induced by \( \varphi \) is called a real Hilbert space. All the results in this section hold for complex Hilbert spaces as well as for real Hilbert spaces. We state all results for the complex case only, since they also apply to the real case, and since the proofs in the complex case need a little more care.

Example 26.1. The space \( l^2 \) of all countably infinite sequences \( x = (x_i)_{i \in \mathbb{N}} \) of complex numbers such that \( \sum_{i=0}^{\infty} |x_i|^2 < \infty \) is a Hilbert space. It will be shown later that the map \( \varphi: l^2 \times l^2 \to \mathbb{C} \) defined such that

\[
\varphi ((x_i)_{i \in \mathbb{N}}, (y_i)_{i \in \mathbb{N}}) = \sum_{i=0}^{\infty} x_i y_i
\]

is well defined, and that \( l^2 \) is a Hilbert space under \( \varphi \). In fact, we will prove a more general result (Proposition 26.3).

Example 26.2. The set \( C^\infty[a,b] \) of smooth functions \( f: [a,b] \to \mathbb{C} \) is a Hermitian space under the Hermitian form

\[
\langle f, g \rangle = \int_a^b f(x) \overline{g(x)} dx,
\]

but it is not a Hilbert space because it is not complete (see Section 24.7 for the definition of the integral of a complex-valued function). It is possible to construct its completion \( L^2([a,b]) \), which turns out to be the space of Lebesgue square-integrable functions on \([a,b]\).

One of the most important facts about finite-dimensional Hermitian (and Euclidean) spaces is that they have orthonormal bases. This implies that, up to isomorphism, every finite-dimensional Hermitian space is isomorphic to \( \mathbb{C}^n \) (for some \( n \in \mathbb{N} \)) and that the inner product is given by

\[
\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} x_i \overline{y_i}.
\]

Furthermore, every subspace \( W \) has an orthogonal complement \( W^\perp \), and the inner product induces a natural duality between \( E \) and \( E^* \), where \( E^* \) is the space of linear forms on \( E \).

When \( E \) is a Hilbert space, \( E \) may be infinite dimensional, often of uncountable dimension. Thus, we can’t expect that \( E \) always have an orthonormal basis. However, if we modify the notion of basis so that a “Hilbert basis” is an orthogonal family that is also dense in \( E \), i.e., every \( v \in E \) is the limit of a sequence of finite combinations of vectors from the Hilbert basis, then we can recover most of the “nice” properties of finite-dimensional Hermitian spaces. For instance, if \( (u_k)_{k \in K} \) is a Hilbert basis, for every \( v \in E \), we can define the Fourier coefficients \( c_k = \langle v, u_k \rangle / \|u_k\| \), and then, \( v \) is the “sum” of its Fourier series \( \sum_{k \in K} c_k u_k \). However, the cardinality of the index set \( K \) can be very large, and it is necessary to define
what it means for a family of vectors indexed by \( K \) to be summable. It turns out that every Hilbert space is isomorphic to a space of the form \( l^2(K) \), where \( l^2(K) \) is a generalization of the space of Example 26.1 (see Theorem 26.7, usually called the Riesz-Fischer theorem).

**Definition 26.6.** Given a Hilbert space \( E \), a family \((u_k)_{k \in K}\) of nonnull vectors is an orthogonal family iff the \( u_k \) are pairwise orthogonal, i.e., \( \langle u_i, u_j \rangle = 0 \) for all \( i \neq j \) \((i, j \in K)\), and an orthonormal family iff \( \langle u_i, u_j \rangle = \delta_{i,j} \), for all \( i, j \in K \). A total orthogonal family (or system) or Hilbert basis is an orthogonal family that is dense in \( E \). This means that for every \( v \in E \), for every \( \epsilon > 0 \), there is some finite subset \( I \subseteq K \) and some family \((\lambda_i)_{i \in I}\) of complex numbers, such that

\[
\| v - \sum_{i \in I} \lambda_i u_i \| < \epsilon.
\]

Given an orthogonal family \((u_k)_{k \in K}\), for every \( v \in E \), for every \( k \in K \), the scalar \( c_k = \langle v, u_k \rangle / \| u_k \|^2 \) is called the \( k \)-th Fourier coefficient of \( v \) over \((u_k)_{k \in K}\).

**Remark:** The terminology Hilbert basis is misleading, because a Hilbert basis \((u_k)_{k \in K}\) is not necessarily a basis in the algebraic sense. Indeed, in general, \((u_k)_{k \in K}\) does not span \( E \). Intuitively, it takes linear combinations of the \( u_k \)'s with infinitely many nonnull coefficients to span \( E \). Technically, this is achieved in terms of limits. In order to avoid the confusion between bases in the algebraic sense and Hilbert bases, some authors refer to algebraic bases as Hamel bases and to total orthogonal families (or Hilbert bases) as Schauder bases.

Given an orthogonal family \((u_k)_{k \in K}\), for any finite subset \( I \) of \( K \), we often call sums of the form \( \sum_{i \in I} \lambda_i u_i \) partial sums of Fourier series, and if these partial sums converge to a limit denoted as \( \sum_{k \in K} c_k u_k \), we call \( \sum_{k \in K} c_k u_k \) a Fourier series.

However, we have to make sense of such sums! Indeed, when \( K \) is unordered or uncountable, the notion of limit or sum has not been defined. This can be done as follows (for more details, see Dixmier [57]):

**Definition 26.7.** Given a normed vector space \( E \) (say, a Hilbert space), for any nonempty index set \( K \), we say that a family \((u_k)_{k \in K}\) of vectors in \( E \) is summable with sum \( v \in E \) iff for every \( \epsilon > 0 \), there is some finite subset \( I \) of \( K \), such that,

\[
\left\| v - \sum_{j \in J} u_j \right\| < \epsilon
\]

for every finite subset \( J \) with \( I \subseteq J \subseteq K \). We say that the family \((u_k)_{k \in K}\) is summable iff there is some \( v \in E \) such that \((u_k)_{k \in K}\) is summable with sum \( v \). A family \((u_k)_{k \in K}\) is a Cauchy family iff for every \( \epsilon > 0 \), there is a finite subset \( I \) of \( K \), such that,

\[
\left\| \sum_{j \in J} u_j \right\| < \epsilon
\]

for every finite subset \( J \) of \( K \) with \( I \cap J = \emptyset \),
26.1. HILBERT SPACES AND HILBERT SUMS

If \((u_k)_{k \in K}\) is summable with sum \(v\), we usually denote \(v\) as \(\sum_{k \in K} u_k\).

**Remark:** The notion of summability implies that the sum of a family \((u_k)_{k \in K}\) is independent of any order on \(K\). In this sense, it is a kind of “commutative summability”. More precisely, it is easy to show that for every bijection \(\varphi: K \to K\) (intuitively, a reordering of \(K\)), the family \((u_k)_{k \in K}\) is summable iff the family \((u_{\varphi(k)})_{k \in \varphi(K)}\) is summable, and if so, they have the same sum.

To state some important properties of Fourier coefficients the following technical proposition, whose proof is found in Bourbaki [27], will be needed:

**Proposition 26.1.** Let \(E\) be a complete normed vector space (say, a Hilbert space).

1. For any nonempty index set \(K\), a family \((u_k)_{k \in K}\) is summable iff it is a Cauchy family.

2. Given a family \((r_k)_{k \in K}\) of nonnegative reals \(r_k \geq 0\) such that \(\sum_{i \in I} r_i < B\) for every finite subset \(I\) of \(K\), then \((r_k)_{k \in K}\) is summable and \(\sum_{k \in K} r_k = r\), where \(r\) is least upper bound of the set of finite sums \(\sum_{i \in I} r_i\) (\(I \subseteq K\)).

The following proposition gives some of the main properties of Fourier coefficients. Among other things, at most countably many of the Fourier coefficient may be nonnull, and the partial sums of a Fourier series converge. Given an orthogonal family \((u_k)_{k \in K}\), we let \(U_k = \mathbb{C}u_k\).

**Proposition 26.2.** Let \(E\) be a Hilbert space, \((u_k)_{k \in K}\) an orthogonal family in \(E\), and \(V\) the closure of the subspace generated by \((u_k)_{k \in K}\). The following properties hold:

1. For every \(v \in E\), for every finite subset \(I \subseteq K\), we have
   \[\sum_{i \in I} |c_i|^2 \leq \|v\|^2,\]
   where the \(c_k\) are the Fourier coefficients of \(v\).

2. For every vector \(v \in E\), if \((c_k)_{k \in K}\) are the Fourier coefficients of \(v\), the following conditions are equivalent:
   
   (2a) \(v \in V\)
   
   (2b) The family \((c_k u_k)_{k \in K}\) is summable and \(v = \sum_{k \in K} c_k u_k\).
   
   (2c) The family \((|c_k|^2)_{k \in K}\) is summable and \(\|v\|^2 = \sum_{k \in K} |c_k|^2\);

3. The family \((|c_k|^2)_{k \in K}\) is summable, and we have the Bessel inequality:
   \[\sum_{k \in K} |c_k|^2 \leq \|v\|^2.\]

As a consequence, at most countably many of the \(c_k\) may be nonzero. The family \((c_k u_k)_{k \in K}\) forms a Cauchy family, and thus, the Fourier series \(\sum_{k \in K} c_k u_k\) converges in \(E\) to some vector \(u = \sum_{k \in K} c_k u_k\).
Proof. (1) Let
\[ u_I = \sum_{i \in I} c_i u_i \]
for any finite subset \( I \) of \( K \). We claim that \( v - u_I \) is orthogonal to \( u_i \) for every \( i \in I \). Indeed,
\[ \langle v - u_I, u_i \rangle = \left\langle v - \sum_{j \in I} c_j u_j, u_i \right\rangle \]
\[ = \langle v, u_i \rangle - \sum_{j \in I} c_j \langle u_j, u_i \rangle \]
\[ = \langle v, u_i \rangle - c_i \| u_i \|^2 \]
\[ = \langle v, u_i \rangle - \langle v, u_i \rangle = 0, \]
since \( \langle u_j, u_i \rangle = 0 \) for all \( i \neq j \) and \( c_i = \langle v, u_i \rangle / \| u_i \|^2 \). As a consequence, we have
\[ \| v \|^2 = \left\| v - \sum_{i \in I} c_i u_i \right\|^2 \]
\[ = \left\| v - \sum_{i \in I} c_i u_i \right\|^2 + \left\| \sum_{i \in I} c_i u_i \right\|^2 \]
\[ = \left\| v - \sum_{i \in I} c_i u_i \right\|^2 + \sum_{i \in I} | c_i |^2, \]
since the \( u_i \) are pairwise orthogonal, that is,
\[ \sum_{i \in I} | c_i |^2 \leq \| v \|^2, \]
Thus,
\[ \sum_{i \in I} | c_i |^2 \leq \| v \|^2, \]
as claimed.

(2) We prove the chain of implications \((a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a)\).

\((a) \Rightarrow (b)\): If \( v \in V \), since \( V \) is the closure of the subspace spanned by \((u_k)_{k \in K}\), for every \( \epsilon > 0 \), there is some finite subset \( I \) of \( K \) and some family \((\lambda_i)_{i \in I}\) of complex numbers, such that
\[ \left\| v - \sum_{i \in I} \lambda_i u_i \right\| < \epsilon. \]
Now, for every finite subset \( J \) of \( K \) such that \( I \subseteq J \), we have
\[ \left\| v - \sum_{i \in I} \lambda_i u_i \right\|^2 = \left\| v - \sum_{j \in J} c_j u_j + \sum_{j \in J} c_j u_j - \sum_{i \in I} \lambda_i u_i \right\|^2 \]
\[ = \left\| v - \sum_{j \in J} c_j u_j \right\|^2 + \left\| \sum_{j \in J} c_j u_j - \sum_{i \in I} \lambda_i u_i \right\|^2, \]
since $I \subseteq J$ and the $u_j$ (with $j \in J$) are orthogonal to $v - \sum_{j \in J} c_j u_j$ by the argument in (1), which shows that
\[
\left\| v - \sum_{j \in J} c_j u_j \right\| \leq \left\| v - \sum_{i \in I} \lambda_i u_i \right\| < \epsilon,
\]
and thus, that the family $(c_k u_k)_{k \in K}$ is summable with sum $v$, so that
\[
v = \sum_{k \in K} c_k u_k.
\]

(b) $\Rightarrow$ (c): If $v = \sum_{k \in K} c_k u_k$, then for every $\epsilon > 0$, there some finite subset $I$ of $K$, such that
\[
\left\| v - \sum_{j \in J} c_j u_j \right\| < \sqrt{\epsilon},
\]
for every finite subset $J$ of $K$ such that $I \subseteq J$, and since we proved in (1) that
\[
\|v\|^2 = \left\| v - \sum_{j \in J} c_j u_j \right\|^2 + \sum_{j \in J} |c_j|^2,
\]
we get
\[
\|v\|^2 - \sum_{j \in J} |c_j|^2 < \epsilon,
\]
which proves that $(|c_k|^2)_{k \in K}$ is summable with sum $\|v\|^2$.

(c) $\Rightarrow$ (a): Finally, if $(|c_k|^2)_{k \in K}$ is summable with sum $\|v\|^2$, for every $\epsilon > 0$, there is some finite subset $I$ of $K$ such that
\[
\|v\|^2 - \sum_{j \in J} |c_j|^2 < \epsilon^2,
\]
for every finite subset $J$ of $K$ such that $I \subseteq J$, and again, using the fact that
\[
\|v\|^2 = \left\| v - \sum_{j \in J} c_j u_j \right\|^2 + \sum_{j \in J} |c_j|^2,
\]
we get
\[
\left\| v - \sum_{j \in J} c_j u_j \right\| < \epsilon,
\]
which proves that $(c_k u_k)_{k \in K}$ is summable with sum $\sum_{k \in K} c_k u_k = v$, and $v \in V$.

(3) Since $\sum_{i \in I} |c_i|^2 \leq \|v\|^2$ for every finite subset $I$ of $K$, by Proposition 26.1, the family $(|c_k|^2)_{k \in K}$ is summable. The Bessel inequality
\[
\sum_{k \in K} |c_k|^2 \leq \|v\|^2
\]
is an obvious consequence of the inequality \( \sum_{i \in I} |c_i|^2 \leq \|v\|^2 \) (for every finite \( I \subseteq K \)). Now, for every natural number \( n \geq 1 \), if \( K_n \) is the subset of \( K \) consisting of all \( c_k \) such that \( |c_k| \geq 1/n \), the number of elements in \( K_n \) is at most
\[
\sum_{k \in K_n} |nc_k|^2 \leq n^2 \sum_{k \in K} |c_k|^2 \leq n^2 \|v\|^2,
\]
which is finite, and thus, at most a countable number of the \( c_k \) may be nonzero.

Since \((|c_k|^2)_{k \in K}\) is summable with sum \( c \), for every \( \epsilon > 0 \), there is some finite subset \( I \) of \( K \) such that
\[
\sum_{j \in J} |c_j|^2 < \epsilon^2
\]
for every finite subset \( J \) of \( K \) such that \( I \cap J = \emptyset \). Since
\[
\left\| \sum_{j \in J} c_j u_j \right\|^2 = \sum_{j \in J} |c_j|^2,
\]
we get
\[
\left\| \sum_{j \in J} c_j u_j \right\| < \epsilon.
\]
This proves that \((c_k u_k)_{k \in K}\) is a Cauchy family, which, by Proposition 26.1, implies that \((c_k u_k)_{k \in K}\) is summable, since \( E \) is complete. Thus, the Fourier series \( \sum_{k \in K} c_k u_k \) is summable, with its sum denoted \( u \in V \).

Proposition 26.2 suggests looking at the space of sequences \((z_k)_{k \in K}\) (where \( z_k \in \mathbb{C} \)) such that \((|z_k|^2)_{k \in K}\) is summable. Indeed, such spaces are Hilbert spaces, and it turns out that every Hilbert space is isomorphic to one of those. Such spaces are the infinite-dimensional version of the spaces \( \mathbb{C}^n \) under the usual Euclidean norm.

**Definition 26.8.** Given any nonempty index set \( K \), the space \( l^2(K) \) is the set of all sequences \((z_k)_{k \in K}\), where \( z_k \in \mathbb{C} \), such that \((|z_k|^2)_{k \in K}\) is summable, i.e., \( \sum_{k \in K} |z_k|^2 < \infty \).

**Remarks:**

1. When \( K \) is a finite set of cardinality \( n \), \( l^2(K) \) is isomorphic to \( \mathbb{C}^n \).

2. When \( K = \mathbb{N} \), the space \( l^2(\mathbb{N}) \) corresponds to the space \( l^2 \) of Example 2. In that example we claimed that \( l^2 \) was a Hermitian space, and in fact, a Hilbert space. We now state this fact for any index set \( K \). For a proof of Proposition 26.3 we refer the reader to Schwartz [155]).
Proposition 26.3. Given any nonempty index set \( K \), the space \( l^2(K) \) is a Hilbert space under the Hermitian product
\[
\langle (x_k)_{k \in K}, (y_k)_{k \in K} \rangle = \sum_{k \in K} x_k \overline{y}_k.
\]
The subspace consisting of sequences \( (z_k)_{k \in K} \) such that \( z_k = 0 \), except perhaps for finitely many \( k \), is a dense subspace of \( l^2(K) \).

We just need two more propositions before being able to prove that every Hilbert space is isomorphic to some \( l^2(K) \).

Proposition 26.4. Let \( E \) be a Hilbert space, and \( (u_k)_{k \in K} \) an orthogonal family in \( E \). The following properties hold:

1. For every family \( (\lambda_k)_{k \in K} \in l^2(K) \), the family \( (\lambda_k u_k)_{k \in K} \) is summable. Furthermore, \( v = \sum_{k \in K} \lambda_k u_k \) is the only vector such that \( c_k = \lambda_k \) for all \( k \in K \), where the \( c_k \) are the Fourier coefficients of \( v \).

2. For any two families \( (\lambda_k)_{k \in K} \in l^2(K) \) and \( (\mu_k)_{k \in K} \in l^2(K) \), if \( v = \sum_{k \in K} \lambda_k u_k \) and \( w = \sum_{k \in K} \mu_k u_k \), we have the following equation, also called Parseval identity:
\[
\langle v, w \rangle = \sum_{k \in K} \lambda_k \overline{\mu}_k.
\]

Proof. (1) The fact that \( (\lambda_k)_{k \in K} \in l^2(K) \) means that \( (|\lambda_k|^2)_{k \in K} \) is summable. The proof given in Proposition 26.2 (3) applies to the family \( (|\lambda_k|^2)_{k \in K} \) (instead of \( (|c_k|^2)_{k \in K} \)), and yields the fact that \( (\lambda_k u_k)_{k \in K} \) is summable. Letting \( v = \sum_{k \in K} \lambda_k u_k \), recall that \( c_k = \langle v, u_k \rangle / \|u_k\|^2 \). Pick some \( k \in K \). Since \( \langle -,- \rangle \) is continuous, for every \( \epsilon > 0 \), there is some \( \eta > 0 \) such that
\[
|\langle v, u_k \rangle - \langle w, u_k \rangle| < \epsilon \|u_k\|^2
\]
whenever
\[
\|v - w\| < \eta.
\]
However, since for every \( \eta > 0 \), there is some finite subset \( I \) of \( K \) such that
\[
\left\|v - \sum_{j \in J} \lambda_j u_j\right\| < \eta
\]
for every finite subset \( J \) of \( K \) such that \( I \subseteq J \), we can pick \( J = I \cup \{k\} \), and letting \( w = \sum_{j \in J} \lambda_j u_j \), we get
\[
\left|\langle v, u_k \rangle - \left\langle \sum_{j \in J} \lambda_j u_j, u_k \right\rangle\right| < \epsilon \|u_k\|^2.
\]
However, 
\[ \langle v, u_k \rangle = c_k \|u_k\|^2 \quad \text{and} \quad \left\langle \sum_{j \in J} \lambda_j u_j, u_k \right\rangle = \lambda_k \|u_k\|^2, \]
and thus, the above proves that \( |c_k - \lambda_k| < \epsilon \) for every \( \epsilon > 0 \), and thus, that \( c_k = \lambda_k \).

(2) Since \( \langle -,- \rangle \) is continuous, for every \( \epsilon > 0 \), there are some \( \eta_1 > 0 \) and \( \eta_2 > 0 \), such that
\[ |\langle x, y \rangle| < \epsilon \]
whenever \( \|x\| < \eta_1 \) and \( \|y\| < \eta_2 \). Since \( v = \sum_{k \in K} \lambda_k u_k \) and \( w = \sum_{k \in K} \mu_k u_k \), there is some finite subset \( I_1 \) of \( K \) such that
\[ \|v - \sum_{j \in J} \lambda_j u_j\| < \eta_1 \]
for every finite subset \( J \) of \( K \) such that \( I_1 \subseteq J \), and there is some finite subset \( I_2 \) of \( K \) such that
\[ \|w - \sum_{j \in J} \mu_j u_j\| < \eta_2 \]
for every finite subset \( J \) of \( K \) such that \( I_2 \subseteq J \). Letting \( I = I_1 \cup I_2 \), we get
\[ \left| \left\langle v - \sum_{i \in I} \lambda_i u_i, w - \sum_{i \in I} \mu_i u_i \right\rangle \right| < \epsilon. \]

Furthermore,
\[ \langle v, w \rangle = \left\langle \sum_{i \in I} \lambda_i u_i + \sum_{i \in I} \lambda_i u_i, \sum_{i \in I} \mu_i u_i + \sum_{i \in I} \mu_i u_i \right\rangle \]
\[ = \left\langle \sum_{i \in I} \lambda_i u_i, \sum_{i \in I} \mu_i u_i \right\rangle + \sum_{i \in I} \lambda_i \mu_i, \]
since the \( u_i \) are orthogonal to \( v - \sum_{i \in I} \lambda_i u_i \) and \( w - \sum_{i \in I} \mu_i u_i \) for all \( i \in I \). This proves that for every \( \epsilon > 0 \), there is some finite subset \( I \) of \( K \) such that
\[ \left| \langle v, w \rangle - \sum_{i \in I} \lambda_i \mu_i \right| < \epsilon. \]

We already know from Proposition 26.3 that \( (\lambda_k \overline{\mu_k})_{k \in K} \) is summable, and since \( \epsilon > 0 \) is arbitrary, we get
\[ \langle v, w \rangle = \sum_{k \in K} \lambda_k \overline{\mu_k}. \]
\[ \square \]
Proposition 26.5. Let $E$ be a Hilbert space, and let $(u_k)_{k \in K}$ be an orthogonal family in $E$. The following properties are equivalent:

(1) The family $(u_k)_{k \in K}$ is a total orthogonal family.

(2) For every vector $v \in E$, if $(c_k)_{k \in K}$ are the Fourier coefficients of $v$, then the family $(c_k u_k)_{k \in K}$ is summable and $v = \sum_{k \in K} c_k u_k$.

(3) For every vector $v \in E$, we have the Parseval identity:

$$\|v\|^2 = \sum_{k \in K} |c_k|^2.$$  

(4) For every vector $u \in E$, if $\langle u, u_k \rangle = 0$ for all $k \in K$, then $u = 0$.

Proof. The equivalence of (1), (2), and (3), is an immediate consequence of Proposition 26.2 and Proposition 26.4.

(4) If $(u_k)_{k \in K}$ is a total orthogonal family and $\langle u, u_k \rangle = 0$ for all $k \in K$, since $u = \sum_{k \in K} c_k u_k$ where $c_k = \langle u, u_k \rangle / \|u_k\|^2$, we have $c_k = 0$ for all $k \in K$, and $u = 0$.

Conversely, assume that the closure $V$ of $(u_k)_{k \in K}$ is different from $E$. Then we have $E = V \oplus V^\perp$, where $V^\perp$ is the orthogonal complement of $V$, and $V^\perp$ is nontrivial since $V \neq E$. As a consequence, there is some nonnull vector $u \in V^\perp$. But then, $u$ is orthogonal to every vector in $V$, and in particular,

$$\langle u, u_k \rangle = 0$$

for all $k \in K$, which, by assumption, implies that $u = 0$, contradicting the fact that $u \neq 0$. \qed

At last, we can prove that every Hilbert space is isomorphic to some Hilbert space $l^2(K)$ for some suitable $K$.

First, we need the fact that every Hilbert space has some Hilbert basis. This proof uses Zorn’s Lemma (see Rudin [148]).

Proposition 26.6. Let $E$ be a Hilbert space. Given any orthogonal family $(u_k)_{k \in K}$ in $E$, there is a total orthogonal family $(u_l)_{l \in L}$ containing $(u_k)_{k \in K}$.

All Hilbert bases for a Hilbert space $E$ have index sets $K$ of the same cardinality. For a proof, see Bourbaki [27].

Definition 26.9. A Hilbert space $E$ is separable if its Hilbert bases are countable.
Theorem 26.7. (Riesz-Fischer) For every Hilbert space $E$, there is some nonempty set $K$ such that $E$ is isomorphic to the Hilbert space $l^2(K)$. More specifically, for any Hilbert basis $(u_k)_{k \in K}$ of $E$, the maps $f : l^2(K) \to E$ and $g : E \to l^2(K)$ defined such that

$$f((\lambda_k)_{k \in K}) = \sum_{k \in K} \lambda_k u_k \quad \text{and} \quad g(u) = \left( \frac{\langle u, u_k \rangle}{\|u_k\|^2} \right)_{k \in K} = (c_k)_{k \in K},$$

are bijective linear isometries such that $g \circ f = \text{id}$ and $f \circ g = \text{id}$.

Proof. By Proposition 26.4 (1), the map $f$ is well defined, and it is clearly linear. By Proposition 26.2 (3), the map $g$ is well defined, and it is also clearly linear. By Proposition 26.2 (2b), we have

$$f(g(u)) = u = \sum_{k \in K} c_k u_k,$$

and by Proposition 26.4 (1), we have

$$g(f((\lambda_k)_{k \in K})) = (\lambda_k)_{k \in K},$$

and thus $g \circ f = \text{id}$ and $f \circ g = \text{id}$. By Proposition 26.4 (2), the linear map $g$ is an isometry. Therefore, $f$ is a linear bijection and an isometry between $l^2(K)$ and $E$, with inverse $g$. □

Remark: The surjectivity of the map $g : E \to l^2(K)$ is known as the Riesz-Fischer theorem.

Having done all this hard work, we sketch how these results apply to Fourier series. Again, we refer the readers to Rudin [148] or Lang [112, 113] for a comprehensive exposition.

Let $C(T)$ denote the set of all periodic continuous functions $f : [-\pi, \pi] \to \mathbb{C}$ with period $2\pi$. There is a Hilbert space $L^2(T)$ containing $C(T)$ and such that $C(T)$ is dense in $L^2(T)$, whose inner product is given by

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}dx.$$

The Hilbert space $L^2(T)$ is the space of Lebesgue square-integrable periodic functions (of period $2\pi$).

It turns out that the family $(e^{ikx})_{k \in \mathbb{Z}}$ is a total orthogonal family in $L^2(T)$, because it is already dense in $C(T)$ (for instance, see Rudin [148]). Then, the Riesz-Fischer theorem says that for every family $(c_k)_{k \in \mathbb{Z}}$ of complex numbers such that

$$\sum_{k \in \mathbb{Z}} |c_k|^2 < \infty,$$

there is a unique function $f \in L^2(T)$ such that $f$ is equal to its Fourier series

$$f(x) = \sum_{k \in \mathbb{Z}} c_k e^{ikx}.$$
where the Fourier coefficients \( c_k \) of \( f \) are given by the formula
\[
c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{-ikt}dt.
\]
The Parseval theorem says that
\[
\sum_{k=-\infty}^{+\infty} c_k \overline{d_k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)\overline{g(t)}dt
\]
for all \( f, g \in L^2(T) \), where \( c_k \) and \( d_k \) are the Fourier coefficients of \( f \) and \( g \).

Thus, there is an isomorphism between the two Hilbert spaces \( L^2(T) \) and \( l^2(\mathbb{Z}) \), which is the deep reason why the Fourier coefficients “work”. Theorem 26.7 implies that the Fourier series \( \sum_{k \in \mathbb{Z}} c_k e^{ikx} \) of a function \( f \in L^2(T) \) converges to \( f \) in the \( L^2 \)-sense, i.e., in the mean-square sense. This does not necessarily imply that the Fourier series converges to \( f \) pointwise! This is a subtle issue, and for more on this subject, the reader is referred to Lang [112, 113] or Schwartz [157, 158].

An alternative Hilbert basis for \( L^2(T) \) is given by \( \{\cos kx, \sin kx\}_{k=0}^{\infty} \). This particular Hilbert basis will play an important role representing the spherical harmonics on \( S^1 \) as seen the next section.

## 26.2 Spherical Harmonics on the Circle

For the remainder of this chapter we discuss spherical harmonics and take a glimpse at the linear representation of Lie groups. Spherical harmonics on the sphere \( S^2 \) have interesting applications in computer graphics and computer vision so this material is not only important for theoretical reasons but also for practical reasons.

Joseph Fourier (1768-1830) invented Fourier series in order to solve the heat equation [68]. Using Fourier series, every square-integrable periodic function \( f \) (of period \( 2\pi \)) can be expressed uniquely as the sum of a power series of the form
\[
f(\theta) = a_0 + \sum_{k=1}^{\infty} (a_k \cos k\theta + b_k \cos k\theta),
\]
where the Fourier coefficients \( a_k, b_k \) of \( f \) are given by the formulae
\[
a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) d\theta, \quad a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) \cos k\theta d\theta, \quad b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) \sin k\theta d\theta,
\]
for \( k \geq 1 \). The reader will find the above formulae in Fourier’s famous book [68] in Chapter III, Section 233, page 256, essentially using the notation that we use nowdays.
This remarkable discovery has many theoretical and practical applications in physics, signal processing, engineering, etc. We can describe Fourier series in a more conceptual manner if we introduce the following inner product on square-integrable functions:

\[ \langle f, g \rangle = \int_{-\pi}^{\pi} f(\theta)g(\theta) \, d\theta, \]

which we will also denote by

\[ \langle f, g \rangle = \int_{S^1} f(\theta)g(\theta) \, d\theta, \]

where \( S^1 \) denotes the unit circle. After all, periodic functions of (period 2\( \pi \)) can be viewed as functions on the circle. With this inner product, the space \( L^2(S^1) \) is a complete normed vector space, that is, a Hilbert space. Furthermore, if we define the subspaces \( \mathcal{H}_k(S^1) \) of \( L^2(S^1) \) so that \( \mathcal{H}_0(S^1) (= \mathbb{R}) \) is the set of constant functions and \( \mathcal{H}_k(S^1) \) is the two-dimensional space spanned by the functions \( \cos k\theta \) and \( \sin k\theta \), then it turns out that we have a Hilbert sum decomposition

\[ L^2(S^1) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^1) \]

into pairwise orthogonal subspaces, where \( \bigcup_{k=0}^{\infty} \mathcal{H}_k(S^1) \) is dense in \( L^2(S^1) \). The functions \( \cos k\theta \) and \( \sin k\theta \) are also orthogonal in \( \mathcal{H}_k(S^1) \).

Now, it turns out that the spaces \( \mathcal{H}_k(S^1) \) arise naturally when we look for homogeneous solutions of the Laplace equation \( \Delta f = 0 \) in \( \mathbb{R}^2 \) (Pierre-Simon Laplace, 1749-1827). Roughly speaking, a homogeneous function in \( \mathbb{R}^2 \) is a function that can be expressed in polar coordinates \((r, \theta)\) as

\[ f(r, \theta) = r^k g(\theta). \]

Recall that the Laplacian on \( \mathbb{R}^2 \) expressed in cartesian coordinates \((x, y)\) is given by

\[ \Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}, \]

where \( f: \mathbb{R}^2 \to \mathbb{R} \) is a function which is at least of class \( C^2 \). In polar coordinates \((r, \theta)\), where \((x, y) = (r \cos \theta, r \sin \theta)\) and \( r > 0 \), the Laplacian is given by

\[ \Delta f = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2}. \]

If we restrict \( f \) to the unit circle \( S^1 \), then the Laplacian on \( S^1 \) is given by

\[ \Delta_{S^1} f = \frac{\partial^2 f}{\partial \theta^2}. \]

It turns out that the space \( \mathcal{H}_k(S^1) \) is the eigenspace of \( \Delta_{S^1} \) for the eigenvalue \(-k^2\).
To show this, we consider another question, namely what are the harmonic functions on \( \mathbb{R}^2 \); that is, the functions \( f \) that are solutions of the Laplace equation

\[
\Delta f = 0.
\]

Our ancestors had the idea that the above equation can be solved by separation of variables. This means that we write \( f(r, \theta) = F(r)g(\theta) \), where \( F(r) \) and \( g(\theta) \) are independent functions. To make things easier, let us assume that \( F(r) = r^k \) for some integer \( k \geq 0 \), which means that we assume that \( f \) is a homogeneous function of degree \( k \). Recall that a function \( f: \mathbb{R}^2 \to \mathbb{R} \) is homogeneous of degree \( k \) iff

\[
f(tx, ty) = t^k f(x, y) \quad \text{for all } t > 0.
\]

Now, using the Laplacian in polar coordinates, we get

\[
\Delta f = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial (r^k g(\theta))}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 (r^k g(\theta))}{\partial \theta^2}
\]

\[
= \frac{1}{r} \frac{\partial}{\partial r} \left( kr^k g \right) + r^{k-2} \frac{\partial^2 g}{\partial \theta^2}
\]

\[
= r^{k-2} k^2 g + r^{k-2} \frac{\partial^2 g}{\partial \theta^2}
\]

\[
= r^{k-2} (k^2 g + \Delta_{S^1} g).
\]

Thus, we deduce that

\[
\Delta f = 0 \iff \Delta_{S^1} g = -k^2 g;
\]

that is, \( g \) is an eigenfunction of \( \Delta_{S^1} \) for the eigenvalue \( -k^2 \). But, the above equation is equivalent to the second-order differential equation

\[
\frac{d^2 g}{d \theta^2} + k^2 g = 0,
\]

whose general solution is given by

\[
g(\theta) = a_n \cos k\theta + b_n \sin k\theta.
\]

In summary, we found that the integers \( 0, -1, -4, -9, \ldots, -k^2, \ldots \) are eigenvalues of \( \Delta_{S^1} \), and that the functions \( \cos k\theta \) and \( \sin k\theta \) are eigenfunctions for the eigenvalue \( -k^2 \), with \( k \geq 0 \). So, it looks like the dimension of the eigenspace corresponding to the eigenvalue \( -k^2 \) is 1 when \( k = 0 \), and 2 when \( k \geq 1 \).

It can indeed be shown that \( \Delta_{S^1} \) has no other eigenvalues and that the dimensions claimed for the eigenspaces are correct. Observe that if we go back to our homogeneous harmonic functions \( f(r, \theta) = r^k g(\theta) \), we see that this space is spanned by the functions

\[
u_k = r^k \cos k\theta, \quad v_k = r^k \sin k\theta.
\]
Now, \((x + iy)^k = r^k (\cos k\theta + i \sin k\theta)\), and since \(\Re(x + iy)^k\) and \(\Im(x + iy)^k\) are homogeneous polynomials, we see that \(u_k\) and \(v_k\) are homogeneous polynomials called \textit{harmonic polynomials}. For example, here is a list of a basis for the harmonic polynomials (in two variables) of degree \(k = 0, 1, 2, 3, 4\):

\[
\begin{align*}
  k = 0 & \quad 1 \\
  k = 1 & \quad x, y \\
  k = 2 & \quad x^2 - y^2, xy \\
  k = 3 & \quad x^3 - 3xy^2, 3x^2y - y^3 \\
  k = 4 & \quad x^4 - 6x^2y^2 + y^4, x^3y - xy^3.
\end{align*}
\]

Therefore, the eigenfunctions of the Laplacian on \(S^1\) are the restrictions of the harmonic polynomials on \(\mathbb{R}^2\) to \(S^1\), and we have a Hilbert sum decomposition \(L^2(S^1) = \bigoplus_{k=0}^\infty \mathcal{H}_k(S^1)\). It turns out that this phenomenon generalizes to the sphere \(S^n \subseteq \mathbb{R}^{n+1}\) for all \(n \geq 1\).

Let us take a look at next case, \(n = 2\).

### 26.3 Spherical Harmonics on the 2-Sphere

The material of section is very classical and can be found in many places, for example Andrews, Askey and Roy [3] (Chapter 9), Sansone [152] (Chapter III), Hochstadt [93] (Chapter 6), and Lebedev [116] (Chapter ). We recommend the exposition in Lebedev [116] because we find it particularly clear and uncluttered. We have also borrowed heavily from some lecture notes by Hermann Gluck for a course he offered in 1997-1998.

Our goal is to find the homogeneous solutions of the Laplace equation \(\Delta f = 0\) in \(\mathbb{R}^3\), and to show that they correspond to spaces \(\mathcal{H}_k(S^2)\) of eigenfunctions of the Laplacian \(\Delta_{S^2}\) on the 2-sphere

\[
S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.
\]

Then, the spaces \(\mathcal{H}_k(S^2)\) will give us a Hilbert sum decomposition of the Hilbert space \(L^2(S^2)\) of square-integrable functions on \(S^2\). This is the generalization of Fourier series to the 2-sphere and the functions in the spaces \(\mathcal{H}_k(S^2)\) are called \textit{spherical harmonics}.

The Laplacian in \(\mathbb{R}^3\) is of course given by

\[
\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.
\]

If we use spherical coordinates

\[
\begin{align*}
  x &= r \sin \theta \cos \varphi \\
  y &= r \sin \theta \sin \varphi \\
  z &= r \cos \theta,
\end{align*}
\]

then the Laplacian in spherical coordinates is

\[
\Delta f = r^{-1} \sum_{\ell = 0}^\infty \sum_{m = -\ell}^\ell Y_{\ell m}(r, \theta, \varphi) \Delta_{S^2} Y_{\ell m}(r, \theta, \varphi).
\]
in $\mathbb{R}^3$, where $0 \leq \theta < \pi$, $0 \leq \varphi < 2\pi$ and $r > 0$ (recall that $\varphi$ is the so-called azimuthal angle in the $xy$-plane originating at the $x$-axis and $\theta$ is the so-called polar angle from the $z$-axis, angle defined in the plane obtained by rotating the $xz$-plane around the $z$-axis by the angle $\varphi$), then the Laplacian in spherical coordinates is given by

$$\Delta f = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^2} f,$$

where

$$\Delta_{S^2} f = \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2},$$

is the Laplacian on the sphere $S^2$ (for example, see Lebedev [116], Chapter 8, or Section 26.4 where we derive this formula). Let us look for homogeneous harmonic functions $f(r, \theta, \varphi) = r^k g(\theta, \varphi)$ on $\mathbb{R}^3$; that is, solutions of the Laplace equation

$$\Delta f = 0.$$

We get

$$\Delta f = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial (r^k g)}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^2} (r^k g) = \frac{1}{r^2} \frac{\partial}{\partial r} (k r^{k+1} g) + r^{k-2} \Delta_{S^2} g = r^{k-2} (k(k+1)g + \Delta_{S^2} g).$$

Therefore,

$$\Delta f = 0 \iff \Delta_{S^2} g = -k(k+1)g,$$

that is, $g$ is an eigenfunction of $\Delta_{S^2}$ for the eigenvalue $-k(k+1)$.

We can look for solutions of the above equation using the separation of variables method. If we let $g(\theta, \varphi) = \Theta(\theta)\Phi(\varphi)$, then we get the equation

$$\frac{\Phi}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \Theta}{\partial \theta} \right) + \frac{\Theta}{\sin^2 \theta} \frac{\partial^2 \Phi}{\partial \varphi^2} = -k(k+1) \Theta \Phi,$$

that is, dividing by $\Theta \Phi$ and multiplying by $\sin^2 \theta$,

$$\sin \theta \frac{\partial}{\Theta} \left( \sin \theta \frac{\partial \Theta}{\partial \theta} \right) + k(k+1) \sin^2 \theta = -\frac{1}{\Phi} \frac{\partial^2 \Phi}{\partial \varphi^2}.$$

Since $\Theta$ and $\Phi$ are independent functions, the above is possible only if both sides are equal to a constant, say $\mu$. This leads to two equations

$$\frac{\partial^2 \Phi}{\partial \varphi^2} + \mu \Phi = 0$$

$$\sin \theta \frac{\partial}{\Theta} \left( \sin \theta \frac{\partial \Theta}{\partial \theta} \right) + k(k+1) \sin^2 \theta - \mu = 0.$$
However, we want $\Phi$ to be periodic in $\varphi$ since we are considering functions on the sphere, so $\mu$ must be of the form $\mu = m^2$ for some non-negative integer $m$. Then, we know that the space of solutions of the equation

$$\frac{\partial^2 \Phi}{\partial \varphi^2} + m^2 \Phi = 0$$

is two-dimensional and is spanned by the two functions

$$\Phi(\varphi) = \cos m\varphi, \quad \Phi(\varphi) = \sin m\varphi.$$

We still have to solve the equation

$$\sin \theta \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \Theta}{\partial \theta} \right) + (k(k+1) \sin^2 \theta - m^2) \Theta = 0,$$

which is equivalent to

$$\sin^2 \theta \Theta'' + \sin \theta \cos \theta \Theta' + (k(k+1) \sin^2 \theta - m^2) \Theta = 0.$$

a variant of Legendre’s equation. For this, we use the change of variable $t = \cos \theta$, and we consider the function $u$ given by $u(\cos \theta) = \Theta(\theta)$ (recall that $0 \leq \theta < \pi$), so we get the second-order differential equation

$$(1 - t^2)u'' - 2tu' + \left(k(k+1) - \frac{m^2}{1 - t^2}\right) u = 0$$

sometimes called the general Legendre equation (Adrien-Marie Legendre, 1752-1833). The trick to solve this equation is to make the substitution

$$u(t) = (1 - t^2)^m v(t);$$

see Lebedev [116], Chapter 7, Section 7.12. Then, we get

$$(1 - t^2)v'' - 2(m + 1)tv' + (k(k+1) - m(m+1))v = 0.$$ 

When $m = 0$, we get the Legendre equation:

$$(1 - t^2)v'' - 2tv' + k(k+1)v = 0;$$

see Lebedev [116], Chapter 7, Section 7.3.

This equation has two fundamental solution $P_k(t)$ and $Q_k(t)$ called the Legendre functions of the first and second kinds. The $P_k(t)$ are actually polynomials and the $Q_k(t)$ are given by power series that diverge for $t = 1$, so we only keep the Legendre polynomials $P_k(t)$. The Legendre polynomials can be defined in various ways. One definition is in terms of Rodrigues’ formula:

$$P_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n}(t^2 - 1)^n;$$
see Lebedev [116], Chapter 4, Section 4.2. In this version of the Legendre polynomials they are normalized so that \( P_n(1) = 1 \). There is also the following recurrence relation:

\[
\begin{align*}
P_0 &= 1 \\
P_1 &= t \\
(n + 1)P_{n+1} &= (2n + 1)tP_n - nP_{n-1} \quad n \geq 1;
\end{align*}
\]

see Lebedev [116], Chapter 4, Section 4.3. For example, the first six Legendre polynomials are:

\[
\begin{align*}
P_0 &= 1 \\
&P_1 = t \\
&P_2 = \frac{1}{2}(3t^2 - 1) \\
&P_3 = \frac{1}{2}(5t^3 - 3t) \\
&P_4 = \frac{1}{8}(35t^4 - 30t^2 + 3) \\
&P_5 = \frac{1}{8}(63t^5 - 70t^3 + 15t).
\end{align*}
\]

Let us now return to our differential equation

\[
(1 - t^2)v'' - 2(m + 1)tv' + (k(k + 1) - m(m + 1))v = 0. \tag{*}
\]

Observe that if we differentiate with respect to \( t \), we get the equation

\[
(1 - t^2)v''' - 2(m + 2)tv'' + (k(k + 1) - (m + 1)(m + 2))v' = 0.
\]

This shows that if \( v \) is a solution of our equation (\( * \)) for given \( k \) and \( m \), then \( v' \) is a solution of the same equation for \( k \) and \( m + 1 \). Thus, if \( P_k(t) \) solves (\( * \)) for given \( k \) and \( m = 0 \), then \( P'_k(t) \) solves (\( * \)) for the same \( k \) and \( m = 1 \), \( P''_k(t) \) solves (\( * \)) for the same \( k \) and \( m = 2 \), and in general \( \frac{d^m}{dt^m}(P_k(t)) \) solves (\( * \)) for \( k \) and \( m \). Therefore, our original equation

\[
(1 - t^2)u'' - 2tu' + \left( k(k + 1) - \frac{m^2}{1 - t^2} \right) u = 0 \tag{\dagger}
\]

has the solution

\[
u(t) = (1 - t^2)^{\frac{m}{2}} \frac{d^m}{dt^m}(P_k(t)).
\]

The function \( u(t) \) is traditionally denoted \( P'^m_k(t) \) and called an associated Legendre function; see Lebedev [116], Chapter 7, Section 7.12. The index \( k \) is often called the band index. Obviously, \( P'^m_k(t) \equiv 0 \) if \( m > k \) and \( P'^0_k(t) = P_k(t) \), the Legendre polynomial of degree \( k \). An associated Legendre function is not a polynomial in general, and because of the factor \( (1 - t^2)^{\frac{m}{2}} \), it is only defined on the closed interval \([-1, 1]\).
Certain authors add the factor \((-1)^m\) in front of the expression for the associated Legendre function \(P^m_k(t)\), as in Lebedev [116], Chapter 7, Section 7.12, see also footnote 29 on page 193. This seems to be common practice in the quantum mechanics literature where it is called the Condon Shortley phase factor.

The associated Legendre functions satisfy various recurrence relations that allows us to compute them. For example, for fixed \(m \geq 0\), we have (see Lebedev [116], Chapter 7, Section 7.12) the recurrence
\[
(k - m + 1)P^m_{k+1}(t) = (2k + 1)tP^m_k(t) - (k + m)P^m_{k-1}(t), \quad k \geq 1,
\]
and for fixed \(k \geq 2\), we have
\[
P^{m+2}_k(t) = \frac{2(m+1)t}{(t^2 - 1)^{1/2}}P^{m+1}_k(t) + (k - m)(k + m + 1)P^m_k(t), \quad 0 \leq m \leq k - 2,
\]
which can also be used to compute \(P^m_k\) starting from
\[
P^0_k(t) = P_k(t),
\]
\[
P^1_k(t) = \frac{kt}{(t^2 - 1)^{1/2}}P_k(t) - \frac{k}{(t^2 - 1)^{1/2}}P_{k-1}(t).
\]

Observe that the recurrence relation for \(m\) fixed yields the following equation for \(k = m\) (as \(P^m_{m-1} = 0\)):
\[
P^m_{m+1}(t) = (2m + 1)tP^m_m(t).
\]
It it also easy to see that
\[
P^m_m(t) = \frac{(2m)!}{2^m m!} (1 - t^2)^{m/2}.
\]
Observe that
\[
\frac{(2m)!}{2^m m!} = (2m - 1)(2m - 3) \cdots 5 \cdot 3 \cdot 1,
\]
an expression that is sometimes denoted \((2m - 1)!!\) and called the double factorial.

Beware that some papers in computer graphics adopt the definition of associated Legendre functions with the scale factor \((-1)^m\) added, so this factor is present in these papers, for example Green [78].

The equation above allows us to “lift” \(P^m_m\) to the higher band \(m + 1\). The computer graphics community (see Green [78]) uses the following three rules to compute \(P^m_k(t)\) where \(0 \leq m \leq k\):

1. Compute
\[
P^m_m(t) = \frac{(2m)!}{2^m m!} (1 - t^2)^{m/2}.
\]
If \(m = k\), stop. Otherwise do step 2 once:
(2) Compute \( P_{m+1}^n(t) = (2m + 1)tP_m^n(t) \). If \( k = m + 1 \), stop. Otherwise, iterate step 3:

(3) Starting from \( i = m + 1 \), compute

\[
(i - m + 1)P_{i+1}^m(t) = (2i + 1)tP_i^m(t) - (i + m)P_{i-1}^m(t)
\]

until \( i + 1 = k \).

If we recall that equation (†) was obtained from the equation

\[
\sin^2 \theta \Theta'' + \sin \theta \cos \theta \Theta' + (k(k + 1) \sin^2 \theta - m^2)\Theta = 0
\]

using the substitution \( u(\cos \theta) = \Theta(\theta) \), we see that \( \Theta(\theta) = P_k^m(\cos \theta) \) is a solution of the above equation. Putting everything together, as \( f(r, \theta, \varphi) = r^k \Theta(\theta)\Phi(\varphi) \), we proved that the homogeneous functions

\[
f(r, \theta, \varphi) = r^k \cos m\varphi P_k^m(\cos \theta), \quad f(r, \theta, \varphi) = r^k \sin m\varphi P_k^m(\cos \theta)
\]

are solutions of the Laplacian \( \Delta \) in \( \mathbb{R}^3 \), and that the functions

\[
\cos m\varphi P_k^m(\cos \theta), \quad \sin m\varphi P_k^m(\cos \theta)
\]

are eigenfunctions of the Laplacian \( \Delta_{S^2} \) on the sphere for the eigenvalue \(-k(k + 1)\). For \( k \) fixed, as \( 0 \leq m \leq k \), we get \( 2k + 1 \) linearly independent functions.

The notation for the above functions varies quite a bit, essentially because of the choice of normalization factors used in various fields (such as physics, seismology, geodesy, spectral analysis, magnetics, quantum mechanics etc.). We will adopt the notation \( y_l^m \), where \( l \) is a nonnegative integer but \( m \) is allowed to be negative, with \(-l \leq m \leq l\). Thus, we set

\[
y_l^m(\theta, \varphi) = \begin{cases} N_l^0 P_l(\cos \theta) & \text{if } m = 0 \\ \sqrt{2}N_l^m \cos m\varphi P_l^m(\cos \theta) & \text{if } m > 0 \\ \sqrt{2}N_l^m \sin(-m\varphi) P_l^{-m}(\cos \theta) & \text{if } m < 0 \end{cases}
\]

for \( l = 0, 1, 2, \ldots \), and where the \( N_l^m \) are scaling factors. In physics and computer graphics, \( N_l^m \) is chosen to be

\[
N_l^m = \sqrt{(2l + 1)(l - |m|)! \over 4\pi(l + |m|)!}. 
\]

The functions \( y_l^m \) are called the real spherical harmonics of degree \( l \) and order \( m \). The index \( l \) is called the band index.
The functions, \( y^m_l \), have some very nice properties, but to explain these we need to recall the Hilbert space structure of the space \( L^2(S^2) \) of square-integrable functions on the sphere. Recall that we have an inner product on \( L^2(S^2) \) given by

\[
\langle f, g \rangle = \int_{S^2} f g \, \text{Vol}_{S^2} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \varphi) g(\theta, \varphi) \sin \theta d\theta d\varphi,
\]

where \( f, g \in L^2(S^2) \) and where \( \text{Vol}_{S^2} \) is the volume form on \( S^2 \) (induced by the metric on \( \mathbb{R}^3 \)). With this inner product, \( L^2(S^2) \) is a complete normed vector space using the norm \( \| f \| = \sqrt{\langle f, f \rangle} \) associated with this inner product; that is, \( L^2(S^2) \) is a Hilbert space. Now, it can be shown that the Laplacian \( \Delta_{S^2} \) on the sphere is a self-adjoint linear operator with respect to this inner product. As the functions \( y^m_{l_1} \) and \( y^m_{l_2} \) with \( l_1 \neq l_2 \) are eigenfunctions corresponding to distinct eigenvalues \( -(l_1(l_1 + 1)) \) and \( -(l_2(l_2 + 1)) \), they are orthogonal; that is,

\[
\langle y^m_{l_1}, y^m_{l_2} \rangle = 0, \quad \text{if } l_1 \neq l_2.
\]

It is also not hard to show that for a fixed \( l \),

\[
\langle y^m_{l_1}, y^m_{l_2} \rangle = \delta_{m_1, m_2};
\]

that is, the functions \( y^m_l \) with \( -l \leq m \leq l \) form an orthonormal system, and we denote by \( \mathcal{H}_l(S^2) \) the \((2l + 1)\)-dimensional space spanned by these functions.

It turns out that the functions \( y^m_l \) form a basis of the eigenspace \( E_l \) of \( \Delta_{S^2} \) associated with the eigenvalue \( -l(l+1) \), so that \( E_l = \mathcal{H}_l(S^2) \), and that \( \Delta_{S^2} \) has no other eigenvalues. More is true. It turns out that \( L^2(S^2) \) is the orthogonal Hilbert sum of the eigenspaces \( \mathcal{H}_l(S^2) \). This means that the \( \mathcal{H}_l(S^2) \) are

1. mutually orthogonal
2. closed, and
3. The space \( L^2(S^2) \) is the Hilbert sum \( \bigoplus_{l=0}^{\infty} \mathcal{H}_l(S^2) \), which means that for every function \( f \in L^2(S^2) \), there is a unique sequence of spherical harmonics \( f_j \in \mathcal{H}_l(S^2) \) so that

\[
f = \sum_{l=0}^{\infty} f_l;
\]

that is, the sequence \( \sum_{j=0}^{l} f_j \) converges to \( f \) (in the norm on \( L^2(S^2) \)). Observe that each \( f_l \) is a unique linear combination \( f_l = \sum_{m} a_{m_l} y^m_l \).

Therefore, (3) gives us a Fourier decomposition on the sphere generalizing the familiar Fourier decomposition on the circle. Furthermore, the Fourier coefficients \( a_{m_l} \) can be computed using the fact that the \( y^m_l \) form an orthonormal Hilbert basis:

\[
a_{m_l} = \langle f, y^m_l \rangle.
\]
We also have the corresponding homogeneous harmonic functions $H_l^m(r, \theta, \varphi)$ on $\mathbb{R}^3$ given by

$$H_l^m(r, \theta, \varphi) = r^l y_l^m(\theta, \varphi).$$

If one starts computing explicitly the $H_l^m$ for small values of $l$ and $m$, one finds that it is always possible to express these functions in terms of the cartesian coordinates $x, y, z$ as homogeneous polynomials! This remarkable fact holds in general: The eigenfunctions of the Laplacian $\Delta S^2$, and thus the spherical harmonics, are the restrictions of homogeneous harmonic polynomials in $\mathbb{R}^3$. Here is a list of bases of the homogeneous harmonic polynomials of degree $k$ in three variables up to $k = 4$ (thanks to Herman Gluck):

<table>
<thead>
<tr>
<th>$k$</th>
<th>Basis Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1$</td>
</tr>
<tr>
<td>1</td>
<td>$x, y, z$</td>
</tr>
<tr>
<td>2</td>
<td>$x^2 - y^2, x^2 - z^2, xy, xz, yz$</td>
</tr>
<tr>
<td>3</td>
<td>$x^3 - 3xy^2, 3x^2y - y^3, x^3 - 3xz^2, 3x^2z - z^3, y^3 - 3yz^2, 3y^2z - z^3, xyz$</td>
</tr>
<tr>
<td>4</td>
<td>$x^4 - 6x^2y^2 + y^4, x^4 - 6x^2z^2 + z^4, y^4 - 6y^2z^2 + z^4, x^3y - xy^3, x^3z - xz^3, y^3z - yz^3, 3x^2yz - yz^3, 3xy^2z - xz^3, 3xyz^2 - x^3y$</td>
</tr>
</tbody>
</table>

Subsequent sections will be devoted to a proof of the important facts stated earlier.

## 26.4 The Laplace-Beltrami Operator

In order to define rigorously the Laplacian on the sphere $S^m \subseteq \mathbb{R}^{n+1}$ and establish its relationship with the Laplacian on $\mathbb{R}^{n+1}$, we need the definition of the Laplacian on a Riemannian manifold $(M, g)$, the Laplace-Beltrami operator (Eugenio Beltrami, 1835-1900). A more general definition of the the Laplace-Beltrami operator as an operator on differential forms is given in Section 27.3. In this chapter we only need the definition of the Laplacian on functions.

Recall that a Riemannian metric $g$ on a manifold $M$ is a smooth family of inner products $g = (g_p)$, where $g_p$ is an inner product on the tangent space $T_pM$ for every $p \in M$. The inner product $g_p$ on $T_pM$ establishes a canonical duality between $T_pM$ and $T^*_pM$, namely, we have the isomorphism $\flat: T_pM \rightarrow T^*_pM$ defined such that for every $u \in T_pM$, the linear form $u^\flat \in T_p^*M$ is given by

$$u^\flat(v) = g_p(u, v), \quad v \in T_pM.$$  

The inverse isomorphism $\sharp: T_p^*M \rightarrow T_pM$ is defined such that for every $\omega \in T_p^*M$, the vector $\omega^\sharp$ is the unique vector in $T_pM$ so that

$$g_p(\omega^\sharp, v) = \omega(v), \quad v \in T_pM.$$
The isomorphisms $♭$ and $♯$ induce isomorphisms between vector fields $X \in \mathfrak{X}(M)$ and one-forms $\omega \in \mathcal{A}^1(M)$. In particular, for every smooth function $f \in C^\infty(M)$, the vector field corresponding to the one-form $df$ is the gradient $\text{grad} f$ of $f$. The gradient of $f$ is uniquely determined by the condition

$$g_p((\text{grad} f)_p, v) = df_p(v), \quad v \in T_pM, \, p \in M.$$  

**Definition 26.10.** Let $(M, g)$ be a Riemannian manifold. If $\nabla_X$ is the covariant derivative associated with the Levi-Civita connection induced by the metric $g$, then the divergence of a vector field $X \in \mathfrak{X}(M)$ is the function $\text{div} X : M \to \mathbb{R}$ defined so that

$$(\text{div} X)(p) = \text{tr}(Y(p) \mapsto (\nabla_Y X)_p);$$

namely, for every $p$, $(\text{div} X)(p)$ is the trace of the linear map $Y(p) \mapsto (\nabla_Y X)_p$. Then, the Laplace-Beltrami operator, for short, Laplacian, is the linear operator $\Delta : C^\infty(M) \to C^\infty(M)$ given by

$$\Delta f = \text{div} \, \text{grad} \, f.$$  

**Remark:** The definition just given differs from the definition given in Section 27.3 by a negative sign. We adopted this sign convention to conform with most of the literature on spherical harmonics (where the negative sign is omitted). A consequence of this choice is that the eigenvalues of the Laplacian are negative.

For more details on the Laplace-Beltrami operator, we refer the reader to Chapter 27 or to Gallot, Hulin and Lafontaine [73] (Chapter 4) or O’Neill [138] (Chapter 3), Postnikov [144] (Chapter 13), Helgason [87] (Chapter 2) or Warner [175] (Chapters 4 and 6).

All this being rather abstract, it is useful to know how $\text{grad} f$, $\text{div} X$, and $\Delta f$ are expressed in a chart. If $(U, \varphi)$ is a chart of $M$, with $p \in M$, and if as usual

$$\left(\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p\right)$$

denotes the basis of $T_pM$ induced by $\varphi$, the local expression of the metric $g$ at $p$ is given by the $n \times n$ matrix $(g_{ij})_p$, with

$$(g_{ij})_p = g_p\left(\left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p\right).$$

The matrix $(g_{ij})_p$ is symmetric, positive definite, and its inverse is denoted $(g^{ij})_p$. We also let $|g|_p = \text{det}(g_{ij})_p$. For simplicity of notation we often omit the subscript $p$. Then it can be shown that for every function $f \in C^\infty(M)$, in local coordinates given by the chart $(U, \varphi)$, we have

$$\text{grad} f = \sum_{ij} g^{ij} \frac{\partial f}{\partial x_j} \frac{\partial}{\partial x_i},$$
where as usual
\[
\frac{\partial f}{\partial x_j}(p) = \left( \frac{\partial}{\partial x_j} \right)_p f = \frac{\partial(f \circ \varphi^{-1})}{\partial u_j}(\varphi(p)),
\]
and \((u_1, \ldots, u_n)\) are the coordinate functions in \(\mathbb{R}^n\). There are formulae for \(\text{div } X\) and \(\Delta f\) involving the Christoffel symbols. Let
\[
X = \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i},
\]
be a vector field expressed over a chart \((U, \varphi)\). Recall that the Christoffel symbol \(\Gamma^k_{ij}\) is defined as
\[
\Gamma^k_{ij} = \frac{1}{2} \sum_{l=1}^{n} g^{kl} \left( \partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij} \right), \tag{*}
\]
where \(\partial_k g_{ij} = \frac{\partial}{\partial x_k} (g_{ij})\); see Section 11.3. Then
\[
\text{div } X = \sum_{i=1}^{n} \left[ \frac{\partial X_i}{\partial x_i} + \sum_{j=1}^{n} \Gamma^i_{ij} X_j \right],
\]
and
\[
\Delta f = \sum_{i,j} g^{ij} \left[ \frac{\partial^2 f}{\partial x_i \partial x_j} - \sum_{k=1}^{n} \Gamma^k_{ij} \frac{\partial f}{\partial x_k} \right],
\]
whenever \(f \in C^\infty(M)\); see Pages 86 and 87 of O'Neill [138].

We take a moment to use O'Neill formula to re-derive the expression for the Laplacian on \(\mathbb{R}^2\) in terms of polar coordinates \((r, \theta)\), where \(x = r \cos \theta\), and \(y = r \sin \theta\). Note that
\[
\frac{\partial}{\partial x_1} = \frac{\partial}{\partial r} = (\cos \theta, \sin \theta)
\]
\[
\frac{\partial}{\partial x_2} = \frac{\partial}{\partial \theta} = (-r \sin \theta, r \cos \theta),
\]
which in turn gives
\[
g_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \quad g^{ij} = \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix}.
\]

In Section 11.3 we found that the only nonzero Christoffel symbols were
\[
\Gamma^2_{12} = \Gamma^2_{21} = \frac{1}{r} \quad \Gamma^1_{22} = -r.
\]
Hence
\[
\Delta f = \sum_{i,j=1}^{2} g^{ij} \left[ \frac{\partial^2 f}{\partial x_i \partial x_j} - \sum_{k=1}^{2} \Gamma^k_{ij} \frac{\partial f}{\partial x_k} \right]
\]
\[
= g^{11} \left[ \frac{\partial^2 f}{\partial x_1^2} - \sum_{k=1}^{2} \Gamma^k_{11} \frac{\partial f}{\partial x_k} \right] + g^{22} \left[ \frac{\partial^2 f}{\partial x_2^2} - \sum_{k=1}^{2} \Gamma^k_{22} \frac{\partial f}{\partial x_k} \right]
\]
\[
= \frac{\partial^2 f}{\partial r^2} + \frac{1}{r^2} \left[ \frac{\partial^2 f}{\partial \theta^2} - \Gamma^1_{12} \frac{\partial f}{\partial r} \right]
\]
\[
= \frac{\partial^2 f}{\partial r^2} + \frac{1}{r^2} \left[ \frac{\partial^2 f}{\partial \theta^2} + \Gamma^1_{12} \frac{\partial f}{\partial r} \right]
\]
\[
= \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r} \frac{\partial}{\partial r} \left( \frac{r f}{r} \right).
\]

O’Neill’s formula may also be used to re-derive the expression for the Laplacian on \( \mathbb{R}^3 \) in terms of spherical coordinates \((r, \theta, \varphi)\) where
\[
\begin{align*}
x &= r \sin \theta \cos \varphi \\
y &= r \sin \theta \sin \varphi \\
z &= r \cos \theta.
\end{align*}
\]

We have
\[
\begin{align*}
\frac{\partial}{\partial x_1} &= \frac{\partial}{\partial r} = \sin \theta \cos \varphi \frac{\partial}{\partial x} + \sin \theta \sin \varphi \frac{\partial}{\partial y} + \cos \theta \frac{\partial}{\partial z} = \hat{r} \\
\frac{\partial}{\partial x_2} &= \frac{\partial}{\partial \theta} = r \left( \cos \theta \cos \varphi \frac{\partial}{\partial x} + \cos \theta \sin \varphi \frac{\partial}{\partial y} - \sin \theta \frac{\partial}{\partial z} \right) = r \hat{\theta} \\
\frac{\partial}{\partial x_3} &= \frac{\partial}{\partial \varphi} = r \left( - \sin \theta \sin \varphi \frac{\partial}{\partial x} + \sin \theta \cos \varphi \frac{\partial}{\partial y} \right) = r \hat{\varphi}.
\end{align*}
\]

Observe that \(\hat{r}, \hat{\theta}\) and \(\hat{\varphi}\) are pairwise orthogonal. Therefore, the matrix \((g_{ij})\) is given by
\[
(g_{ij}) = \begin{pmatrix}
1 & 0 & 0 \\
0 & r^2 & 0 \\
0 & 0 & r^2 \sin^2 \theta
\end{pmatrix}
\]
and \(|g| = r^4 \sin^2 \theta\). The inverse of \((g_{ij})\) is
\[
(g^{ij}) = \begin{pmatrix}
1 & 0 & 0 \\
0 & r^{-2} & 0 \\
0 & 0 & r^{-2} \sin^{-2} \theta
\end{pmatrix}.
\]
By using Line (*), it is not hard to show that $\Gamma^k_{ij} = 0$ except for:

$$
\begin{align*}
\Gamma^1_{22} &= \frac{1}{2} g^{11} \partial_1 g_{22} = \frac{1}{2} \frac{\partial}{\partial r} r^2 = -r \\
\Gamma^1_{33} &= \frac{1}{2} g^{11} \partial_1 g_{33} = -\frac{1}{2} \frac{\partial}{\partial r} r^2 \sin^2 \theta = -r \sin^2 \theta \\
\Gamma^2_{12} &= \Gamma^2_{21} = \frac{1}{2} g^{22} \partial_2 g_{12} = \frac{1}{2r^2} \frac{\partial}{\partial r} r^2 = \frac{1}{r} \\
\Gamma^2_{33} &= \frac{1}{2} g^{22} [-\partial_2 g_{33}] = -\frac{1}{2r^2} \frac{\partial}{\partial \theta} r^2 \sin^2 \theta = -\sin \theta \cos \theta \\
\Gamma^3_{13} &= \Gamma^3_{31} = \frac{1}{2} g^{33} \partial_3 g_{13} = \frac{1}{2r^2 \sin^2 \theta} \frac{\partial}{\partial r} r^2 \sin^2 \theta = \frac{1}{r} \\
\Gamma^3_{23} &= \Gamma^3_{32} = \frac{1}{2} g^{33} \partial_3 g_{23} = \frac{1}{2r^2 \sin^2 \theta} \frac{\partial}{\partial \theta} r^2 \sin^2 \theta = \cot \theta.
\end{align*}
$$

Then

$$
\Delta f = \sum_{i,j=1}^{3} g^{ij} \left[ \frac{\partial^2 f}{\partial x_i \partial x_j} - \sum_{k=1}^{3} \Gamma^k_{ij} \frac{\partial f}{\partial x_k} \right]
= g^{11} \left[ \frac{\partial^2 f}{\partial x_1^2} - \sum_{k=1}^{3} \Gamma^k_{11} \frac{\partial f}{\partial x_k} \right] + g^{22} \left[ \frac{\partial^2 f}{\partial x_2^2} - \sum_{k=1}^{3} \Gamma^k_{22} \frac{\partial f}{\partial x_k} \right] + g^{33} \left[ \frac{\partial^2 f}{\partial x_3^2} - \sum_{k=1}^{3} \Gamma^k_{33} \frac{\partial f}{\partial x_k} \right]
= \frac{\partial^2 f}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} - \Gamma^1_{22} \frac{\partial f}{\partial r} + \frac{1}{r^2 \sin^2 \theta} \left[ \frac{\partial^2 f}{\partial \varphi^2} + r \sin^2 \theta \frac{\partial f}{\partial r} + \sin \theta \cos \theta \frac{\partial f}{\partial \theta} \right]
= \frac{\partial^2 f}{\partial r^2} + \frac{2}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \left[ \frac{\partial^2 f}{\partial \theta^2} + \cos \theta \frac{\partial f}{\partial \theta} \right] + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2}
= \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2}
$$

O'Neill’s formulae for the divergence and the Laplacian can be tedious to calculate since they involve knowing the Christoffel symbols. Fortunately there are other formulas for the divergence and the Laplacian which only involve $(g_{ij})$ and $(g^{ij})$ and hence will be more convenient for our purposes: For every vector field $X \in \mathfrak{X}(M)$ expressed in local coordinates as

$$
X = \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i},
$$

we have

$$
\text{div } X = \frac{1}{\sqrt{|g|}} \sum_{i=1}^{n} \frac{\partial}{\partial x_i} \left( \sqrt{|g|} X_i \right),
$$
and for every function $f \in C^\infty(M)$, the Laplacian $\Delta f$ is given by

$$\Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i} \left( \sqrt{|g|} g^{ij} \frac{\partial f}{\partial x_j} \right).$$

(\text{**)\)

The above formula is proved in Proposition 27.8, assuming $M$ is orientable. A different derivation is given in Postnikov [144] (Chapter 13, Section 5).

One should check that for $M = \mathbb{R}^n$ with its standard coordinates, the Laplacian is given by the familiar formula

$$\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}.$$

By using Equation (\text{**}), we quickly rediscover the Laplacian in spherical coordinates, namely

$$\Delta f = \frac{1}{r^2 \sin \theta} \sum_{i=1}^3 \sum_{j=1}^3 \frac{\partial}{\partial x_i} \left( r^2 \sin \theta \theta^{ij} \frac{\partial f}{\partial x_j} \right)$$

$$= \frac{1}{r^2 \sin \theta} \left[ \frac{\partial}{\partial r} \left( r^2 \sin \theta \frac{\partial f}{\partial r} \right) + \frac{\partial}{\partial \theta} \left( r^2 \sin \theta r^{-2} \frac{\partial f}{\partial \theta} \right) + \frac{\partial}{\partial \varphi} \left( r^2 \sin \theta r^{-2} \frac{\partial f}{\partial \varphi} \right) \right]$$

$$= \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2}.$$

Since $(\theta, \varphi)$ are coordinates on the sphere $S^2$ via

$$x = \sin \theta \cos \varphi$$
$$y = \sin \theta \sin \varphi$$
$$z = \cos \theta,$$

we see that in these coordinates, the metric $(\tilde{g}_{ij})$ on $S^2$ is given by the matrix

$$(\tilde{g}_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2 \theta \end{pmatrix},$$

that $|\tilde{g}| = \sin^2 \theta$, and that the inverse of $(\tilde{g}_{ij})$ is

$$(\tilde{g}^{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & \sin^{-2} \theta \end{pmatrix}.$$

It follows immediately that

$$\Delta_{S^2} f = \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 f}{\partial \varphi^2}. $$
so we have verified that
\[ \Delta f = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^2} f. \]

Let us now generalize the above formula to the Laplacian \( \Delta \) on \( \mathbb{R}^{n+1} \), and the Laplacian \( \Delta_{S^n} \) on \( S^n \), where
\[ S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1 \}. \]
Following Morimoto [132] (Chapter 2, Section 2), let us use “polar coordinates.” The map from \( \mathbb{R}_+ \times S^n \) to \( \mathbb{R}^{n+1} - \{0\} \) given by
\[ (r, \sigma) \mapsto r\sigma \]
is clearly a diffeomorphism. Thus, for any system of coordinates \((u_1, \ldots, u_n)\) on \( S^n \), the tuple \((u_1, \ldots, u_n, r)\) is a system of coordinates on \( \mathbb{R}^{n+1} - \{0\} \) called polar coordinates. Let us establish the relationship between the Laplacian \( \Delta \), on \( \mathbb{R}^{n+1} - \{0\} \) in polar coordinates and the Laplacian \( \Delta_{S^n} \) on \( S^n \) in local coordinates \((u_1, \ldots, u_n)\).

**Proposition 26.8.** If \( \Delta \) is the Laplacian on \( \mathbb{R}^{n+1} - \{0\} \) in polar coordinates \((u_1, \ldots, u_n, r)\) and \( \Delta_{S^n} \) is the Laplacian on the sphere \( S^n \) in local coordinates \((u_1, \ldots, u_n)\), then
\[ \Delta f = \frac{1}{r^n} \frac{\partial}{\partial r} \left( r^n \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^n} f. \]

**Proof.** Let us compute the \((n + 1) \times (n + 1)\) matrix \( G = (g_{ij}) \) expressing the metric on \( \mathbb{R}^{n+1} \) in polar coordinates and the \( n \times n \) matrix \( \tilde{G} = (\tilde{g}_{ij}) \) expressing the metric on \( S^n \). Recall that if \( \sigma \in S^n \), then \( \sigma \cdot \sigma = 1 \), and so
\[ \frac{\partial \sigma}{\partial u_i} \cdot \sigma = 0, \]
as
\[ \frac{\partial \sigma}{\partial u_i} \cdot \sigma = \frac{1}{2} \frac{\partial (\sigma \cdot \sigma)}{\partial u_i} = 0. \]
If \( x = r\sigma \) with \( \sigma \in S^n \), we have
\[ \frac{\partial x}{\partial u_i} = r \frac{\partial \sigma}{\partial u_i}, \quad 1 \leq i \leq n, \]
and
\[ \frac{\partial x}{\partial r} = \sigma. \]
It follows that
\[ g_{ij} = \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial u_j} = r^2 \frac{\partial \sigma}{\partial u_i} \cdot \frac{\partial \sigma}{\partial u_j} = r^2 \tilde{g}_{ij} \]
\[ g_{in+1} = \frac{\partial x}{\partial u_i} \cdot \frac{\partial x}{\partial r} = r \frac{\partial \sigma}{\partial u_i} \cdot \sigma = 0 \]
\[ g_{n+1n+1} = \frac{\partial x}{\partial r} \cdot \frac{\partial x}{\partial r} = \sigma \cdot \sigma = 1. \]
Consequently, we get

\[ G = \begin{pmatrix} r^2 \tilde{G} & 0 \\ 0 & 1 \end{pmatrix}, \]

\[ |g| = r^{2n} |\tilde{g}|, \text{ and} \]

\[ G^{-1} = \begin{pmatrix} r^{-2} \tilde{G}^{-1} & 0 \\ 0 & 1 \end{pmatrix}. \]

Using the above equations and

\[ \Delta f = \frac{1}{\sqrt{|g|}} \sum_{i,j} \partial_{x_i} \left( \sqrt{|g|} g^{ij} \frac{\partial f}{\partial x_j} \right), \]

we get

\[
\Delta f = \frac{1}{r^n \sqrt{|g|}} \sum_{i,j=1}^{n} \partial_{x_i} \left( r^n \sqrt{|g|} \frac{1}{r^2} \tilde{g}^{ij} \frac{\partial f}{\partial x_j} \right) + \frac{1}{r^n \sqrt{|g|}} \partial_r \left( r^n \sqrt{|g|} \frac{\partial f}{\partial r} \right) \\
= \frac{1}{r^2 \sqrt{|g|}} \sum_{i,j=1}^{n} \partial_{x_i} \left( \sqrt{|g|} \tilde{g}^{ij} \frac{\partial f}{\partial x_j} \right) + \frac{1}{r^n} \partial_r \left( r^n \frac{\partial f}{\partial r} \right) \\
= \frac{1}{r^2} \Delta_{S^n} f + \frac{1}{r^n} \frac{\partial}{\partial r} \left( r^n \frac{\partial f}{\partial r} \right),
\]

as claimed. \( \square \)

It is also possible to express \( \Delta_{S^n} \) in terms of \( \Delta_{S^{n-1}} \). If \( e_{n+1} = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1} \), then we can view \( S^{n-1} \) as the intersection of \( S^n \) with the hyperplane \( x_{n+1} = 0 \); that is, as the set

\[ S^{n-1} = \{ \sigma \in S^n \mid \sigma \cdot e_{n+1} = 0 \}. \]

If \( (u_1, \ldots, u_{n-1}) \) are local coordinates on \( S^{n-1} \), then \( (u_1, \ldots, u_{n-1}, \theta) \) are local coordinates on \( S^n \), by setting

\[ \sigma = \sin \theta \tilde{\sigma} + \cos \theta e_{n+1}, \]

with \( \tilde{\sigma} \in S^{n-1} \) and \( 0 \leq \theta < \pi \). Note that \( \tilde{\sigma} \cdot \tilde{\sigma} = 1 \), which in turn implies

\[ \frac{\partial \tilde{\sigma}}{\partial u_i} = 0. \]

Furthermore, \( \tilde{\sigma} \cdot e_{n+1} = 0 \), and hence

\[ \frac{\partial \tilde{\sigma}}{\partial u_i} \cdot e_{n+1} = 0. \]

By using these local coordinate systems, we find the relationship between \( \Delta_{S^n} \) and \( \Delta_{S^{n-1}} \) as follows: First observe that

\[ \frac{\partial \sigma}{\partial u_i} = \sin \theta \frac{\partial \tilde{\sigma}}{\partial u_i} + 0 e_{n+1} \quad \frac{\partial \sigma}{\partial \theta} = \cos \theta \tilde{\sigma} - \sin \theta e_{n+1}. \]
26.4. THE LAPLACE-BELTRAMI OPERATOR

If \( \tilde{G} = (\tilde{g}_{ij}) \) represents the metric on \( S^n \) and \( \check{G} = (\check{g}_{ij}) \) is the restriction of this metric to \( S^{n-1} \) as defined above then for \( 1 \leq i, j \leq n-1 \), we have

\[
\tilde{g}_{ij} = \frac{\partial \sigma}{\partial u_i} \cdot \frac{\partial \sigma}{\partial u_j} = \sin^2 \theta \frac{\partial \check{\sigma}}{\partial u_i} \cdot \frac{\partial \check{\sigma}}{\partial u_j} = \sin^2 \theta \check{g}_{ij}
\]

\[
\tilde{g}_{i,n} = \frac{\partial \sigma}{\partial u_i} \cdot \frac{\partial \sigma}{\partial \theta} = \left( \sin \theta \frac{\partial \check{\sigma}}{\partial u_i} + 0 e_{n+1} \right) (\cos \theta \check{\sigma} - \sin \theta e_{n+1}) = 0
\]

\[
\tilde{g}_{n,n} = \frac{\partial \sigma}{\partial \theta} \cdot \frac{\partial \sigma}{\partial \theta} = (\cos \theta \check{\sigma} - \sin \theta e_{n+1}) \cdot (\cos \theta \check{\sigma} - \sin \theta e_{n+1}) = \cos^2 \theta + \sin^2 \theta = 1.
\]

These calculations imply that

\[
\tilde{G} = \begin{pmatrix} \sin^2 \theta & 0 \\ 0 & 1 \end{pmatrix},
\]

\[|\tilde{g}| = \sin^{2n-2} \theta |\check{g}|, \text{ and that}
\]

\[
\tilde{G}^{-1} = \begin{pmatrix} \sin^{-2} \theta & \tilde{G}^{-1} \\ 0 & 1 \end{pmatrix}.
\]

Hence

\[
\Delta_{S^n} f = \frac{1}{\sin^{n-1} \theta \sqrt{|\tilde{g}|}} \sum_{i,j=1}^{n-1} \frac{\partial}{\partial u_i} \left( \sin^{n-1} \theta \sqrt{|\check{g}|} \frac{1}{\sin^2 \theta} \check{g}^{ij} \frac{\partial f}{\partial u_j} \right)
\]

\[
+ \frac{1}{\sin^{n-1} \theta \sqrt{|\tilde{g}|}} \frac{\partial}{\partial \theta} \left( \sin^{n-1} \theta \sqrt{|\check{g}|} \frac{\partial f}{\partial \theta} \right)
\]

\[
= \frac{1}{\sin^{n-1} \theta} \frac{\partial}{\partial \theta} \left( \sin^{n-1} \theta \frac{\partial f}{\partial \theta} \right)
+ \frac{1}{\sin^2 \theta \sqrt{|\tilde{g}|}} \sum_{i,j=1}^{n-1} \frac{\partial}{\partial u_i} \left( \sqrt{|\tilde{g}|} \tilde{g}^{ij} \frac{\partial f}{\partial u_j} \right)
\]

\[
= \frac{1}{\sin^{n-1} \theta} \frac{\partial}{\partial \theta} \left( \sin^{n-1} \theta \frac{\partial f}{\partial \theta} \right)
+ \frac{1}{\sin^2 \theta} \Delta_{S^{n-1}} f.
\]

A fundamental property of the divergence is known as Green’s Formula. There are actually two Greens’ Formulae, but we will only need the version for an orientable manifold without boundary given in Proposition 27.11. Recall that Green’s Formula states that if \( M \) is a compact, orientable, Riemannian manifold without boundary, then, for every smooth vector field \( X \in \mathfrak{X}(M) \), we have

\[
\int_M (\text{div } X) \text{ Vol}_M = 0,
\]

where \( \text{Vol}_M \) is the volume form on \( M \) induced by the metric.

If \( M \) is a compact, orientable Riemannian manifold, then for any two smooth functions \( f, h \in C^\infty(M) \), we define \( \langle f, h \rangle_M \) by

\[
\langle f, h \rangle_M = \int_M f h \text{ Vol}_M.
\]
Then, it is not hard to show that $\langle -,- \rangle_M$ is an inner product on $C^\infty(M)$.

An important property of the Laplacian on a compact, orientable Riemannian manifold is that it is a self-adjoint operator. This fact is proved in the more general case of an inner product on differential forms in Proposition 27.6, but it is instructive to give another proof in the special case of functions using Green’s Formula.

For this, we prove the following properties: For any two functions $f, h \in C^\infty(M)$, and any vector field $X \in \mathfrak{X}(M)$, we have:

\[
\text{div}(fX) = f \text{div} X + X(f) = f \text{div} X + g(\text{grad} f, X) \\
\text{grad} f (h) = g(\text{grad} f, \text{grad} h) = \text{grad} h (f).
\]

Using the above identities, we obtain the following important result.

**Proposition 26.9.** Let $M$ be a compact, orientable, Riemannian manifold without boundary. The Laplacian on $M$ is self-adjoint; that is, for any two functions $f, h \in C^\infty(M)$, we have

\[
\langle \Delta f, h \rangle_M = \langle f, \Delta h \rangle_M,
\]

or equivalently

\[
\int_M f \Delta h \, \text{Vol}_M = \int_M h \Delta f \, \text{Vol}_M.
\]

**Proof.** By the two identities before Proposition 26.9,

\[
f \Delta h = f \text{div} \text{grad} h = \text{div}(f \text{grad} h) - g(\text{grad} f, \text{grad} h)
\]

and

\[
h \Delta f = h \text{div} \text{grad} f = \text{div}(h \text{grad} f) - g(\text{grad} h, \text{grad} f),
\]

so we get

\[
f \Delta h - h \Delta f = \text{div}(f \text{grad} h - h \text{grad} f).
\]

By Green’s Formula,

\[
\int_M (f \Delta h - h \Delta f) \text{Vol}_M = \int_M \text{div}(f \text{grad} h - h \text{grad} f) \text{Vol}_M = 0,
\]

which proves that $\Delta$ is self-adjoint. \qed

The importance of Proposition 26.9 lies in the fact that as $\langle -,- \rangle_M$ is an inner product on $C^\infty(M)$, the eigenspaces of $\Delta$ for distinct eigenvalues are pairwise orthogonal. We will make heavy use of this property in the next section on harmonic polynomials.
26.5 Harmonic Polynomials, Spherical Harmonics and $L^2(S^n)$

Harmonic homogeneous polynomials and their restrictions to $S^n$, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} | x_1^2 + \cdots + x_{n+1}^2 = 1\},$$

turn out to play a crucial role in understanding the structure of the eigenspaces of the Laplacian on $S^n$ (with $n \geq 1$). The results in this section appear in one form or another in Stein and Weiss [165] (Chapter 4), Morimoto [132] (Chapter 2), Helgason [87] (Introduction, Section 3), Dieudonné [51] (Chapter 7), Axler, Bourdon and Ramey [15] (Chapter 5), and Vilenkin [173] (Chapter IX). Some of these sources assume a fair amount of mathematical background, and consequently uninitiated readers will probably find the exposition rather condensed, especially Helgason. We tried hard to make our presentation more “user-friendly.”

Recall that a homogeneous polynomial $P$ of degree $k$ in $n$ variables $x_1, \ldots, x_n$ is an expression of the form

$$P = \sum_{\alpha_1 + \cdots + \alpha_n = k} a_{\alpha_1, \ldots, \alpha_n} x_1^{\alpha_1} \cdots x_n^{\alpha_n},$$

where the coefficients $a_{\alpha_1, \ldots, \alpha_n}$ are either real or complex numbers. We view such a homogeneous polynomial as a function $P: \mathbb{R}^n \rightarrow \mathbb{C}$, or as a function $P: \mathbb{R}^n \rightarrow \mathbb{R}$ when the coefficients are all real. The Laplacian $\Delta P$ of $P$ is defined by

$$\Delta P = \sum_{\alpha_1 + \cdots + \alpha_n = k} a_{\alpha_1, \ldots, \alpha_n} \left( \frac{\partial^2}{\partial x_1^{\alpha_1}} + \cdots + \frac{\partial^2}{\partial x_n^{\alpha_n}} \right) (x_1^{\alpha_1} \cdots x_n^{\alpha_n}).$$

**Definition 26.11.** Let $\mathcal{P}_k(n+1)$ (resp. $\mathcal{P}_k^C(n+1)$) denote the space of homogeneous polynomials of degree $k$ in $n+1$ variables with real coefficients (resp. complex coefficients), and let $\mathcal{P}_k(S^n)$ (resp. $\mathcal{P}_k^C(S^n)$) denote the restrictions of homogeneous polynomials in $\mathcal{P}_k(n+1)$ to $S^n$ (resp. the restrictions of homogeneous polynomials in $\mathcal{P}_k^C(n+1)$ to $S^n$). Let $\mathcal{H}_k(n+1)$ (resp. $\mathcal{H}_k^C(n+1)$) denote the space of (real) harmonic polynomials (resp. complex harmonic polynomials), with

$$\mathcal{H}_k(n+1) = \{ P \in \mathcal{P}_k(n+1) | \Delta P = 0 \}$$

and

$$\mathcal{H}_k^C(n+1) = \{ P \in \mathcal{P}_k^C(n+1) | \Delta P = 0 \}.$$
A function \( f : \mathbb{R}^n \to \mathbb{R} \) (resp. \( f : \mathbb{R}^n \to \mathbb{C} \)) is \textit{homogeneous of degree} \( k \) iff
\[
f(tx) = t^k f(x), \quad \text{for all} \ x \in \mathbb{R}^n \ \text{and} \ t > 0.
\]

The restriction map \( \rho : \mathcal{H}_k(n+1) \to \mathcal{H}_k(S^n) \) is a surjective linear map. In fact, it is a bijection. Indeed, if \( P \in \mathcal{H}_k(n+1) \), observe that
\[
P(x) = \|x\|^k P \left( \frac{x}{\|x\|} \right), \quad \text{with} \quad \frac{x}{\|x\|} \in S^n,
\]
for all \( x \neq 0 \). Consequently, if \( P \upharpoonright S^n = Q \upharpoonright S^n \), that is \( P(\sigma) = Q(\sigma) \) for all \( \sigma \in S^n \), then
\[
P(x) = \|x\|^k P \left( \frac{x}{\|x\|} \right) = \|x\|^k Q \left( \frac{x}{\|x\|} \right) = Q(x)
\]
for all \( x \neq 0 \), which implies \( P = Q \) (as \( P \) and \( Q \) are polynomials). Therefore, we have a linear isomorphism between \( \mathcal{H}_k(n+1) \) and \( \mathcal{H}_k(S^n) \) (and between \( \mathcal{H}_k^C(n+1) \) and \( \mathcal{H}_k^C(S^n) \)).

It will be convenient to introduce some notation to deal with homogeneous polynomials. Given \( n \geq 1 \) variables \( x_1, \ldots, x_n \), and any \( n \)-tuple of nonnegative integers \( \alpha = (\alpha_1, \ldots, \alpha_n) \), let
\[
|\alpha| = \alpha_1 + \cdots + \alpha_n, \quad x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}, \quad \text{and let} \ \alpha! = \alpha_1! \cdots \alpha_n!.
\]
Then, every homogeneous polynomial \( P \) of degree \( k \) in the variables \( x_1, \ldots, x_n \) can be written uniquely as
\[
P = \sum_{|\alpha| = k} c_\alpha x^\alpha,
\]
with \( c_\alpha \in \mathbb{R} \) or \( c_\alpha \in \mathbb{C} \). It is well known that \( \mathcal{P}_k(n) \) is a (real) vector space of dimension
\[
d_k = \binom{n + k - 1}{k}
\]
and \( \mathcal{P}_k^C(n) \) is a complex vector space of the same dimension \( d_k \).

We can define an Hermitian inner product on \( \mathcal{P}_k^C(n) \) whose restriction to \( \mathcal{P}_k(n) \) is an inner product by viewing a homogeneous polynomial as a differential operator as follows: For every \( P = \sum_{|\alpha| = k} c_\alpha x^\alpha \in \mathcal{P}_k^C(n) \), let
\[
\partial(P) = \sum_{|\alpha| = k} c_\alpha \frac{\partial^k}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.
\]
Then, for any two polynomials \( P, Q \in \mathcal{P}_k^C(n) \), let
\[
\langle P, Q \rangle = \partial(P) \overline{Q}.
\]
A simple computation shows that
\[
\left\langle \sum_{|\alpha| = k} a_\alpha x^\alpha, \sum_{|\alpha| = k} b_\alpha x^\alpha \right\rangle = \sum_{|\alpha| = k} \alpha! a_\alpha \overline{b_\alpha}.
\]
Therefore, \( \langle P, Q \rangle \) is indeed an inner product. Also observe that
\[
\partial(x_1^2 + \cdots + x_n^2) = \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2} = \Delta.
\]

Another useful property of our inner product is this:
\[
\langle P, QR \rangle = \langle Q, \partial(R)P \rangle.
\]
Indeed.
\[
\langle P, QR \rangle = \langle QR, P \rangle = \partial(Q)\langle \partial(R)P \rangle = \partial(R)\langle Q, \partial(P) \rangle = \langle R, \partial(Q)P \rangle = \langle \partial(Q)P, R \rangle.
\]

In particular,
\[
\langle (x_1^2 + \cdots + x_n^2)P, Q \rangle = \langle P, \partial(x_1^2 + \cdots + x_n^2)Q \rangle = \langle P, \Delta Q \rangle.
\]

Let us write \( \|x\|^2 \) for \( x_1^2 + \cdots + x_n^2 \). Using our inner product, we can prove the following important theorem:

**Theorem 26.10.** The map \( \Delta: \mathcal{P}_k(n) \to \mathcal{P}_{k-2}(n) \) is surjective for all \( n, k \geq 2 \) (and similarly for \( \Delta: \mathcal{P}_k^C(n) \to \mathcal{P}_{k-2}^C(n) \)). Furthermore, we have the following orthogonal direct sum decompositions:

\[
\mathcal{P}_k(n) = \mathcal{H}_k(n) \oplus \|x\|^2 \mathcal{H}_{k-2}(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{k/2}(n)
\]

and

\[
\mathcal{P}_k^C(n) = \mathcal{H}_k^C(n) \oplus \|x\|^2 \mathcal{H}_{k-2}^C(n) \oplus \cdots \oplus \|x\|^{2j} \mathcal{H}_{k-2j}^C(n) \oplus \cdots \oplus \|x\|^{2[k/2]} \mathcal{H}_{k/2}^C(n),
\]

with the understanding that only the first term occurs on the right-hand side when \( k < 2 \).

**Proof.** If the map \( \Delta: \mathcal{P}_k^C(n) \to \mathcal{P}_{k-2}^C(n) \) is not surjective, then some nonzero polynomial \( Q \in \mathcal{P}_{k-2}^C(n) \) is orthogonal to the image of \( \Delta \). In particular, \( Q \) must be orthogonal to \( \Delta P \) with \( P = \|x\|^2 Q \in \mathcal{P}_k^C(n) \). So, using a fact established earlier,
\[
0 = \langle Q, \Delta P \rangle = \langle \|x\|^2 Q, P \rangle = \langle P, P \rangle,
\]
which implies that \( P = \|x\|^2 Q = 0 \), and thus \( Q = 0 \), a contradiction. The same proof is valid in the real case.
We claim that we have an orthogonal direct sum decomposition
\[ P^C_k(n) = H^C_k(n) \oplus \|x\|^2 P^C_{k-2}(n), \]
and similarly in the real case, with the understanding that the second term is missing if \( k < 2 \).

If \( k = 0, 1 \), then \( P^C_k(n) = H^C_k(n) \), so this case is trivial. Assume \( k \geq 2 \). Since \( \text{Ker} \Delta = H^C_k(n) \) and \( \Delta \) is surjective, \( \dim(P^C_k(n)) = \dim(H^C_k(n)) + \dim(P^C_{k-2}(n)) \), so it is sufficient to prove that \( H^C_k(n) \) is orthogonal to \( \|x\|^2 P^C_{k-2}(n) \). Now, if \( H \in H^C_k(n) \) and \( P = \|x\|^2 Q \in \|x\|^2 P^C_{k-2}(n) \), we have
\[ \langle \|x\|^2 Q, H \rangle = \langle Q, \Delta H \rangle = 0, \]
so \( H^C_k(n) \) and \( \|x\|^2 P^C_{k-2}(n) \) are indeed orthogonal. Using induction, we immediately get the orthogonal direct sum decomposition
\[ P^C_k(n) = H^C_k(n) \oplus \|x\|^2 H^C_{k-2}(n) \oplus \cdots \oplus \|x\|^{2j} H^C_{k-2j}(n) \oplus \cdots \oplus \|x\|^{2[k/2]} H^C_{k/2}(n) \]
and the corresponding real version.

**Remark:** Theorem 26.10 also holds for \( n = 1 \).

Theorem 26.10 has some important corollaries. Since every polynomial in \( n+1 \) variables is the sum of homogeneous polynomials, we get:

**Corollary 26.11.** The restriction to \( S^n \) of every polynomial (resp. complex polynomial) in \( n+1 \geq 2 \) variables is a sum of restrictions to \( S^n \) of harmonic polynomials (resp. complex harmonic polynomials).

We can also derive a formula for the dimension of \( H_k(n) \) (and \( H^C_k(n) \)).

**Corollary 26.12.** The dimension \( a_{k,n} \) of the space of harmonic polynomials \( H_k(n) \) is given by the formula
\[ a_{k,n} = \binom{n+k-1}{k} - \binom{n+k-3}{k-2} \]
if \( n, k \geq 2 \), with \( a_{0,n} = 1 \) and \( a_{1,n} = n \), and similarly for \( H^C_k(n) \). As \( H_k(n+1) \) is isomorphic to \( H_k(S^n) \) (and \( H^C_k(n+1) \) is isomorphic to \( H^C_k(S^n) \)) we have
\[ \dim(H^C_k(S^n)) = \dim(H_k(S^n)) = a_{k,n+1} = \binom{n+k}{k} - \binom{n+k-2}{k-2} \].

**Proof.** The cases \( k = 0 \) and \( k = 1 \) are trivial, since in this case \( H_k(n) = P_k(n) \). For \( k \geq 2 \), the result follows from the direct sum decomposition
\[ P_k(n) = H_k(n) \oplus \|x\|^2 P_{k-2}(n) \]
proved earlier. The proof is identical in the complex case.
Observe that when \( n = 2 \), we get \( a_{k,2} = 2 \) for \( k \geq 1 \), and when \( n = 3 \), we get \( a_{k,3} = 2k + 1 \) for all \( k \geq 0 \), which we already knew from Section 26.3. The formula even applies for \( n = 1 \) and yields \( a_{k,1} = 0 \) for \( k \geq 2 \).

**Remark:** It is easy to show that

\[
a_{k,n+1} = \binom{n+k-1}{n-1} + \binom{n+k-2}{n-1}
\]

for \( k \geq 2 \); see Morimoto [132] (Chapter 2, Theorem 2.4) or Dieudonné [51] (Chapter 7, formula 99), where a different proof technique is used.

Let \( L^2(S^n) \) be the space of (real) square-integrable functions on the sphere \( S^n \). We have an inner product on \( L^2(S^n) \) given by

\[
\langle f, g \rangle_{S^n} = \int_{S^n} fg \, \text{Vol}_{S^n},
\]

where \( f, g \in L^2(S^n) \) and where \( \text{Vol}_{S^n} \) is the volume form on \( S^n \) (induced by the metric on \( \mathbb{R}^{n+1} \)). With this inner product, \( L^2(S^n) \) is a complete normed vector space using the norm \( \|f\| = \|f\|_2 = \sqrt{\langle f, f \rangle} \) associated with this inner product; that is, \( L^2(S^n) \) is a Hilbert space. In the case of complex-valued functions, we use the Hermitian inner product

\[
\langle f, g \rangle_{S^n} = \int_{S^n} f \overline{g} \, \text{Vol}_{S^n},
\]

and we get the complex Hilbert space \( L^2_{\mathbb{C}}(S^n) \) (see Section 24.7 for the definition of the integral of a complex-valued function). We also denote by \( C(S^n) \) the space of continuous (real) functions on \( S^n \) with the \( L^\infty \)-norm; that is,

\[
\|f\|_{\infty} = \sup\{|f(x)|\}_{x \in S^n},
\]

and by \( C_{\mathbb{C}}(S^n) \) the space of continuous complex-valued functions on \( S^n \) also with the \( L^\infty \) norm. Recall that \( C(S^n) \) is dense in \( L^2(S^n) \) (and \( C_{\mathbb{C}}(S^n) \) is dense in \( L^2_{\mathbb{C}}(S^n) \)). The following proposition shows why the spherical harmonics play an important role:

**Proposition 26.13.** The set of all finite linear combinations of elements in \( \bigcup_{k=0}^{\infty} \mathcal{H}_k(S^n) \) (resp. \( \bigcup_{k=0}^{\infty} \mathcal{H}_{k}^{\mathbb{C}}(S^n) \)) is

(i) dense in \( C(S^n) \) (resp. in \( C_{\mathbb{C}}(S^n) \)) with respect to the \( L^\infty \)-norm;

(ii) dense in \( L^2(S^n) \) (resp. dense in \( L^2_{\mathbb{C}}(S^n) \)).

**Proof.** (i) As \( S^n \) is compact, by the Stone-Weierstrass approximation theorem (Lang [112], Chapter III, Corollary 1.3), if \( g \) is continuous on \( S^n \), then it can be approximated uniformly by polynomials \( P_j \) restricted to \( S^n \). By Corollary 26.11, the restriction of each \( P_j \) to \( S^n \) is a linear combination of elements in \( \bigcup_{k=0}^{\infty} \mathcal{H}_k(S^n) \).
(ii) We use the fact that \( C(S^n) \) is dense in \( L^2(S^n) \). Given \( f \in L^2(S^n) \), for every \( \epsilon > 0 \), we can choose a continuous function \( g \) so that \( \|f - g\|_2 < \epsilon/2 \). By (i), we can find a linear combination \( h \) of elements in \( \bigcup_{k=0}^{\infty} H_k(S^n) \) so that \( \|g - h\|_\infty < \epsilon/(2\sqrt{\text{vol}(S^n)}) \), where \( \text{vol}(S^n) \) is the volume of \( S^n \) (really, area). Thus, we get

\[
\|f - h\|_2 \leq \|f - g\|_2 + \|g - h\|_2 < \epsilon/2 + \sqrt{\text{vol}(S^n)} \|g - h\|_\infty < \epsilon/2 + \epsilon/2 = \epsilon,
\]

which proves (ii). The proof in the complex case is identical.

We need one more proposition before showing that the spaces \( H_k(S^n) \) constitute an orthogonal Hilbert space decomposition of \( L^2(S^n) \).

**Proposition 26.14.** For every harmonic polynomial \( P \in \mathcal{H}_k(n+1) \) (resp. \( P \in \mathcal{H}_k^C(n+1) \)), the restriction \( H \in \mathcal{H}_k(S^n) \) (resp. \( H \in \mathcal{H}_k^C(S^n) \)) of \( P \) to \( S^n \) is an eigenfunction of \( \Delta_{S^n} \) for the eigenvalue \(-k(n + k - 1)\).

**Proof.** We have

\[
P(r\sigma) = r^k H(\sigma), \quad r > 0, \sigma \in S^n,
\]

and by Proposition 26.8, for any \( f \in C^\infty(\mathbb{R}^{n+1}) \), we have

\[
\Delta f = \frac{1}{r^n} \frac{\partial}{\partial r} \left( r^n \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^n} f.
\]

Consequently,

\[
\Delta P = \Delta(r^k H) = \frac{1}{r^n} \frac{\partial}{\partial r} \left( r^n \frac{\partial (r^k H)}{\partial r} \right) + \frac{1}{r^2} \Delta_{S^n} (r^k H)
\]

\[
= \frac{1}{r^n} \frac{\partial}{\partial r} \left( k r^{n+k-1} H \right) + r^{k-2} \Delta_{S^n} H
\]

\[
= \frac{1}{r^n} k(n + k - 1) r^{n+k-2} H + r^{k-2} \Delta_{S^n} H
\]

\[
= r^{k-2} (k(n + k - 1) H + \Delta_{S^n} H).
\]

Thus,

\[
\Delta P = 0 \quad \text{iff} \quad \Delta_{S^n} H = -k(n + k - 1) H,
\]

as claimed. \(\square\)

From Proposition 26.14, we deduce that the space \( \mathcal{H}_k(S^n) \) is a subspace of the eigenspace \( E_k \) of \( \Delta_{S^n} \) associated with the eigenvalue \(-k(n + k - 1)\) (and similarly for \( \mathcal{H}_k^C(S^n) \)). Remarkably, \( E_k = \mathcal{H}_k(S^n) \), but it will take more work to prove this.

What we can deduce immediately is that \( \mathcal{H}_k(S^n) \) and \( \mathcal{H}_l(S^n) \) are pairwise orthogonal whenever \( k \neq l \). This is because, by Proposition 26.9, the Laplacian is self-adjoint, and thus any two eigenspaces \( E_k \) and \( E_l \) are pairwise orthogonal whenever \( k \neq l \), and as \( \mathcal{H}_k(S^n) \subseteq \).
$E_k$ and $\mathcal{H}_k(S^n) \subseteq E_1$, our claim is indeed true. Furthermore, by Proposition 26.12, each $\mathcal{H}_k(S^n)$ is finite-dimensional, and thus closed. Finally, we know from Proposition 26.13 that $\bigcup_{k=0}^{\infty} \mathcal{H}_k(S^n)$ is dense in $L^2(S^n)$. But then, we can apply a standard result from Hilbert space theory (for example, see Lang [112], Chapter V, Proposition 1.9) to deduce the following important result:

**Theorem 26.15.** The family of spaces $\mathcal{H}_k(S^n)$ (resp. $\mathcal{H}_k^C(S^n)$) yields a Hilbert space direct sum decomposition

$$L^2(S^n) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k(S^n) \quad \text{(resp.} \quad L^2_C(S^n) = \bigoplus_{k=0}^{\infty} \mathcal{H}_k^C(S^n)),
$$

which means that the summands are closed, pairwise orthogonal, and that every $f \in L^2(S^n)$ (resp. $f \in L^2_C(S^n)$) is the sum of a converging series

$$f = \sum_{k=0}^{\infty} f_k$$

in the $L^2$-norm, where the $f_k \in \mathcal{H}_k(S^n)$ (resp. $f_k \in \mathcal{H}_k^C(S^n)$) are uniquely determined functions. Furthermore, given any orthonormal basis $(Y_{k,1}, \ldots, Y_{k,a_{k,n+1}})$ of $\mathcal{H}_k(S^n)$, we have

$$f_k = \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_{k,m_k}^{m_k}, \quad \text{with} \quad c_{k,m_k} = \langle f, Y_{k,m_k}^{m_k} \rangle_{S^n}.$$

The coefficients $c_{k,m_k}$ are “generalized” Fourier coefficients with respect to the Hilbert basis $\{Y_{k,m_k}^{m_k} \mid 1 \leq m_k \leq a_{k,n+1}, k \geq 0\}$. We can finally prove the main theorem of this section.

**Theorem 26.16.**

1. The eigenspaces (resp. complex eigenspaces) of the Laplacian $\Delta_{S^n}$ on $S^n$ are the spaces of spherical harmonics

$$E_k = \mathcal{H}_k(S^n) \quad \text{(resp.} \quad E_k = \mathcal{H}_k^C(S^n)),$

and $E_k$ corresponds to the eigenvalue $-k(n+k-1)$.

2. We have the Hilbert space direct sum decompositions

$$L^2(S^n) = \bigoplus_{k=0}^{\infty} E_k \quad \text{(resp.} \quad L^2_C(S^n) = \bigoplus_{k=0}^{\infty} E_k).$$

3. The complex polynomials of the form $(c_1 x_1 + \cdots + c_{n+1} x_{n+1})^k$, with $c_1^2 + \cdots + c_{n+1}^2 = 0$, span the space $\mathcal{H}_k^C(n+1)$, for $k \geq 1$.  

Proof. We follow essentially the proof in Helgason [87] (Introduction, Theorem 3.1). In (1) and (2) we only deal with the real case, the proof in the complex case being identical.

(1) We already know that the integers \(-k(n + k - 1)\) are eigenvalues of \(\Delta_{S^n}\) and that \(\mathcal{H}_k(S^n) \subseteq E_k\). We will prove that \(\Delta_{S^n}\) has no other eigenvalues and no other eigenvectors using the Hilbert basis \(\{Y_{mk}^{\pm} \mid 1 \leq m_k \leq a_{k,n+1}, k \geq 0\}\) given by Theorem 26.15. Let \(\lambda\) be any eigenvalue of \(\Delta_{S^n}\) and let \(f \in L^2(S^n)\) be any eigenfunction associated with \(\lambda\) so that \(\Delta f = \lambda f\).

We have a unique series expansion

\[
f = \sum_{k=0}^{\infty} a_{k,n+1} \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_{k,m_k}^{mk},
\]

with \(c_{k,m_k} = \langle f, Y_{k,m_k}^{mk} \rangle_{S^n}\). Now, as \(\Delta_{S^n}\) is self-adjoint and \(\Delta Y_{k,m_k}^{mk} = -k(n + k - 1)Y_{k,m_k}^{mk}\), the Fourier coefficients \(d_{k,m_k}\) of \(\Delta f\) are given by

\[
d_{k,m_k} = \langle \Delta f, Y_{k,m_k}^{mk} \rangle_{S^n} = \langle f, \Delta Y_{k,m_k}^{mk} \rangle_{S^n} = -k(n + k - 1)\langle f, Y_{k,m_k}^{mk} \rangle_{S^n} = -k(n + k - 1)c_{k,m_k}.
\]

On the other hand, as \(\Delta f = \lambda f\), the Fourier coefficients of \(\Delta f\) are given by

\[
d_{k,m_k} = \lambda c_{k,m_k}.
\]

By uniqueness of the Fourier expansion, we must have

\[
\lambda c_{k,m_k} = -k(n + k - 1)c_{k,m_k} \quad \text{for all } k \geq 0.
\]

Since \(f \neq 0\), there some \(k\) such that \(c_{k,m_k} \neq 0\), and we must have

\[
\lambda = -k(n + k - 1)
\]

for any such \(k\). However, the function \(k \mapsto -k(n + k - 1)\) reaches its maximum for \(k = -\frac{n-1}{2}\), and as \(n \geq 1\), it is strictly decreasing for \(k \geq 0\), which implies that \(k\) is unique and that

\[
c_{j,m_j} = 0 \quad \text{for all } j \neq k.
\]

Therefore \(f \in \mathcal{H}_k(S^n)\), and the eigenvalues of \(\Delta_{S^n}\) are exactly the integers \(-k(n + k - 1)\), so \(E_k = \mathcal{H}_k(S^n)\) as claimed.

Since we just proved that \(E_k = \mathcal{H}_k(S^n)\), (2) follows immediately from the Hilbert decomposition given by Theorem 26.15.

(3) If \(H = (c_1 x_1 + \cdots + c_{n+1} x_{n+1})^k\), with \(c_1^2 + \cdots + c_{n+1}^2 = 0\), then for \(k \leq 1\) is is obvious that \(\Delta H = 0\), and for \(k \geq 2\) we have

\[
\Delta H = k(k-1)(c_1^2 + \cdots + c_{n+1}^2)(c_1 x_1 + \cdots + c_{n+1} x_{n+1})^{k-2} = 0,
\]
so \( H \in \mathcal{H}_k^C(n+1) \). A simple computation shows that for every \( Q \in \mathcal{P}_k^C(n+1) \), if \( c = (c_1, \ldots, c_{n+1}) \), then we have

\[
\partial(Q)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^m = m(m-1) \cdots (m-k+1)Q(c)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^{m-k},
\]

for all \( m \geq k \geq 1 \).

Assume that \( \mathcal{H}_k^C(n+1) \) is not spanned by the complex polynomials of the form \( (c_1x_1 + \cdots + c_{n+1}x_{n+1})^k \), with \( c_1^2 + \cdots + c_{n+1}^2 = 0 \), for \( k \geq 1 \). Then, some \( Q \in \mathcal{H}_k^C(n+1) \) is orthogonal to all polynomials of the form \( H = (c_1x_1 + \cdots + c_{n+1}x_{n+1})^k \), with \( c_1^2 + \cdots + c_{n+1}^2 = 0 \). Recall that

\[
\langle P, \partial(Q)H \rangle = \langle QP, H \rangle
\]

and apply this equation to \( P = Q(c) \), \( H \) and \( Q \). Since

\[
\partial(Q)H = \partial(Q)(c_1x_1 + \cdots + c_{n+1}x_{n+1})^k = k!Q(c),
\]

as \( Q \) is orthogonal to \( H \), we get

\[
k!\langle Q(c), Q(c) \rangle = \langle Q(c), k!Q(c) \rangle = \langle Q(c), \partial(Q)H \rangle = \langle QQ(c), H \rangle = Q(c)\langle Q, H \rangle = 0,
\]

which implies \( Q(c) = 0 \). Consequently, \( Q(x_1, \ldots, x_{n+1}) \) vanishes on the complex algebraic variety

\[
\{(x_1, \ldots, x_{n+1}) \in \mathbb{C}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 0\}.
\]

By the Hilbert Nullstellensatz, some power \( Q^m \) belongs to the ideal \( (x_1^2 + \cdots + x_{n+1}^2) \) generated by \( x_1^2 + \cdots + x_{n+1}^2 \). Now, if \( n \geq 2 \), it is well-known that the polynomial \( x_1^2 + \cdots + x_{n+1}^2 \) is irreducible so the ideal \( (x_1^2 + \cdots + x_{n+1}^2) \) is a prime ideal, and thus \( Q \) is divisible by \( x_1^2 + \cdots + x_{n+1}^2 \). However, we know from the proof of Theorem 26.10 that we have an orthogonal direct sum

\[
\mathcal{P}_k^C(n+1) = \mathcal{H}_k^C(n+1) \oplus \|x\|^2 \mathcal{P}_k^C(n+2)\).
\]

Since \( Q \in \mathcal{H}_k^C(n+1) \) and \( Q \) is divisible by \( x_1^2 + \cdots + x_{n+1}^2 \), we must have \( Q = 0 \). Therefore, if \( n \geq 2 \), we proved (3). However, when \( n = 1 \), we know from Section 26.2 that the complex harmonic homogeneous polynomials in two variables \( P(x, y) \) are spanned by the real and imaginary parts \( U_k, V_k \) of the polynomial \( (x + iy)^k = U_k + iV_k \). Since \( (x - iy)^k = U_k - iV_k \) we see that

\[
U_k = \frac{1}{2} \left( (x + iy)^k + (x - iy)^k \right), \quad V_k = \frac{1}{2i} \left( (x + iy)^k - (x - iy)^k \right),
\]

and as \( 1 + i^2 = 1 + (-i)^2 = 0 \), the space \( \mathcal{H}_k^C(\mathbb{R}^2) \) is spanned by \( (x + iy)^k \) and \( (x - iy)^k \) (for \( k \geq 1 \)), so (3) holds for \( n = 1 \) as well.

As an illustration of part (3) of Theorem 26.16, the polynomials \( (x_1 + i \cos \theta x_2 + i \sin \theta x_3)^k \) are harmonic. Of course, the real and imaginary part of a complex harmonic polynomial \( (c_1x_1 + \cdots + c_{n+1}x_{n+1})^k \) are real harmonic polynomials.
CHAPTER 26. SPHERICAL HARMONICS AND LINEAR REPRESENTATIONS

26.6 Reproducing Kernel, Zonal Spherical Functions and Gegenbauer Polynomials

In this section we describe the zonal spherical functions on $S^n$ and show that they essentially come from certain polynomials generalizing the Legendre polynomials known as the Gegenbauer Polynomials. Most proofs will be omitted. We refer the reader to Stein and Weiss [165] (Chapter 4) and Morimoto [132] (Chapter 2) for a complete exposition with proofs.

In order to define spherical zonal harmonics we will need the following proposition.

**Proposition 26.17.** If $P$ is any (complex) polynomial in $n$ variables such that $P(R(x)) = P(x)$ for all rotations $R \in SO(n)$, and all $x \in \mathbb{R}^n$, then $P$ is of the form

$$P(x) = \sum_{j=0}^{m} c_j (x_1^2 + \cdots + x_n^2)^j,$$

for some $c_0, \ldots, c_m \in \mathbb{C}$.

**Proof.** Write $P$ as the sum of its homogeneous pieces $P = \sum_{l=0}^{k} Q_l$, where $Q_l$ is homogeneous of degree $l$. For every $\epsilon > 0$ and every rotation $R$, we have

$$\sum_{l=0}^{k} \epsilon^l Q_l(x) = P(\epsilon x) = P(R(\epsilon x)) = P(\epsilon R(x)) = \sum_{l=0}^{k} \epsilon^l Q_l(R(x)),$$

which implies that

$$Q_l(R(x)) = Q_l(x), \quad l = 0, \ldots, k.$$ 

If we let $F_l(x) = \|x\|^{-l} Q_l(x)$, then $F_l$ is a homogeneous function of degree 0, and $F_l$ is invariant under all rotations. This is only possible if $F_l$ is a constant function, thus $F_l(x) = a_l$ for all $x \in \mathbb{R}^n$. But then, $Q_l(x) = a_l \|x\|^l$. Since $Q_l$ is a polynomial, $l$ must be even whenever $a_l \neq 0$. It follows that

$$P(x) = \sum_{j=0}^{m} c_j \|x\|^{2j}$$

with $c_j = a_{2j}$ for $j = 0, \ldots, m$, and where $m$ is the largest integer $\leq k/2$. \qed

Proposition 26.17 implies that if a polynomial function on the sphere $S^n$, in particular a spherical harmonic, is invariant under all rotations, then it is a constant. If we relax this condition to invariance under all rotations leaving some given point $\tau \in S^n$ invariant, then we obtain zonal harmonics.

The following theorem from Morimoto [132] (Chapter 2, Theorem 2.24) gives the relationship between zonal harmonics and the Gegenbauer polynomials:
**Theorem 26.18.** Fix any $\tau \in S^n$. For every constant $c \in \mathbb{C}$, there is a unique homogeneous harmonic polynomial $Z_k^\tau \in \mathcal{H}_k^C(n+1)$ satisfying the following conditions:

1. $Z_k^\tau(\tau) = c$;

2. For every rotation $R \in \text{SO}(n+1)$, if $R\tau = \tau$, then $Z_k^\tau(R(x)) = Z_k^\tau(x)$ for all $x \in \mathbb{R}^{n+1}$.

Furthermore, we have

$$Z_k^\tau(x) = c \|x\|^k P_{k,n}(\frac{x}{\|x\|} \cdot \tau),$$

for some polynomial $P_{k,n}(t)$ of degree $k$.

**Remark:** The proof given in Morimoto [132] is essentially the same as the proof of Theorem 2.12 in Stein and Weiss [165] (Chapter 4), but Morimoto makes an implicit use of Proposition 26.17 above. Also, Morimoto states Theorem 26.18 only for $c = 1$, but the proof goes through for any $c \in \mathbb{C}$, including $c = 0$, and we will need this extra generality in the proof of the Funk-Hecke formula.

**Proof.** Let $e_{n+1} = (0, \ldots, 0, 1) \in \mathbb{R}^{n+1}$, and for any $\tau \in S^n$, let $R_\tau$ be some rotation such that $R_\tau(e_{n+1}) = \tau$. Assume $Z \in \mathcal{H}_k^C(n+1)$ satisfies conditions (1) and (2), and let $Z'$ be given by $Z'(x) = Z(R_\tau(x))$. As $R_\tau(e_{n+1}) = \tau$, we have $Z'(e_{n+1}) = Z(\tau) = c$. Furthermore, for any rotation $S$ such that $S(e_{n+1}) = e_{n+1}$, observe that

$$R_\tau \circ S \circ R_\tau^{-1}(\tau) = R_\tau \circ S(e_{n+1}) = R_\tau(e_{n+1}) = \tau,$$

and so, as $Z$ satisfies property (2) for the rotation $R_\tau \circ S \circ R_\tau^{-1}$, we get

$$Z'(S(x)) = Z(R_\tau \circ S(x)) = Z(R_\tau \circ S \circ R_\tau^{-1} \circ R_\tau(x)) = Z(R_\tau(x)) = Z'(x),$$

which proves that $Z'$ is a harmonic polynomial satisfying properties (1) and (2) with respect to $e_{n+1}$. Therefore, we may assume that $\tau = e_{n+1}$.

Write

$$Z(x) = \sum_{j=0}^{k} x_{n+1}^{k-j} P_j(x_1, \ldots, x_n),$$

where $P_j(x_1, \ldots, x_n)$ is a homogeneous polynomial of degree $j$. Since $Z$ is invariant under every rotation $R$ fixing $e_{n+1}$, and since the monomials $x_{n+1}^{k-j}$ are clearly invariant under such a rotation, we deduce that every $P_j(x_1, \ldots, x_n)$ is invariant under all rotations of $\mathbb{R}^{n}$. (Clearly, there is a one-to-one correspondence between the rotations of $\mathbb{R}^{n+1}$ fixing $e_{n+1}$ and the rotations of $\mathbb{R}^{n}$). By Proposition 26.17, we conclude that

$$P_j(x_1, \ldots, x_n) = c_j(x_1^2 + \cdots + x_n^2)^{\frac{j}{2}},$$
which implies that $P_j = 0$ if $j$ is odd. Thus, we can write

$$Z(x) = \sum_{i=0}^{[k/2]} c_i x_{n+1}^{k-2i} (x_1^2 + \cdots + x_n^2)^i$$

where $[k/2]$ is the greatest integer $m$ such that $2m \leq k$. If $k < 2$, then $Z = c_0$, so $c_0 = c$ and $Z$ is uniquely determined. If $k \geq 2$, we know that $Z$ is a harmonic polynomial so we assert that $\Delta Z = 0$. For $i \leq j \leq n$,

$$\frac{\partial}{\partial x_j} (x_1^2 + \cdots + x_j^2 + \cdots x_n^2)^i = 2ix_j (x_1^2 + \cdots x_n^2)^{i-1},$$

and

$$\frac{\partial^2}{\partial x_j^2} (x_1^2 + \cdots + x_j^2 + \cdots x_n^2)^i = 2i(x_1^2 + \cdots x_n^2)^{i-1} + 4x_j^2 i(i-1)(x_1^2 + \cdots x_n^2)^{i-2}
$$

$$= 2i(x_1^2 + \cdots x_n^2)^{i-2} [x_1^2 + \cdots + 2(i-1)x_j^2].$$

Since $\Delta(x_1^2 + \cdots + x_n^2)^i = \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2} (x_1^2 + \cdots + x_j^2 + \cdots x_n^2)^i$, we find that

$$\Delta(x_1^2 + \cdots + x_n^2)^i = 2i(x_1^2 + \cdots x_n^2)^{i-2} \sum_{j=1}^{n} [x_1^2 + \cdots x_n^2 + 2(i-1)x_j^2]
$$

$$= 2i(x_1^2 + \cdots x_n^2)^{i-2} \left[ n(x_1^2 + \cdots x_n^2) + 2(i-1)\sum_{j=1}^{n} x_j^2 \right]
$$

$$= 2i(x_1^2 + \cdots x_n^2)^{i-2} [n(x_1^2 + \cdots x_n^2) + 2(i-1)(x_1^2 + \cdots x_n^2)]
$$

$$= 2i(n+2i-2)(x_1^2 + \cdots + x_n^2)^{i-1}.$$ 

Thus

$$\Delta x_{n+1}^{k-2i} (x_1^2 + \cdots + x_n^2)^i = (k-2i)(k-2i-1)x_{n+1}^{k-2i-2} (x_1^2 + \cdots + x_n^2)^i
$$

$$+ x_{n+1}^{k-2i} \Delta(x_1^2 + \cdots + x_n^2)^i
$$

$$= (k-2i)(k-2i-1)x_{n+1}^{k-2i-2} (x_1^2 + \cdots + x_n^2)^i
$$

$$+ 2i(n+2i-2)x_{n+1}^{k-2i} (x_1^2 + \cdots + x_n^2)^{i-1},$$

and so we get

$$\Delta Z = \sum_{i=0}^{[k/2]-1} ((k-2i)(k-2i-1)c_i + 2(i+1)(n+2i)c_{i+1}) x_{n+1}^{k-2i-2} (x_1^2 + \cdots + x_n^2)^i.$$

Then $\Delta Z = 0$ yields the relations

$$2(i+1)(n+2i)c_{i+1} = -(k-2i)(k-2i-1)c_i, \quad i = 0, \ldots, [k/2] - 1.$$
which shows that $Z$ is uniquely determined up to the constant $c_0$. Since we are requiring $Z(e_{n+1}) = c$, we get $c_0 = c$, and $Z$ is uniquely determined. Now on $S^n$ we have $x_1^2 + \cdots + x_{n+1}^2 = 1$, so if we let $t = x_{n+1}$, for $c_0 = 1$, we get a polynomial in one variable

$$P_{k,n}(t) = \sum_{i=0}^{[k/2]} c_i t^{k-2i} (1 - t^2)^i.$$ 

Thus, we proved that when $Z(e_{n+1}) = c$, we have

$$Z(x) = c \|x\|^k P_{k,n} \left( \frac{x_{n+1}}{\|x\|} \right) = c \|x\|^k P_{k,n} \left( \frac{x}{\|x\|} \cdot e_{n+1} \right).$$

When $Z(\tau) = c$, we write $Z = Z' \circ R^{-1}_\tau$ with $Z' = Z \circ R_\tau$ and where $R_\tau$ is a rotation such that $R_\tau(e_{n+1}) = \tau$. Then, as $Z'(e_{n+1}) = c$, using the formula above for $Z'$, we have

$$Z(x) = Z'(R^{-1}_\tau(x)) = c \|R^{-1}_\tau(x)\|^k P_{k,n} \left( \frac{R^{-1}_\tau(x)}{\|R^{-1}_\tau(x)\|} \cdot e_{n+1} \right) = c \|x\|^k P_{k,n} \left( \frac{x}{\|x\|} \cdot R_\tau(e_{n+1}) \right) = c \|x\|^k P_{k,n} \left( \frac{x}{\|x\|} \cdot \tau \right),$$

since $R_\tau$ is an isometry.

\[\square\]

**Definition 26.12.** The function, $Z'_k$, is called a *zonal function* and its restriction to $S^n$ is a *zonal spherical function*. The polynomial $P_{k,n}(t)$ is called the *Gegenbauer polynomial* of degree $k$ and dimension $n + 1$ or *ultraspherical polynomial*. By definition, $P_{k,n}(1) = 1$.

The proof of Theorem 26.18 shows that for $k$ even, say $k = 2m$, the polynomial $P_{2m,n}$ is of the form

$$P_{2m,n}(t) = \sum_{j=0}^{m} c_{m-j} t^{2j} (1 - t^2)^{m-j},$$

and for $k$ odd, say $k = 2m + 1$, the polynomial $P_{2m+1,n}$ is of the form

$$P_{2m+1,n}(t) = \sum_{j=0}^{m} c_{m-j} t^{2j+1} (1 - t^2)^{m-j}.$$

Consequently, $P_{k,n}(-t) = (-1)^k P_{k,n}(t)$, for all $k \geq 0$. The proof also shows that the “natural basis” for these polynomials consists of the polynomials, $t^i (1 - t^2)^{k/2}$, with $k - i$ even. Indeed, with this basis, there are simple recurrence equations for computing the coefficients of $P_{k,n}(t)$.

**Remark:** Morimoto [132] calls the polynomials $P_{k,n}(t)$ “Legendre polynomials.” For $n = 2$, they are indeed the Legendre polynomials. Stein and Weiss denotes our (and Morimoto’s) $P_{k,n}(t)$ by $P^{[n]}_k (t)$ (up to a constant factor), and Dieudonné [51] (Chapter 7) by $G_{k,n+1}(t)$. 

When \( n = 2 \), using the notation of Section 26.3, the zonal spherical functions on \( S^2 \) are the spherical harmonics \( y^0_l \) for which \( m = 0 \); that is (up to a constant factor),

\[
y^0_l(\theta, \varphi) = \sqrt{\frac{(2l+1)}{4\pi}} P_l(\cos \theta),
\]

where \( P_l \) is the Legendre polynomial of degree \( l \). For example, for \( l = 2 \), \( P_2(t) = \frac{1}{2}(3t^2 - 1) \).

Zonal spherical functions have many important properties. One such property is associated with the reproducing kernel of \( \mathcal{H}_k^C(S^n) \).

**Definition 26.13.** Let \( \mathcal{H}_k^C(S^n) \) be the space of spherical harmonics. Let \( a_{k,n+1} \) be the dimension of \( \mathcal{H}_k^C(S^n) \) where

\[
a_{k,n+1} = \binom{n+k}{k} - \binom{n+k-2}{k-2},
\]

if \( n \geq 1 \) and \( k \geq 2 \), with \( a_{0,n+1} = 1 \) and \( a_{1,n+1} = n + 1 \). Let \( (Y^1_k, \ldots, Y^{a_{k,n+1}}_k) \) be any orthonormal basis of \( \mathcal{H}_k^C(S^n) \), and define \( F_k(\sigma, \tau) \) by

\[
F_k(\sigma, \tau) = \sum_{i=1}^{a_{k,n+1}} Y^i_k(\sigma)Y^i_k(\tau), \quad \sigma, \tau \in S^n.
\]

The function \( F_k(\sigma, \tau) \) is the reproducing kernel of \( \mathcal{H}_k^C(S^n) \).

The following proposition is easy to prove (see Morimoto [132], Chapter 2, Lemma 1.19 and Lemma 2.20):

**Proposition 26.19.** The function \( F_k \) is independent of the choice of orthonormal basis. Furthermore, for every orthogonal transformation \( R \in O(n+1) \), we have

\[
F_k(R\sigma, R\tau) = F_k(\sigma, \tau), \quad \sigma, \tau \in S^n.
\]

Clearly, \( F_k \) is a symmetric function. Since we can pick an orthonormal basis of real orthogonal functions for \( \mathcal{H}_k^C(S^n) \) (pick a basis of \( \mathcal{H}_k(S^n) \)), Proposition 26.19 shows that \( F_k \) is a real-valued function.

The function \( F_k \) satisfies the following property which justifies its name as the reproducing kernel for \( \mathcal{H}_k^C(S^n) \):

**Proposition 26.20.** For every spherical harmonic \( H \in \mathcal{H}_j^C(S^n) \), we have

\[
\int_{S^n} H(\tau)F_k(\sigma, \tau) \text{Vol}_{S^n} = \delta_{j,k} \delta(\sigma), \quad \sigma, \tau \in S^n,
\]

for all \( j, k \geq 0 \).
26.6. REPRODUCING KERNEL AND ZONAL SPHERICAL FUNCTIONS

Proof. When $j \neq k$, since $\mathcal{H}_j^C(S^n)$ and $\mathcal{H}_k^C(S^n)$ are orthogonal and since $F_k(\sigma, \tau) = \sum_{i=1}^{a_k,n+1} Y^i_k(\sigma)Y^i_k(\tau)$, it is clear that the integral in Proposition 26.20 vanishes. When $j = k$, we have

$$\int_{S^n} H(\tau) F_k(\sigma, \tau) \text{Vol}_{S^n} = \int_{S^n} H(\tau) \sum_{i=1}^{a_k,n+1} Y^i_k(\sigma)Y^i_k(\tau) \text{Vol}_{S^n}$$

$$= \sum_{i=1}^{a_k,n+1} Y^i_k(\sigma) \int_{S^n} H(\tau)Y^i_k(\tau) \text{Vol}_{S^n}$$

$$= \sum_{i=1}^{a_k,n+1} Y^i_k(\sigma) \langle H, Y^i_k \rangle$$

$$= H(\sigma),$$

since $(Y^1_k, \ldots, Y^{a_k,n+1}_k)$ is an orthonormal basis. \hfill \square

Remark: In Stein and Weiss [165] (Chapter 4), the function $F_k(\sigma, \tau)$ is denoted by $Z^k_\sigma(\tau)$ and it is called the zonal harmonic of degree $k$ with pole $\sigma$.

Before we investigate the relationship between $F_k(\sigma, \tau)$ and $Z^k_\tau(\sigma)$, we need two technical propositions. Both are proven in Morimoto [132]. The first, Morimoto [132] (Chapter 2, Lemma 2.21), is needed to prove the second, Morimoto [132] (Chapter 2, Lemma 2.23).

Proposition 26.21. For all $\sigma, \tau, \sigma', \tau' \in S^n$, with $n \geq 1$, the following two conditions are equivalent:

(i) There is some orthogonal transformation $R \in O(n + 1)$ such that $R(\sigma) = \sigma'$ and $R(\tau) = \tau'$.

(ii) $\sigma \cdot \tau = \sigma' \cdot \tau'$.

Propositions 26.20 and 26.21 immediately yield

Proposition 26.22. For all $\sigma, \tau, \sigma', \tau' \in S^n$, if $\sigma \cdot \tau = \sigma' \cdot \tau'$, then $F_k(\sigma, \tau) = F_k(\sigma', \tau')$. Consequently, there is some function $\varphi: \mathbb{R} \to \mathbb{R}$ such that $F_k(\sigma, \tau) = \varphi(\sigma \cdot \tau)$.

We claim that the $\varphi(\sigma \cdot \tau)$ of Proposition 26.22 is a zonal spherical function $Z^k_\tau(\sigma)$. To see why this is true, define $Z(r^k(\sigma)) := r^k F_k(\sigma, \tau)$ for a fixed $\tau$. By the definition of $F_k(\sigma, \tau)$, it is clear that $Z$ is a homogeneous harmonic polynomial. The value $F_k(\tau, \tau)$ does not depend of $\tau$, because by transitivity of the action of $SO(n + 1)$ on $S^n$, for any other $\sigma \in S^n$, there is some rotation $R$ so that $R\tau = \sigma$, and by Proposition 26.19, we have $F_k(\sigma, \sigma) = F_k(\tau, R\tau) = F_k(\tau, \tau)$. To compute $F_k(\tau, \tau)$, since

$$F_k(\tau, \tau) = \sum_{i=1}^{a_k,n+1} \|Y^i_k(\tau)\|^2,$$
and since \((Y_k^1, \ldots, Y_k^{a_k,n+1})\) is an orthonormal basis of \(H^C_k(S^n)\), observe that

\[
a_{k,n+1} = \sum_{i=1}^{a_k,n+1} \langle Y_k^i, Y_k^i \rangle \tag{26.1}
\]

\[
= \sum_{i=1}^{a_k,n+1} \int_{S^n} \|Y_k^i(\tau)\|^2 \text{Vol}_{S^n} \tag{26.2}
\]

\[
= \int_{S^n} \left( \sum_{i=1}^{a_k,n+1} \|Y_k^i(\tau)\|^2 \right) \text{Vol}_{S^n} \tag{26.3}
\]

\[
= \int_{S^n} F_k(\tau, \tau) \text{Vol}_{S^n} = F_k(\tau, \tau) \text{vol}(S^n). \tag{26.4}
\]

Therefore,

\[
F_k(\tau, \tau) = \frac{a_{k,n+1}}{\text{vol}(S^n)}. \tag{26.5}
\]

\(\mathbb{Z}\) Beware that Morimoto [132] uses the normalized measure on \(S^n\), so the factor involving \(\text{vol}(S^n)\) does not appear.

**Remark:** Recall that

\[
\text{vol}(S^{2d}) = \frac{2^{d+1} \pi^d}{1 \cdot 3 \cdots (2d-1)} \quad \text{if} \quad d \geq 1 \quad \text{and} \quad \text{vol}(S^{2d+1}) = \frac{2\pi^{d+1}}{d!} \quad \text{if} \quad d \geq 0.
\]

Now, if \(R\tau = \tau\), Proposition 26.19 shows that

\[
Z(R^k(\sigma)) = Z(r^k\sigma) = r^k F_k(R\sigma, \tau) = r^k F_k(R\sigma, R\tau) = r^k F_k(\sigma, \tau) = Z(r^k\sigma).
\]

Therefore, the function \(Z\) satisfies conditions (1) and (2) of Theorem 26.18 with \(c = \frac{a_{k,n+1}}{\text{vol}(S^n)}\), and by uniqueness, we conclude that \(Z\) is the zonal function \(Z_k^\tau\) whose restriction to \(S^n\) is the zonal spherical function

\[
F_k(\sigma, \tau) = \frac{a_{k,n+1}}{\text{vol}(S^n)} P_{k,n}(\sigma \cdot \tau).
\]

Consequently, we have obtained the so-called *addition formula*:

**Proposition 26.23. (Addition Formula)** If \((Y_k^1, \ldots, Y_k^{a_k,n+1})\) is any orthonormal basis of \(H^C_k(S^n)\), then

\[
P_{k,n}(\sigma \cdot \tau) = \frac{\text{vol}(S^n)}{a_{k,n+1}} \sum_{i=1}^{a_k,n+1} Y_k^i(\sigma) Y_k^i(\tau).
\]
Again, beware that Morimoto [132] does not have the factor \( \text{vol}(S^n) \).

For \( n = 1 \), we can write \( \sigma = (\cos \theta, \sin \theta) \) and \( \tau = (\cos \varphi, \sin \varphi) \), and it is easy to see that the addition formula reduces to

\[
P_{k,1}(\cos(\theta - \varphi)) = \cos k\theta \cos k\varphi + \sin k\theta \sin k\varphi = \cos k(\theta - \varphi),
\]

the standard addition formula for trigonometric functions.

Proposition 26.23 implies that \( P_{k,n}(t) \) has real coefficients. Furthermore Proposition 26.20 is reformulated as

\[
\frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} P_{k,n}(\sigma \cdot \tau) H(\sigma) \, \text{Vol}_{S^n} = \delta_{jk} H(\sigma),
\]

showing that the Gegenbauer polynomials are reproducing kernels. A neat application of this formula is a formula for obtaining the \( k \)th spherical harmonic component of a function \( f \in L^2_C(S^n) \).

**Proposition 26.24.** For every function \( f \in L^2_C(S^n) \), if \( f = \sum_{k=0}^{\infty} f_k \) is the unique decomposition of \( f \) over the Hilbert sum \( \bigoplus_{k=0}^{\infty} \mathcal{H}^C_k(S^k) \), then \( f_k \) is given by

\[
f_k(\sigma) = \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} f(\tau) P_{k,n}(\sigma \cdot \tau) \, \text{Vol}_{S^n},
\]

for all \( \sigma \in S^n \).

**Proof.** If we recall that \( \mathcal{H}^C_k(S^k) \) and \( \mathcal{H}^C_j(S^k) \) are orthogonal for all \( j \neq k \), using the formula (rk), we have

\[
\frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} f(\tau) P_{k,n}(\sigma \cdot \tau) \, \text{Vol}_{S^n} = \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} \sum_{j=0}^{\infty} f_j(\tau) P_{k,n}(\sigma \cdot \tau) \, \text{Vol}_{S^n}
\]

\[
= \frac{a_{k,n+1}}{\text{vol}(S^n)} \sum_{j=0}^{\infty} \int_{S^n} f_j(\tau) P_{k,n}(\sigma \cdot \tau) \, \text{Vol}_{S^n}
\]

\[
= \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} f_k(\tau) P_{k,n}(\sigma \cdot \tau) \, \text{Vol}_{S^n}
\]

\[
= f_k(\sigma),
\]

as claimed. \( \square \)

Another important property of the zonal spherical functions \( Z_k^l \) is that they generate \( \mathcal{H}^C_k(S^n) \). In order to prove this fact, we use the following proposition:

**Proposition 26.25.** If \( H_1, \ldots, H_m \in \mathcal{H}^C_k(S^n) \) are linearly independent, then there are \( m \) points \( \sigma_1, \ldots, \sigma_m \) on \( S^n \) so that the \( m \times m \) matrix \( (H_j(\sigma_i)) \) is invertible.
Proof. We proceed by induction on \(m\). The case \(m = 1\) is trivial. For the induction step, we may assume that we found \(m\) points \(\sigma_1, \ldots, \sigma_m\) on \(S^n\) so that the \(m \times m\) matrix \((H_j(\sigma_i))\) is invertible. Consider the function

\[
\sigma \mapsto \left| \begin{array}{ccc}
H_1(\sigma) & \ldots & H_m(\sigma) & H_{m+1}(\sigma) \\
H_1(\sigma_1) & \ldots & H_m(\sigma_1) & H_{m+1}(\sigma_1) \\
\vdots & \ddots & \vdots & \vdots \\
H_1(\sigma_m) & \ldots & H_m(\sigma_m) & H_{m+1}(\sigma_m)
\end{array} \right|.
\]

Since \(H_1, \ldots, H_{m+1}\) are linearly independent, the above function does not vanish for all \(\sigma\), since otherwise, by expanding this determinant with respect to the first row, we would get a linear dependence among the \(H_j\)'s where the coefficient of \(H_{m+1}\) is nonzero. Therefore, we can find \(\sigma_{m+1}\) so that the \((m+1) \times (m+1)\) matrix \((H_j(\sigma_i))\) is invertible. \(\square\)

**Definition 26.14.** We say that \(a_{k,n+1}\) points, \(\sigma_1, \ldots, \sigma_{a_{k,n+1}}\) on \(S^n\) form a fundamental system iff the \(a_{k,n+1} \times a_{k,n+1}\) matrix \((P_{k,n}(\sigma_i \cdot \sigma_j))\) is invertible.

**Theorem 26.26.** The following properties hold:

(i) There is a fundamental system \(\sigma_1, \ldots, \sigma_{a_{k,n+1}}\) for every \(k \geq 1\).

(ii) Every spherical harmonic \(H \in \mathcal{H}_k^C(S^n)\) can be written as

\[
H(\sigma) = \sum_{j=1}^{a_{k,n+1}} c_j P_{k,n}(\sigma_j \cdot \sigma),
\]

for some unique \(c_j \in \mathbb{C}\).

**Proof.** (i) By the addition formula,

\[
P_{k,n}(\sigma_i \cdot \sigma_j) = \frac{\text{vol}(S^n)}{a_{k,n+1}} \sum_{l=1}^{a_{k,n+1}} Y^i_k(\sigma_i)Y^j_k(\sigma_j)
\]

for any orthonormal basis \((Y^1_k, \ldots, Y^{a_{k,n+1}}_k)\). It follows that the matrix \((P_{k,n}(\sigma_i \cdot \sigma_j))\) can be written as

\[
(P_{k,n}(\sigma_i \cdot \sigma_j)) = \frac{\text{vol}(S^n)}{a_{k,n+1}} YY^*,
\]

where \(Y = (Y^i_k(\sigma_i))\), and by Proposition 26.25, we can find \(\sigma_1, \ldots, \sigma_{a_{k,n+1}} \in S^n\) so that \(Y\) and thus also \(Y^*\) are invertible, and so \((P_{n,k}(\sigma_i \cdot \sigma_j))\) is invertible.

(ii) Again, by the addition formula,

\[
P_{k,n}(\sigma \cdot \sigma_j) = \frac{\text{vol}(S^n)}{a_{k,n+1}} \sum_{i=1}^{a_{k,n+1}} Y^i_k(\sigma)Y^j_k(\sigma_j).
\]

However, as \((Y^1_k, \ldots, Y^{a_{k,n+1}}_k)\) is an orthonormal basis, (i) proved that the matrix \(Y^*\) is invertible, so the \(Y^i_k(\sigma)\) can be expressed uniquely in terms of the \(P_{k,n}(\sigma \cdot \sigma_j)\), as claimed. \(\square\)
Statement (ii) of Theorem 26.26 shows that the set of \( P_{k,n}(\sigma \cdot \tau) = \frac{\text{vol}(S^n)}{a_{k,n+1}} F_k(\sigma, \tau) \) do indeed generate \( \mathcal{H}_k^C(S^n) \).

We end this section with a neat geometric characterization of the zonal spherical functions is given in Stein and Weiss [165]. For this, we need to define the notion of a parallel on \( S^n \). A parallel of \( S^n \) orthogonal to a point \( \tau \in S^n \) is the intersection of \( S^n \) with any (affine) hyperplane orthogonal to the line through the center of \( S^n \) and \( \tau \). Clearly, any rotation \( R \) leaving \( \tau \) fixed leaves every parallel orthogonal to \( \tau \) globally invariant, and for any two points \( \sigma_1 \) and \( \sigma_2 \), on such a parallel, there is a rotation leaving \( \tau \) fixed that maps \( \sigma_1 \) to \( \sigma_2 \). Consequently, the zonal function \( Z^\tau_k \) defined by \( \tau \) is constant on the parallels orthogonal to \( \tau \). In fact, this property characterizes zonal harmonics, up to a constant.

The theorem below is proved in Stein and Weiss [165] (Chapter 4, Theorem 2.12). The proof uses Proposition 26.17 and it is very similar to the proof of Theorem 26.18. To save space, it is omitted.

**Theorem 26.27.** Fix any point \( \tau \in S^n \). A spherical harmonic \( Y \in \mathcal{H}_k^C(S^n) \) is constant on parallels orthogonal to \( \tau \) iff \( Y = cZ^\tau_k \) for some constant \( c \in \mathbb{C} \).

In the next section we show how the Gegenbauer polynomials can actually be computed.

## 26.7 More on the Gegenbauer Polynomials

The Gegenbauer polynomials are characterized by a formula generalizing the Rodrigues formula defining the Legendre polynomials (see Section 26.3). The expression

\[
\left( k + \frac{n-2}{2} \right) \left( k - 1 + \frac{n-2}{2} \right) \cdots \left( 1 + \frac{n-2}{2} \right)
\]

can be expressed in terms of the \( \Gamma \) function as

\[
\frac{\Gamma \left( k + \frac{n}{2} \right)}{\Gamma \left( \frac{n}{2} \right)}.
\]

Recall that the \( \Gamma \) function is a generalization of factorial that satisfies the equation

\[
\Gamma(z + 1) = z\Gamma(z).
\]

For \( z = x + iy \) with \( x > 0 \), \( \Gamma(z) \) is given by

\[
\Gamma(z) = \int_0^\infty t^{z-1}e^{-t} dt,
\]

where the integral converges absolutely. If \( n \) is an integer \( n \geq 0 \), then \( \Gamma(n+1) = n! \).

It is proved in Morimoto [132] (Chapter 2, Theorem 2.35) that
Proposition 26.28. The Gegenbauer polynomial \( P_{k,n} \) is given by Rodrigues’ formula:
\[
P_{k,n}(t) = \frac{(-1)^k}{2^k} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(k + \frac{n}{2}\right)} \frac{1}{(1 - t^2)^{n/2}} \frac{d^k}{dt^k} (1 - t^2)^{k+n/2},
\]
with \( n \geq 2 \).

The Gegenbauer polynomials satisfy the following orthogonality properties with respect to the kernel \((1 - t^2)^{n/2}\) (see Morimoto [132] (Chapter 2, Theorem 2.34):

Proposition 26.29. The Gegenbauer polynomial \( P_{k,n} \) have the following properties:
\[
\int_{-1}^{-1} (P_{k,n}(t))^2 (1 - t^2)^{n/2} dt = \frac{\text{vol}(S^n)}{a_{k,n+1} \text{vol}(S^{n-1})}
\]
\[
\int_{-1}^{-1} P_{k,n}(t) P_{l,n}(t) (1 - t^2)^{n/2} dt = 0, \quad k \neq l.
\]

The Gegenbauer polynomials satisfy a second-order differential equation generalizing the Legendre equation from Section 26.3.

Proposition 26.30. The Gegenbauer polynomial \( P_{k,n} \) are solutions of the differential equation
\[
(1 - t^2)P_{k,n}''(t) - nt P_{k,n}'(t) + k(k + n - 1) P_{k,n}(t) = 0.
\]

Proof. If we let \( \tau = e_{n+1} \), then the function \( H \) given by \( H(\sigma) = P_{k,n}(\sigma \cdot \tau) = P_{k,n}(\cos \theta) \) belongs to \( H_k^C(S^n) \), so
\[
\Delta_{S^n} H = -k(k + n - 1) H.
\]

Recall from Section 26.4 that
\[
\Delta_{S^n} f = \frac{1}{\sin^{n-1} \theta} \frac{\partial}{\partial \theta} \left( \sin^{n-1} \theta \frac{\partial f}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \Delta_{S^{n-1}} f,
\]
in the local coordinates where
\[
\sigma = \sin \theta \tilde{\sigma} + \cos \theta e_{n+1},
\]
with \( \tilde{\sigma} \in S^{n-1} \) and \( 0 \leq \theta < \pi \). If we make the change of variable \( t = \cos \theta \), then it is easy to see that the above formula becomes
\[
\Delta_{S^n} f = (1 - t^2) \frac{\partial^2 f}{\partial t^2} - nt \frac{\partial f}{\partial t} + \frac{1}{1 - t^2} \Delta_{S^{n-1}}
\]
(see Morimoto [132], Chapter 2, Theorem 2.9.) But, \( H \) being zonal, it only depends on \( \theta \), that is on \( t \), so \( \Delta_{S^{n-1}} H = 0 \), and thus
\[
-k(k + n - 1) P_{k,n}(t) = \Delta_{S^n} P_{k,n}(t) = (1 - t^2) \frac{\partial^2 P_{k,n}}{\partial t^2} - nt \frac{\partial P_{k,n}}{\partial t},
\]
which yields our equation. \( \square \)
Note that for \( n = 2 \), the differential equation of Proposition 26.30 is the Legendre equation from Section 26.3.

The Gegenbauer polynomials also appear as coefficients in some simple generating functions. The following proposition is proved in Morimoto [132] (Chapter 2, Theorem 2.53 and Theorem 2.55):

**Proposition 26.31.** For all \( r \) and \( t \) such that \(-1 < r < 1 \) and \(-1 \leq t \leq 1\), for all \( n \geq 1 \), we have the following generating formula:

\[
\sum_{k=0}^{\infty} a_{k,n+1} r^k P_{k,n}(t) = \frac{1 - r^2}{(1 - 2rt + r^2)^{n+1/2}}.
\]

Furthermore, for all \( r \) and \( t \) such that \(0 \leq r < 1 \) and \(-1 \leq t \leq 1\), if \( n = 1 \), then

\[
\sum_{k=1}^{\infty} \frac{r^k}{k} P_{k,1}(t) = -\frac{1}{2} \log(1 - 2rt + r^2),
\]

and if \( n \geq 2 \), then

\[
\sum_{k=0}^{\infty} \frac{n-1}{2k+n-1} a_{k,n+1} r^k P_{k,n}(t) = \frac{1}{(1 - 2rt + r^2)^{n-1/2}}.
\]

In Stein and Weiss [165] (Chapter 4, Section 2), the polynomials \( P^\lambda_k(t) \), where \( \lambda > 0 \), are defined using the following generating formula:

\[
\sum_{k=0}^{\infty} r^k P^\lambda_{k}(t) = \frac{1}{(1 - 2rt + r^2)^{\lambda}}.
\]

Each polynomial \( P^\lambda_k(t) \) has degree \( k \) and is called an *ultraspherical polynomial of degree \( k \) associated with \( \lambda \). In view of Proposition 26.31, we see that

\[
P_k^{\frac{n-1}{2}}(t) = \frac{n-1}{2k+n-1} a_{k,n+1} P_{k,n}(t),
\]

as we mentioned earlier. There is also an integral formula for the Gegenbauer polynomials known as *Laplace representation*; see Morimoto [132] (Chapter 2, Theorem 2.52).

### 26.8 The Funk-Hecke Formula

The Funk-Hecke Formula (also known as Hecke-Funk Formula) basically allows one to perform a sort of convolution of a "kernel function" with a spherical function in a convenient way. Given a measurable function \( K \) on \([-1, 1]\) such that the integral

\[
\int_{-1}^{1} |K(t)|(1 - t^2)^{\frac{n-2}{2}} \, dt
\]
makes sense, given a function \( f \in L^2_\mathbb{C}(S^n) \), we can view the expression

\[
K \star f(\sigma) = \int_{S^n} K(\sigma \cdot \tau) f(\tau) \text{Vol}_{S^n}
\]

as a sort of convolution of \( K \) and \( f \).

Actually, the use of the term convolution is really unfortunate because in a “true” convolution \( f \ast g \), either the argument of \( f \) or the argument of \( g \) should be multiplied by the inverse of the variable of integration, which means that the integration should really be taking place over the group \( \text{SO}(n+1) \). We will come back to this point later. For the time being, let us call the expression \( K \star f \) defined above a pseudo-convolution. Now, if \( f \) is expressed in terms of spherical harmonics as

\[
f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_{k,n+1}} c_{k,m_k} Y_k^{m_k},
\]

then the Funk-Hecke Formula states that

\[
K \star Y_k^{m_k}(\sigma) = \alpha_k Y_k^{m_k}(\sigma),
\]

for some fixed constant \( \alpha_k \), and so

\[
K \star f = \sum_{k=0}^{\infty} \sum_{m_k=1}^{a_k,n+1} \alpha_k c_{k,m_k} Y_k^{m_k}.
\]

Thus, if the constants \( \alpha_k \) are known, then it is “cheap” to compute the pseudo-convolution \( K \star f \).

This method was used in a ground-breaking paper by Basri and Jacobs [17] to compute the reflectance function \( r \) from the lighting function \( \ell \) as a pseudo-convolution \( K \star \ell \) (over \( S^2 \)) with the Lambertian kernel \( K \) given by

\[
K(\sigma \cdot \tau) = \max(\sigma \cdot \tau, 0).
\]

Below, we give a proof of the Funk-Hecke formula due to Morimoto [132] (Chapter 2, Theorem 2.39); see also Andrews, Askey and Roy [3] (Chapter 9). This formula was first published by Funk in 1916 and then by Hecke in 1918.

**Theorem 26.32. (Funk-Hecke Formula)** Given any measurable function \( K \) on \([-1,1]\) such that the integral

\[
\int_{-1}^{1} |K(t)|(1-t^2)^{\frac{n-2}{2}} \, dt
\]

makes sense, for every function \( H \in \mathcal{H}^\mathbb{C}_k(S^n) \), we have

\[
\int_{S^n} K(\sigma \cdot \xi) H(\xi) \text{Vol}_{S^n} = \left( \text{vol}(S_{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t)(1-t^2)^{\frac{n-2}{2}} \, dt \right) H(\sigma).
\]
Observe that when \( n = 2 \), the term \((1 - t^2)^{\frac{n-2}{2}}\) is missing and we are simply requiring that \( \int_{-1}^{1} |K(t)|\,dt \) makes sense.

**Proof.** We first prove the formula in the case where \( H \) is a zonal harmonic, and then use the fact that the \( P_{k,n} \)'s are reproducing kernels (formula \((rk)\)).

For all \( \sigma, \tau \in S^n \), define \( H \) by

\[
H(\sigma) = P_{k,n}(\sigma \cdot \tau),
\]

and \( F \) by

\[
F(\sigma, \tau) = \int_{S^n} K(\sigma \cdot \xi) P_{k,n}(\xi \cdot \tau) \text{Vol}_{S^n}.
\]

Since the volume form on the sphere is invariant under orientation-preserving isometries, for every \( R \in \text{SO}(n + 1) \), we have

\[
F(R\sigma, R\tau) = F(\sigma, \tau).
\]

On the other hand, for \( \sigma \) fixed, it is not hard to see that as a function in \( \tau \), the function \( F(\sigma, -) \) is a spherical harmonic, because \( P_{k,n} \) satisfies a differential equation that implies that \( \Delta_{S^2} F(\sigma, -) = -k(k + n - 1) F(\sigma, -) \). Now, for every rotation \( R \) that fixes \( \sigma \),

\[
F(\sigma, \tau) = F(R\sigma, R\tau) = F(\sigma, R\tau),
\]

which means that \( F(\sigma, -) \) satisfies condition (2) of Theorem 26.18. By Theorem 26.18, we get

\[
F(\sigma, \tau) = F(\sigma, \sigma) P_{k,n}(\sigma \cdot \tau).
\]

If we use local coordinates on \( S^n \) where

\[
\sigma = \sqrt{1 - t^2} \tilde{\sigma} + t e_{n+1},
\]

with \( \tilde{\sigma} \in S^{n-1} \) and \(-1 \leq t \leq 1\), it is not hard to show that the volume form on \( S^n \) is given by

\[
\text{Vol}_{S^n} = (1 - t^2)^{\frac{n-2}{2}} \, dt \text{Vol}_{S^{n-1}}.
\]

Using this, we have

\[
F(\sigma, \sigma) = \int_{S^n} K(\sigma \cdot \xi) P_{k,n}(\xi \cdot \sigma) \text{Vol}_{S^n} = \text{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t)(1 - t^2)^{\frac{n-2}{2}} \, dt,
\]

and thus,

\[
F(\sigma, \tau) = \left( \text{vol}(S^{n-1}) \int_{-1}^{1} K(t) P_{k,n}(t)(1 - t^2)^{\frac{n-2}{2}} \, dt \right) P_{k,n}(\sigma \cdot \tau),
\]

which is the Funk-Hecke formula when \( H(\sigma) = P_{k,n}(\sigma \cdot \tau) \).
Let us now consider any function \( H \in \mathcal{H}_C(S^n) \). Recall that by the reproducing kernel property (rk), we have

\[
\frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} P_{k,n}(\xi \cdot \tau) H(\tau) \text{Vol}_{S^n} = H(\xi).
\]

Then, we can compute \( \int_{S^n} K(\sigma \cdot \xi) H(\xi) \text{Vol}_{S^n} \) using Fubini’s Theorem and the Funk-Hecke formula in the special case where \( H(\sigma) = P_{k,n}(\sigma \cdot \tau) \), as follows:

\[
\int_{S^n} K(\sigma \cdot \xi) H(\xi) \text{Vol}_{S^n} = \int_{S^n} \left( \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} P_{k,n}(\xi \cdot \tau) H(\tau) \text{Vol}_{S^n} \right) \text{Vol}_{S^n}
\]

\[
= \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} H(\tau) \left( \int_{S^n} K(\sigma \cdot \xi) P_{k,n}(\xi \cdot \tau) \text{Vol}_{S^n} \right) \text{Vol}_{S^n}
\]

\[
= \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} H(\tau) \left( \text{vol}(S^{n-1}) \int_{-1}^1 K(t) P_{k,n}(t)(1 - t^2)^{\frac{n-2}{2}} dt \right) P_{k,n}(\sigma \cdot \tau) \text{Vol}_{S^n}
\]

\[
= \left( \text{vol}(S^{n-1}) \int_{-1}^1 K(t) P_{k,n}(t)(1 - t^2)^{\frac{n-2}{2}} dt \right) \left( \frac{a_{k,n+1}}{\text{vol}(S^n)} \int_{S^n} P_{k,n}(\sigma \cdot \tau) \text{Vol}_{S^n} \right)
\]

\[
= \left( \text{vol}(S^{n-1}) \int_{-1}^1 K(t) P_{k,n}(t)(1 - t^2)^{\frac{n-2}{2}} dt \right) H(\sigma),
\]

which proves the Funk-Hecke formula in general.

The Funk-Hecke formula can be used to derive an “addition theorem” for the ultraspherical polynomials (Gegenbauer polynomials). We omit this topic and we refer the interested reader to Andrews, Askey and Roy [3] (Chapter 9, Section 9.8).

Remark: Oddly, in their computation of \( K \ast \ell \), Basri and Jacobs [17] first expand \( K \) in terms of spherical harmonics as

\[
K = \sum_{n=0}^{\infty} k_n Y_n^0,
\]

and then use the Funk-Hecke formula to compute \( K \ast Y_n^m \). They get (see page 222)

\[
K \ast Y_n^m = \alpha_n Y_n^m, \quad \text{with} \quad \alpha_n = \sqrt{\frac{4\pi}{2n+1}} k_n,
\]

for some constant \( k_n \) given on page 230 of their paper (see below). However, there is no need to expand \( K \), as the Funk-Hecke formula yields directly

\[
K \ast Y_n^m(\sigma) = \int_{S^2} K(\sigma \cdot \xi) Y_n^m(\xi) \text{Vol}_{S^n} = \left( \int_{-1}^1 K(t) P_n(t) dt \right) Y_n^m(\sigma),
\]

\[
K \ast Y_n^m(\sigma) = \cos(\theta) \sum_{m=-n}^{n} \frac{4\pi}{2n+1} k_n Y_n^m(\sigma)
\]

for some constant \( k_n \) given on page 230 of their paper (see below). However, there is no need to expand \( K \), as the Funk-Hecke formula yields directly

\[
K \ast Y_n^m(\sigma) = \int_{S^2} K(\sigma \cdot \xi) Y_n^m(\xi) \text{Vol}_{S^n} = \left( \int_{-1}^1 K(t) P_n(t) dt \right) Y_n^m(\sigma),
\]

\[
K \ast Y_n^m(\sigma) = \cos(\theta) \sum_{m=-n}^{n} \frac{4\pi}{2n+1} k_n Y_n^m(\sigma)
\]
where \( P_n(t) \) is the standard Legendre polynomial of degree \( n \), since we are in the case of \( S^2 \). By the definition of \( K \) (\( K(t) = \max(t, 0) \)) and since \( \text{vol}(S^1) = 2\pi \), we get

\[
K \star Y^m_n = \left( 2\pi \int_0^1 tP_n(t) \, dt \right) Y^m_n,
\]

which is equivalent to Basri and Jacobs’ formula (14), since their \( \alpha_n \) on page 222 is given by

\[
\alpha_n = \sqrt{\frac{4\pi}{2n+1} k_n},
\]

but from page 230,

\[
k_n = \sqrt{(2n+1)\pi} \int_0^1 tP_n(t) \, dt.
\]

What remains to be done is to compute \( \int_0^1 tP_n(t) \, dt \), which is done by using the Rodrigues Formula and integrating by parts (see Appendix A of Basri and Jacobs [17]).

In the next section, we try to show how spherical harmonics fit into the broader framework of linear representations of (Lie) groups.

### 26.9 Linear Representations of Compact Lie Groups; A Glimpse

In this section, we indicate briefly how Theorem 26.16 (except part (3)) can be viewed as a special case of a famous theorem known as the Peter–Weyl Theorem about unitary representations of compact Lie groups (Herman, Klauss, Hugo Weyl, 1885-1955). First, we review the notion of a linear representation of a group. A good and easy-going introduction to representations of Lie groups can be found in Hall [84]. We begin with finite-dimensional representations.

**Definition 26.15.** Given a Lie group \( G \) and a vector space \( V \) of dimension \( n \), a linear representation of \( G \) of dimension (or degree) \( n \) is a group homomorphism \( U : G \to \text{GL}(V) \) such that the map \( g \mapsto U(g)(u) \) is continuous for every \( u \in V \), where \( \text{GL}(V) \) denotes the group of invertible linear maps from \( V \) to itself. The space \( V \), called the representation space, may be a real or a complex vector space. If \( V \) has a Hermitian (resp Euclidean) inner product \( \langle - , - \rangle \), we say that \( U : G \to \text{GL}(V) \) is a unitary representation iff

\[
\langle U(g)(u), U(g)(v) \rangle = \langle u, v \rangle, \quad \text{for all } g \in G \text{ and all } u, v \in V.
\]

Thus, a linear representation of \( G \) is a map \( U : G \to \text{GL}(V) \) satisfying the properties:

\[
U(gh) = U(g)U(h), \quad U(g^{-1}) = U(g)^{-1}, \quad U(1) = I.
\]
For simplicity of language, we usually abbreviate \textit{linear representation} as \textit{representation}. The representation space $V$ is also called a \textit{$G$-module}, since the representation $U: G \to \text{GL}(V)$ is equivalent to the left action $\cdot: G \times V \to V$, with $g \cdot v = U(g)(v)$. The representation such that $U(g) = I$ for all $g \in G$ is called the \textit{trivial representation}.

As an example, we describe a class of representations of $\text{SL}(2, \mathbb{C})$, the group of complex matrices with determinant $+1$, $egin{pmatrix} a & b \\ c & d \end{pmatrix}$, $ad - bc = 1$.

Recall that $\mathcal{P}_k^C(2)$ denotes the vector space of complex homogeneous polynomials of degree $k$ in two variables $(z_1, z_2)$. For every matrix $A \in \text{SL}(2, \mathbb{C})$, with $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, for every homogeneous polynomial $Q \in \mathcal{P}_k^C(2)$, we define $U_k(A)(Q(z_1, z_2))$ by

$$U_k(A)(Q(z_1, z_2)) = Q(dz_1 - bz_2, -cz_1 + az_2).$$

If we think of the homogeneous polynomial $Q(z_1, z_2)$ as a function $Q(z_1, z_2)$ of the vector $(z_1, z_2)$, then

$$U_k(A) \left( Q \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right) = QA^{-1} \left( \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \right) = Q \begin{pmatrix} b & -d \\ c & -a \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$ 

The expression above makes it clear that

$$U_k(AB) = U_k(A)U_k(B)$$

for any two matrices $A, B \in \text{SL}(2, \mathbb{C})$, so $U_k$ is indeed a representation of $\text{SL}(2, \mathbb{C})$ into $\mathcal{P}_k^C(2)$.

One might wonder why we considered $\text{SL}(2, \mathbb{C})$ rather than $\text{SL}(2, \mathbb{R})$. This is because it can be shown that $\text{SL}(2, \mathbb{R})$ has \textit{no} nontrivial unitary (finite-dimensional) representations! For more on representations of $\text{SL}(2, \mathbb{R})$, see Dieudonné [51] (Chapter 14).

Given any basis $(e_1, \ldots, e_n)$ of $V$, each $U(g)$ is represented by an $n \times n$ matrix $U(g) = (U_{ij}(g))$. We may think of the scalar functions $g \mapsto U_{ij}(g)$ as \textit{special functions} on $G$. As explained in Dieudonné [51] (see also Vilenkin [173]), essentially all special functions (Legendre polynomials, ultraspherical polynomials, Bessel functions \textit{etc.}) arise in this way by choosing some suitable $G$ and $V$. There is a natural and useful notion of equivalence of representations:

\textbf{Definition 26.16.} Given any two representations $U_1: G \to \text{GL}(V_1)$ and $U_2: G \to \text{GL}(V_2)$, a \textit{$G$-map} (or \textit{morphism of representations}) $\varphi: U_1 \to U_2$ is a linear map $\varphi: V_1 \to V_2$ so that
the following diagram commutes for every \( g \in G \):\[
\begin{array}{ccc}
V_1 & \xrightarrow{U_1(g)} & V_1 \\
\downarrow{\varphi} & & \downarrow{\varphi} \\
V_2 & \xrightarrow{U_2(g)} & V_2
\end{array}
\]

The space of all \( G \)-maps between two representations as above is denoted \( \text{Hom}_G(U_1, U_2) \). Two representations \( U_1: G \to \text{GL}(V_1) \) and \( U_2: G \to \text{GL}(V_2) \) are equivalent iff \( \varphi: V_1 \to V_2 \) is an invertible linear map (which implies that \( \dim V_1 = \dim V_2 \)). In terms of matrices, the representations \( U_1: G \to \text{GL}(V_1) \) and \( U_2: G \to \text{GL}(V_2) \) are equivalent iff there is some invertible \( n \times n \) matrix, \( P \), so that \( U_2(g) = PU_1(g)P^{-1}, \quad g \in G \).

If \( W \subseteq V \) is a subspace of \( V \), then in some cases, a representation \( U: G \to \text{GL}(V) \) yields a representation \( U: G \to \text{GL}(W) \). This is interesting because under certain conditions on \( G \) (e.g., \( G \) compact) every representation may be decomposed into a “sum” of so-called irreducible representations, and thus the study of all representations of \( G \) boils down to the study of irreducible representations of \( G \); for instance, see Knapp [106] (Chapter 4, Corollary 4.7), or Bröcker and tom Dieck [31] (Chapter 2, Proposition 1.9).

**Definition 26.17.** Let \( U: G \to \text{GL}(V) \) be a representation of \( G \). If \( W \subseteq V \) is a subspace of \( V \), then we say that \( W \) is invariant (or stable) under \( U \) iff \( U(g)(w) \in W \), for all \( g \in G \) and all \( w \in W \). If \( W \) is invariant under \( U \), then we have a homomorphism, \( U: G \to \text{GL}(W) \), called a subrepresentation of \( G \). A representation \( U: G \to \text{GL}(V) \) with \( V \neq (0) \) is irreducible iff it only has the two subrepresentations \( U: G \to \text{GL}(W) \) corresponding to \( W = (0) \) or \( W = V \).

It can be shown that the representations \( U_k \) of \( \text{SL}(2, \mathbb{C}) \) defined earlier are irreducible, and that every representation of \( \text{SL}(2, \mathbb{C}) \) is equivalent to one of the \( U_k \)’s (see Bröcker and tom Dieck [31], Chapter 2, Section 5). The representations \( U_k \) are also representations of \( \text{SU}(2) \). Again, they are irreducible representations of \( \text{SU}(2) \), and they constitute all of them (up to equivalence). The reader should consult Hall [84] for more examples of representations of Lie groups.

An easy but crucial lemma about irreducible representations is “Schur’s Lemma.”

**Lemma 26.33. (Schur’s Lemma)** Let \( U_1: G \to \text{GL}(V) \) and \( U_2: G \to \text{GL}(W) \) be any two real or complex representations of a group \( G \). If \( U_1 \) and \( U_2 \) are irreducible, then the following properties hold:

(i) Every \( G \)-map \( \varphi: U_1 \to U_2 \) is either the zero map or an isomorphism.

(ii) If \( U_1 \) is a complex representation, then every \( G \)-map \( \varphi: U_1 \to U_1 \) is of the form \( \varphi = \lambda \text{id} \), for some \( \lambda \in \mathbb{C} \).
Proof. (i) Observe that the kernel $\ker \varphi \subseteq V$ of $\varphi$ is invariant under $U_1$. Indeed, for every $v \in \ker \varphi$ and every $g \in G$, we have

$$\varphi(U_1(g)(v)) = U_2(g)(\varphi(v)) = U_2(g)(0) = 0,$$

so $U_1(g)(v) \in \ker \varphi$. Thus, $U_1 : G \to \text{GL}(\ker \varphi)$ is a subrepresentation of $U_1$, and as $U_1$ is irreducible, either $\ker \varphi = (0)$ or $\ker \varphi = V$. In the second case, $\varphi = 0$. If $\ker \varphi = (0)$, then $\varphi$ is injective. However, $\varphi(V) \subseteq W$ is invariant under $U_2$, since for every $v \in V$ and every $g \in G$,

$$U_2(g)(\varphi(v)) = \varphi(U_1(g)(v)) \in \varphi(V),$$

and as $\varphi(V) \neq (0)$ (as $V \neq (0)$ since $U_1$ is irreducible) and $U_2$ is irreducible, we must have $\varphi(V) = W$; that is, $\varphi$ is an isomorphism.

(ii) Since $V$ is a complex vector space, the linear map $\varphi$ has some eigenvalue $\lambda \in \mathbb{C}$. Let $E_\lambda \subseteq V$ be the eigenspace associated with $\lambda$. The subspace $E_\lambda$ is invariant under $U_1$, since for every $u \in E_\lambda$ and every $g \in G$, we have

$$\varphi(U_1(g)(u)) = U_1(g)(\varphi(u)) = U_1(g)(\lambda u) = \lambda U_1(g)(u),$$

so $U_1 : G \to \text{GL}(E_\lambda)$ is a subrepresentation of $U_1$, and as $U_1$ is irreducible and $E_\lambda \neq (0)$, we must have $E_\lambda = V$. \hfill \square

An interesting corollary of Schur’s Lemma is that every complex irreducible representation of a commutative group is one-dimensional.

Let us now restrict our attention to compact Lie groups. If $G$ is a compact Lie group, then it is known that it has a left and right-invariant volume form $\omega_G$, so we can define the integral of a (real or complex) continuous function $f$ defined on $G$ by

$$\int_G f = \int_G f \omega_G,$$

also denoted $\int_G f \, d\mu_G$ or simply $\int_G f(t) \, dt$, with $\omega_G$ normalized so that $\int_G \omega_G = 1$. (See Section 24.7, or Knapp [106], Chapter 8, or Warner [175], Chapters 4 and 6.) Because $G$ is compact, the Haar measure $\mu_G$ induced by $\omega_G$ is both left and right-invariant ($G$ is a unimodular group), and our integral has the following invariance properties:

$$\int_G f(t) \, dt = \int_G f(st) \, dt = \int_G f(tu) \, dt = \int_G f(t^{-1}) \, dt,$$

for all $s, u \in G$ (see Section 24.7).

Since $G$ is a compact Lie group, we can use an “averaging trick” to show that every (finite-dimensional) representation is equivalent to a unitary representation (see Bröcker and tom Dieck [31] (Chapter 2, Theorem 1.7) or Knapp [106] (Chapter 4, Proposition 4.6).
If we define the Hermitian inner product

$$\langle f, g \rangle = \int_G f \overline{g} \omega_G,$$

then, with this inner product the space of square-integrable functions $L^2_\mathbb{C}(G)$ is a Hilbert space (in fact, a separable Hilbert space).

We can also define the convolution $f \ast g$ of two functions $f, g \in L^2_\mathbb{C}(G)$, by

$$(f \ast g)(x) = \int_G f(xt^{-1})g(t)dt = \int_G f(t)g(t^{-1}x)dt.$$ 

In general, $f \ast g \neq g \ast f$, unless $G$ is commutative. With the convolution product, $L^2_\mathbb{C}(G)$ becomes an associative algebra (non-commutative in general).

This leads us to consider unitary representations of $G$ into the infinite-dimensional vector space $L^2_\mathbb{C}(G)$, and more generally into a Hilbert space $E$.

Given a Hilbert space $E$, the definition of a unitary representation $\rho: G \to \text{Aut}(E)$ is the same as in Definition 26.15, except that $\text{GL}(E)$ is the group of automorphisms (unitary operators) $\text{Aut}(E)$ of the Hilbert space $E$, and

$$\langle \rho(g)(u), \rho(g)(v) \rangle = \langle u, v \rangle$$

with respect to the inner product on $E$. Also, in the definition of an irreducible representation $U: G \to V$, we require that the only closed subrepresentations $U: G \to W$ of the representation $U: G \to V$ correspond to $W = (0)$ or $W = V$.

The Peter–Weyl Theorem gives a decomposition of $L^2_\mathbb{C}(G)$ as a Hilbert sum of spaces that correspond to irreducible unitary representations of $G$. We present a version of the Peter–Weyl Theorem found in Dieudonné [51] (Chapters 3-8) and Dieudonné [52] (Chapter XXI, Sections 1-4), which contains complete proofs. Other versions can be found in Bröcker and tom Dieck [31] (Chapter 3), Knapp [106] (Chapter 4) or Duistermaat and Kolk [64] (Chapter 4). A good preparation for these fairly advanced books is Deitmar [48].

**Theorem 26.34. (Peter–Weyl (1927))** Given a compact Lie group $G$, there is a decomposition of $L^2_\mathbb{C}(G)$ as a Hilbert sum

$$L^2_\mathbb{C}(G) = \bigoplus_{\rho \in R(G)} a_\rho$$

of countably many two-sided ideals $a_\rho$, where each $a_\rho$ is isomorphic to a finite-dimensional algebra of $n_\rho \times n_\rho$ complex matrices, where the set of indices $R(G)$ corresponds to the set of equivalence classes of (finite-dimensional) irreducible representations of $G$. More precisely, for each $\rho \in R(G)$, there is a basis of $a_\rho$ consisting of $n_\rho^2$ pairwise orthogonal continuous functions $m_{ij}^{(\rho)}$ satisfying various properties, including

$$\langle m_{ij}^{(\rho)}, m_{ij}^{(\rho)} \rangle = n_\rho,$$
and if we form the $n_\rho \times n_\rho$ matrix $M_\rho(g)$ given by

$$M_\rho(g) = \frac{1}{n_\rho} \begin{pmatrix} m_{11}^{(\rho)}(g) & \ldots & m_{1n}^{(\rho)}(g) \\ \vdots & \ddots & \vdots \\ m_{n1}^{(\rho)}(g) & \ldots & m_{nn}^{(\rho)}(g) \end{pmatrix},$$

then the map $g \mapsto M_\rho(g)$ is an \textbf{irreducible unitary representation} of $G$ in the vector space $\mathbb{C}^{n_\rho}$. Furthermore, every (finite dimensional) irreducible representation of $G$ is equivalent to some $M_\rho$. The function $u_\rho$ given by

$$u_\rho(g) = \sum_{j=1}^{n_\rho} m_{jj}^{(\rho)}(g) = n_\rho \text{tr}(M_\rho(g))$$

is the unit of the algebra $a_\rho$, and the orthogonal projection of $L^2_G(\mathbb{C})$ onto $a_\rho$ is the map

$$f \mapsto u_\rho * f;$$

that is, convolution with $u_\rho$.

The function $\chi_\rho = \frac{1}{n_\rho} u_\rho = \text{tr}(M_\rho)$ is the \textbf{character} of $G$ associated with the representation $M_\rho$. The functions $\chi_\rho$ satisfy the following properties:

$$\begin{align*}
\chi_\rho(e) &= n_\rho \\
\chi_\rho(sts^{-1}) &= \chi_\rho(t) \quad \text{for all } s,t \in G \\
\chi_\rho(s^{-1}) &= \overline{\chi_\rho(s)} \quad \text{for all } s \in G \\
\chi_\rho \ast \chi_{\rho'} &= 0 \quad \text{if } \rho \neq \rho' \\
\chi_\rho \ast \chi_\rho &= \frac{1}{n_\rho}.
\end{align*}$$

Furthermore, the characters form an orthonormal Hilbert basis of the Hilbert subspace of $L^2_G(\mathbb{C})$ consisting of the \textbf{central functions}, namely those functions $f \in L^2_G(\mathbb{C})$ such that for every $s \in G$,

$$f(sts^{-1}) = f(t) \quad \text{almost everywhere.}$$

So, we have

$$\int_G \chi_\rho(t) \overline{\chi_{\rho'}(t)} \, dt = 0 \quad \text{if } \rho \neq \rho'$$

and

$$\int_G |\chi_\rho(t)|^2 \, dt = 1.$$

If $G$ is commutative, then all representations $M_\rho$ are one-dimensional. Then, each character $s \mapsto \chi_\rho(s)$ is a continuous homomorphism of $G$ into $\text{U}(1)$, the group of unit complex
numbers. For the torus group $S_1 = T = \mathbb{R}/\mathbb{Z}$, the characters are the homomorphisms $\theta \mapsto e^{k2\pi i \theta}$, with $k \in \mathbb{N}$. This is the special case of Fourier analysis on the circle.

**Remark:** The Peter–Weyl theorem actually holds for any compact topological metrizable group, not just for a compact Lie group.

A complete proof of Theorem 26.34 is given in Dieudonné [52], Chapter XXI, Section 2, but see also Sections 3 and 4.

An important corollary of the Peter–Weyl theorem is that every compact Lie group is isomorphic to a matrix group.

**Theorem 26.35.** For every compact Lie group $G$, there is some integer $N \geq 1$ and an isomorphism of $G$ onto a closed subgroup of $U(N)$.

The proof of Theorem 26.35 can be found in Dieudonné [52], Chapter XXI, Theorem 21.13.1 or Knapp [106] (Chapter 4, Corollary 4.22).

There is more to the Peter–Weyl Theorem: It gives a description of all unitary representations of $G$ into a separable Hilbert space. Recall that a Hilbert space is separable if it has a countable total orthogonal family, also called a Hilbert basis; see Definition 26.9.

If $f : E \to E$ is function from the compact Lie group $G$ to a Hilbert space $E$ and if for all $z \in E$ the function $s \mapsto \langle f(s), z \rangle$ is integrable and the function $s \mapsto \|f(s)\|$ is integrable, then it can be shown that there is a unique $y \in E$ such that

$$
\langle y, z \rangle = \int_G \langle f(s), z \rangle ds \quad \text{for all } z \in E;
$$

see Dieudonné [55] (Chapter XIII, Proposition 13.10.4). The vector $y \in E$ as above is denoted by

$$
\int_G f(s) ds
$$

and is called the *weak integral* (for short, *integral*) of $f$.

If $V : G \to \text{Aut}(E)$ is a representation of $G$ in a separable Hilbert space $E$, for every $\rho \in R(G)$ as above, for every $x \in E$ the map

$$
x \mapsto V_{u_\rho}(x) = \int_G u_\rho(s)(V(s)(x)) ds
$$

is an orthogonal projection of $E$ onto a closed subspace $E_\rho$, where the expression on the right-hand side is the weak integral of the function $s \mapsto u_\rho(s)(V(s)(x))$.

Then, $E$ is the Hilbert sum $E = \bigoplus_\rho E_\rho$ of those $E_\rho$ such that $E_\rho \neq (0)$. Each such $E_\rho$ is invariant under $V$, but the subrepresentation of $V$ in $E_\rho$ is not necessarily irreducible. However, each $E_\rho$ is a (finite or countable) Hilbert sum of closed subspaces invariant under $V$, and the subrepresentations of $V$ corresponding to these subspaces of $E_\rho$ are all equivalent.
to \( M_\rho \), where \( M_\rho \) is the representation of \( G \) given by \( M_\rho(g) = \overline{M_\rho(g)} \) for all \( g \in G \). These representations are all irreducible. As a consequence, every irreducible unitary representation of \( G \) is equivalent to some representation of the form \( M_\rho \).

If \( E_\rho \neq (0) \), we say that the irreducible representation \( M_\rho \) is contained in the representation \( V \). If \( E_\rho \) is finite-dimensional, then \( \dim(E_\rho) = d_\rho n_\rho \) for some positive integer \( d_\rho \). The integer \( d_\rho \) is called the multiplicity of \( M_\rho \) in \( V \).

An interesting special case is the case of the so-called regular representation of \( G \) in \( L^2_G(G) \) itself. The (left) regular representation \( R \) of \( G \) in \( L^2_G(G) \) is defined by

\[
(R_s(f))(t) = \lambda_s(f)(t) = f(s^{-1}t), \quad f \in L^2_G(G), \ s, t \in G.
\]

It turns out that we also get the same Hilbert sum

\[
L^2_G(G) = \bigoplus_\rho a_\rho,
\]

but this time, the \( a_\rho \) generally do not correspond to irreducible subrepresentations of \( R \). However, \( a_\rho \) splits into \( d_\rho = n_\rho \) left ideals \( b_j^{(\rho)} \), where \( b_j^{(\rho)} \) is spanned by the \( j \)th column of \( M_\rho \), and all the subrepresentations of \( G \) in \( b_j^{(\rho)} \) are equivalent to \( M_\rho \), and thus are irreducible (see Dieudonné [51], Chapter 3).

Finally, assume that besides the compact Lie group \( G \), we also have a closed subgroup \( K \) of \( G \). Then, we know that \( M = G/K \) is a manifold called a homogeneous space, and \( G \) acts on \( M \) on the left. For example, if \( G = SO(n+1) \) and \( K = SO(n) \), then \( S^n = SO(n+1)/SO(n) \) (see Chapter 5 or Warner [175], Chapter 3). The subspace of \( L^2_G(G) \) consisting of the functions \( f \in L^2_G(G) \) that are right-invariant under the action of \( K \), that is, such that

\[
f(su) = f(s) \quad \text{for all } s \in G \text{ and all } u \in K,
\]

form a closed subspace of \( L^2_G(G) \) denoted \( L^2_G(G/K) \). For example, if \( G = SO(n+1) \) and \( K = SO(n) \), then \( L^2_G(G/K) = L^2_G(S^n) \).

It turns out that \( L^2_G(G/K) \) is invariant under the regular representation \( R \) of \( G \) in \( L^2_G(G) \), so we get a subrepresentation (of the regular representation) of \( G \) in \( L^2_G(G/K) \). Again, the Peter–Weyl theorem gives us a Hilbert sum decomposition of \( L^2_G(G/K) \) of the form

\[
L^2_G(G/K) = \bigoplus_\rho L_\rho = L^2_G(G/K) \cap a_\rho,
\]

for the same \( \rho \)'s as before. However, these subrepresentations of \( R \) in \( L_\rho \) are not necessarily irreducible. What happens is that there is some \( d_\rho \) with \( 0 \leq d_\rho \leq n_\rho \), so that if \( d_\rho \geq 1 \), then \( L_\rho \) is the direct sum of the subspace spanned by the first \( d_\rho \) columns of \( M_\rho \). In fact, \( d_\rho \) is the multiplicity of the trivial representation \( \sigma_0 \) of \( K \) in the restriction of the representation \( M_\rho \) to \( K \); for this reason, \( d_\rho \) is also denoted \( (\rho : \sigma_0) \) (see Dieudonné [51], Chapter 6 and Dieudonné [53], Chapter XXII, Sections 4-5).
26.10. GELFAND PAIRS, SPHERICAL FUNCTIONS, FOURIER TRANSFORM ⊗ 847

We can also consider the subspace of $L^2_\mathbb{C}(G)$ consisting of the functions $f \in L^2_\mathbb{C}(G)$ that are left-invariant under the action of $K$; that is, such that

$$f(ts) = f(s) \quad \text{for all } s \in G \text{ and all } t \in K.$$ 

This is a closed subspace of $L^2_\mathbb{C}(G)$ denoted $L^2_\mathbb{C}(K\backslash G)$. Then, we get a Hilbert sum decomposition of $L^2_\mathbb{C}(K\backslash G)$ of the form

$$L^2_\mathbb{C}(K\backslash G) = \bigoplus_\rho L'_\rho = L^2_\mathbb{C}(K\backslash G) \cap a_\rho,$$

and for the same $d_\rho$ as before, $L'_\rho$ is the direct sum of the subspace spanned by the first $d_\rho$ rows of $M_\rho$. We can also consider

$$L^2_\mathbb{C}(K\backslash G/K) = L^2_\mathbb{C}(G/K) \cap L^2_\mathbb{C}(K\backslash G) = \{f \in L^2_\mathbb{C}(G) \mid f(tsu) = f(s)\} \quad \text{for all } s \in G \text{ and all } t, u \in K.$$ 

From our previous discussion, we see that we have a Hilbert sum decomposition

$$L^2_\mathbb{C}(K\backslash G/K) = \bigoplus_\rho L_\rho \cap L'_\rho$$

and each $L_\rho \cap L'_\rho$ for which $d_\rho \geq 1$ is a matrix algebra of dimension $d^2_\rho$ having as a basis the functions $m_{i,j}^{(\rho)}$ for $1 \leq i, j \leq d_\rho$. As a consequence, the algebra $L^2_\mathbb{C}(K\backslash G/K)$ is commutative iff $d_\rho \leq 1$ for all $\rho$.

26.10 Gelfand Pairs, Spherical Functions, and Fourier Transform ⊗

If the algebra $L^2_\mathbb{C}(K\backslash G/K)$ is commutative (for the convolution product), we say that $(G, K)$ is a Gelfand pair (see Dieudonné [51], Chapter 8 and Dieudonné [53], Chapter XXII, Sections 6-7).

In this case, the $L_\rho$ in the Hilbert sum decomposition of $L^2_\mathbb{C}(G/K)$ are nontrivial of dimension $n_\rho$ iff $(\rho : \sigma_0) = d_\rho = 1$, and the subrepresentation $U$ (of the regular representation) of $G$ into $L_\rho$ is irreducible and equivalent to $M_\rho$. The space $L_\rho$ is generated by the functions $m_{1,1}^{(\rho)}, \ldots, m_{n_\rho,1}^{(\rho)}$, but the function

$$\omega_\rho(s) = \frac{1}{n_\rho} m_{1,1}^{(\rho)}(s)$$

plays a special role. This function called a zonal spherical function has some interesting properties.
First, $\omega$ is a continuous function, even a smooth function since $G$ is a Lie group. The function $\omega$ is such that $\omega_\rho(e) = 1$ (where $e$ is the identity element of the group, $G$), and

$$\omega_\rho(ust) = \omega_\rho(s)$$

for all $s \in G$ and all $u, t \in K$.

In addition, $\omega_\rho$ is of positive type. A function $f : G \to \mathbb{C}$ is of positive type iff

$$\sum_{j,k=1}^{n} f(s_j^{-1}s_k)z_jz_k \geq 0,$$

for every finite set $\{s_1, \ldots, s_n\}$ of elements of $G$ and every finite tuple $(z_1, \ldots, z_n) \in \mathbb{C}^n$.

When $L^2_\mathbb{C}(K\setminus G/K)$ is commutative, it is the Hilbert sum of all the 1-dimensional subspaces $\mathbb{C}\omega_\rho$ for all $\rho \in R(G)$ such that $d_\rho = 1$. The orthogonal projection of $L^2_\mathbb{C}(K\setminus G/K)$ onto $\mathbb{C}\omega_\rho$ is given by

$$g \mapsto g^*\omega_\rho \quad g \in L^2_\mathbb{C}(K\setminus G/K).$$

Since $\mathbb{C}\omega_\rho$ is an ideal in the algebra $L^2_\mathbb{C}(K\setminus G/K)$, there is some homomorphism $\xi_\rho : L^2_\mathbb{C}(K\setminus G/K) \to \mathbb{C}$ such that

$$g^*\omega_\rho = \xi_\rho(g)\omega_\rho \quad g \in L^2_\mathbb{C}(K\setminus G/K).$$

To be more precise, $\xi_\rho$ has the property

$$\xi_\rho(g_1g_2) = \xi_\rho(g_1)\xi_\rho(g_2) \quad \text{for all } g_1, g_2 \in L^2_\mathbb{C}(K\setminus G/K).$$

In other words, $\xi_\rho$ is a character of the algebra $L^2_\mathbb{C}(K\setminus G/K)$.

Because the subrepresentation of $G$ into $L_\rho$ is irreducible, the function $\omega_\rho$ generates $L_\rho$ under left translation. This means the following: If we recall that for any function $f$ on $G$,

$$\lambda_s(f)(t) = f(s^{-1}t), \quad s, t \in G,$$

then $L_\rho$ is generated by the functions $\lambda_s(\omega_\rho)$, as $s$ varies in $G$.

The set of zonal spherical functions on $G/K$ is denoted $S(G/K)$. Because $G$ is compact it is a countable set in bijection with the set of equivalence classes of representations $\rho \in R(G)$ such that $(\rho : \sigma_0) = 1$.

It can be shown that a (non-identically zero) function $\omega$ in the set $C_\mathbb{C}(K\setminus G/K)$ of continuous complex-valued functions in $L^2_\mathbb{C}(K\setminus G/K)$ belongs to $S(G/K)$ iff the functional equation

$$\int_K \omega(xs)ds = \omega(x)\omega(y) \quad (*)$$

holds for all $s \in K$ and all $x, y \in G$.

It is remarkable that fairly general criteria (due to Gelfand) for a pair $(G, K)$ to be a Gelfand pair exist. This is certainly the case if $G$ is commutative and $K = (e)$; this situation
corresponds to commutative harmonic analysis. If \( G \) is a semisimple compact connected Lie group and if \( \sigma : G \to G \) is an involutive automorphism of \( G \), if \( K \) is the subgroup of fixed points of \( \sigma \)

\[
K = \{ s \in G \mid \sigma(s) = s \},
\]

then it can be shown that \((G, K)\) is a Gelfand pair. Involutive automorphisms as above were determined explicitly by E. Cartan.

It turns out that \( G = \text{SO}(n + 1) \) and \( K = \text{SO}(n) \) form a Gelfand pair corresponding to the above situation (see Dieudonné [51], Chapters 7-8 and Dieudonné [54], Chapter XXIII, Section 38). In this particular case, \( \rho = k \) is any nonnegative integer and \( L_\rho = E_k \), the eigenspace of the Laplacian on \( S^n \) corresponding to the eigenvalue \(-k(n + k - 1)\); all this was shown in Section 26.5. Therefore, the regular representation of \( \text{SO}(n) \) into \( E_k \) is irreducible. This can be proved more directly; for example, see Helgason [87] (Introduction, Theorem 3.1) or Bröcker and tom Dieck [31] (Chapter 2, Proposition 5.10).

The zonal spherical harmonics \( \omega_k \) can be expressed in terms of the ultraspherical polynomials (also called Gegenbauer polynomials) \( P_{n+1}^{n-1/2} \) (up to a constant factor); this was discussed in Sections 26.6 and 26.7. The reader should also consult Stein and Weiss [165] (Chapter 4), Morimoto [132] (Chapter 2) and Dieudonné [51] (Chapter 7). For \( n = 2 \), \( P_k^2 \) is just the ordinary Legendre polynomial (up to a constant factor).

If \((G, K)\) is a Gelfand pair (with \( G \) a compact group), it is possible to define the Fourier transform \( \mathcal{F}(f) \) of a function \( f \in L^2(K \setminus G/K) \) as the function \( \mathcal{F}(f) : S(G/K) \to \mathbb{C} \) given by

\[
\mathcal{F}(f)(\omega) = \int_G f(s)\omega(s^{-1}) \, ds \quad \omega \in S(G/K).
\]

More explicitly, because \( \omega_\rho = \frac{1}{n_\rho} m_{\rho,1} \) and \( m_{\rho,1}(s^{-1}) = \overline{m_{\rho,1}(s)} \), the Fourier transform \( \mathcal{F}(f) \) is the countable family

\[
\rho \mapsto \frac{1}{n_\rho} \langle f, m_{\rho,1} \rangle = \int_G f(s)\omega_\rho(s^{-1}) \, ds
\]

for all \( \rho \in R(G) \) such that \( (\rho : \sigma_0) = 1 \). The Fourier transform satisfies the fundamental relation

\[
\mathcal{F}(f * g) = \mathcal{F}(f)\mathcal{F}(g).
\]

The notion of Gelfand pair and of the Fourier transform can be generalized to locally-compact unimodular groups that are not necessary compact, but we will not discuss this here. Let us just say that when \( G \) is a commutative locally-compact group and \( K = (e) \), then equation (*) implies that

\[
\omega(xy) = \omega(x)\omega(y),
\]

which means that the functions \( \omega \) are characters of \( G \), so \( S(G/K) \) is the Pontrjagin dual group \( \hat{G} \) of \( G \), which is the group of characters of \( G \) (continuous homomorphisms of \( G \) into
the group $U(1)$. In this case, the Fourier transform $\mathcal{F}(f)$ is defined for every function $f \in L^1_c(G)$ as a function on the characters of $G$. This is the case of commutative harmonic analysis, as discussed in Folland [67] and Deitmar [48]. For more on Gelfand pairs, curious readers may consult Dieudonné [51] (Chapters 8 and 9) and Dieudonné [53] (Chapter XXII, Sections 6-9). Another approach to spherical functions (not using Gelfand pairs) is discussed in Helgason [87] (Chapter 4). Helgason [86] contains a short section on Gelfand pairs (chapter III, Section 12).

The material in this section belongs to the overlapping areas of representation theory and noncommutative harmonic analysis. These are deep and vast areas. Besides the references cited earlier, for noncommutative harmonic analysis, the reader may consult Knapp [105], Folland [67], Taylor [168], or Varadarajan [172], but they may find the pace rather rapid. Another great survey on both topics is Kirillov [103], although it is not geared for the beginner. In a different direction, namely Fourier analysis on finite groups, Audrey Terras’s book [169] contains some fascinating material.
Chapter 27

The Laplace-Beltrami Operator, Harmonic Forms, The Connection Laplacian and Weitzenböck Formulae

27.1 The Gradient and Hessian Operators on Riemannian Manifolds

The Laplacian is a very important operator because it shows up in many of the equations used in physics to describe natural phenomena such as heat diffusion or wave propagation. Therefore, it is highly desirable to generalize the Laplacian to functions defined on a manifold. Furthermore, in the late 1930’s, George de Rham (inspired by Élie Cartan) realized that it was fruitful to define a version of the Laplacian operating on differential forms, because of a fundamental and almost miraculous relationship between harmonics forms (those in the kernel of the Laplacian) and the de Rham cohomology groups on a (compact, orientable) smooth manifold. Indeed, as we will see in Section 27.4, for every cohomology group $H^k_{DR}(M)$, every cohomology class $[\omega] \in H^k_{DR}(M)$ is represented by a unique harmonic $k$-form $\omega$. This connection between analysis and topology lies deep and has many important consequences. For example, Poincaré duality follows as an “easy” consequence of the Hodge Theorem.

Technically, the Laplacian can be defined on differential forms using the Hodge $*$ operator (Section 22.6). On functions, there are alternate definitions of the Laplacian using only the covariant derivative and obtained by generalizing the notions of gradient and divergence to functions on manifolds.

Another version of the Laplacian can be defined in terms of the adjoint of the connection $\nabla$ on differential forms, viewed as a linear map from $\mathcal{A}^*(M)$ to $\text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$. We obtain the connection Laplacian (also called Bochner Laplacian) $\nabla^* \nabla$. Then, it is natural to wonder how the Hodge Laplacian $\Delta$ differs from the connection Laplacian $\nabla^* \nabla$? Remarkably, there is a formula known as Weitzenböck’s formula (or Bochner’s formula) of

851
the form
\[ \Delta = \nabla^* \nabla + C(R\nabla), \]
where \( C(R\nabla) \) is a contraction of a version of the curvature tensor on differential forms (a fairly complicated term). In the case of one-forms,
\[ \Delta = \nabla^* \nabla + \text{Ric}, \]
where Ric is a suitable version of the Ricci curvature operating on one-forms.

Weitzenböck-type formulae are at the root of the so-called “Bochner Technique,” which consists in exploiting curvature information to deduce topological information. For example, if the Ricci curvature on a compact orientable Riemannian manifold is strictly positive, then \( H^1_{\text{DR}}(M) = (0) \), a theorem due to Bochner.

In preparation for defining the (Hodge) Laplacian, we define the gradient of a function on a Riemannian manifold, as well as the Hessian, which plays an important role in optimization theory. Unlike the situation where \( M \) is a vector space (\( M \) is flat), the Riemannian metric on \( M \) is critically involved in the definition of the gradient and of the Hessian.

If \( (M, \langle - , - \rangle) \) is a Riemannian manifold of dimension \( n \), then for every \( p \in M \), the inner product \( \langle - , - \rangle_p \) on \( T_pM \) yields a canonical isomorphism \( \flat : T_pM \to T^*_pM \), as explained in Sections 21.2 and 29.5. Namely, for any \( u \in T_pM \), \( u^\flat = \flat(u) \) is the linear form in \( T^*_pM \) defined by
\[ u^\flat(v) = \langle u, v \rangle_p, \quad v \in T_pM. \]
Recall that the inverse of the map \( \flat \) is the map \( \sharp : T^*_pM \to T_pM \). As a consequence, for every smooth function \( f \in C^\infty(M) \), we get smooth vector field \( \text{grad } f = (df)^\sharp \) defined so that
\[ (\text{grad } f)_p = (df_p)^\sharp. \]

**Definition 27.1.** For every smooth function \( f \) over a Riemannian manifold \( (M, \langle - , - \rangle) \), the vector field \( \text{grad } f \) defined by
\[ \langle (\text{grad } f)_p, u \rangle_p = df_p(u), \quad \text{for all } u \in T_pM, \text{ and all } p \in M, \]
is the gradient of the function \( f \).

Conversely, a vector field \( X \in \mathfrak{X}(M) \) yields the one-form \( X^\flat \in \mathcal{A}^1(M) \) given by
\[ (X^\flat)_p = (X_p)^\flat. \]
The one-form \( X^\flat \) is uniquely defined by the equation
\[ (X^\flat)_p(v) = \langle X_p, v \rangle_p, \quad \text{for all } p \in M \text{ and all } v \in T_pM. \]

In view of this equation, the one-form \( X^\flat \) is an insertion operator in the sense discussed in Section 22.7 just after Proposition 22.18, so it is also denoted by \( i_Xg \), where \( g = \langle - , - \rangle \) is the Riemannian metric on \( M \).
27.1. **The Gradient and Hessian Operators**

In the special case $X = \text{grad } f$, we have

$$(\text{grad } f)^b_p(v) = \langle (\text{grad } f)_p, v \rangle = df_p(v),$$

and since $dd = 0$, we deduce that

$$d(\text{grad } f)^b = 0.$$ 

Therefore, for an arbitrary vector field $X$, the 2-form $dX^b$ measures the extent to which $X$ is a gradient field.

The Hessian was already defined in Chapter 12, but for the sake of completeness we repeat the definition.

**Definition 27.2.** The Hessian $\text{Hess}(f)$ (or $\nabla^2(f)$) of a function $f \in C^\infty(M)$ is the $(0,2)$-tensor defined by

$$\text{Hess}(f)(X,Y) = X(Y(f)) - (\nabla_XY)(f) = df(Y) - X(\nabla_Y f),$$

for all vector fields $X,Y \in \mathfrak{X}(M)$.

Recall from Proposition 29.5 that the covariant derivative $\nabla_X \theta$ of any one-form $\theta \in \mathcal{A}^1(M)$ is the one-form given by

$$(\nabla_X \theta)(Y) = X(\theta(Y)) - \theta(\nabla_X Y),$$

so the Hessian $\text{Hess}(f)$ is also defined by

$$\text{Hess}(f)(X,Y) = (\nabla_X df)(Y).$$

According to the notational convention for the covariant derivative of a tensor stated just after Proposition 29.5,

$$(\nabla df)(X,Y) = (\nabla_X df)(Y) = \text{Hess}(f)(X,Y),$$

which means that the $(0,2)$-tensor $\text{Hess}(f)$ is given by

$$\text{Hess}(f) = \nabla df.$$ 

Since by definition $\nabla_X f = df(X)$, we can also write $\text{Hess}(f) = \nabla \nabla f$, but we find this expression confusing.

Since $\nabla$ is torsion-free, we get

$$\text{Hess}(f)(X,Y) = X(Y(f)) - (\nabla_X Y)(f) = Y(X(f)) - (\nabla_Y X)(f) = \text{Hess}(f)(Y,X),$$

which means that the Hessian is a symmetric $(0,2)$-tensor. For the convenience of the reader, we repeat Proposition 12.4 (which is proved in Chapter 12).
**Proposition 27.1.** The Hessian is given by the equation
\[
\text{Hess}(f)(X,Y) = \langle \nabla_X (\text{grad } f), Y \rangle, \quad X, Y \in \mathfrak{X}(M).
\]

The Hessian can also be defined in terms of Lie derivatives; this is the approach followed by Petersen [140] (Chapter 2, Section 1.3). We observed in Section 29.4 that the Levi-Civita connection can be defined in terms of the Lie derivative of the Riemannian metric \( g \) on \( M \) by the equation
\[
2g(\nabla_X Y, Z) = (L_Y g)(X, Z) + (d(i_Y g))(X, Z), \quad X, Y, Z \in \mathfrak{X}(M).
\]

**Proposition 27.2.** The Hessian of \( f \) is given by
\[
\text{Hess}(f) = \frac{1}{2} L_{\text{grad } f} g.
\]

**Proof.** To prove the above equation, we use the fact that \( d(i_{\text{grad } f}) = 0 \) and Proposition 12.4. We have
\[
2\text{Hess}(f)(X,Y) = 2g(\nabla_X (\text{grad } f), Y)
= (L_{\text{grad } f} g)(X,Y) + (d(i_{\text{grad } f})(X,Y)
= (L_{\text{grad } f} g)(X,Y),
\]
as claimed. \( \square \)

Since the Hessian is a symmetric bilinear form, it is determined by the quadratic form \( X \mapsto \text{Hess}(f)(X,X) \), and it can be recovered by polarization from this quadratic form. There is also a way to compute \( \text{Hess}(f)(X,X) \) using geodesics. When geodesics are easily computable, this is usually the simplest way to compute the Hessian.

**Proposition 27.3.** Given any \( p \in M \) and any tangent vector \( X \in T_pM \), if \( \gamma \) is a geodesic such that \( \gamma(0) = p \) and \( \gamma'(0) = X \), then at \( p \), we have
\[
\text{Hess}(f)_p(X,X) = \frac{d^2}{dt^2} f(\gamma(t)) \bigg|_{t=0}.
\]

**Proof.** To prove the above formula, following Jost [99], we have
\[
X(X(f))(p) = \gamma'(\langle \text{grad } f \rangle_p, \gamma')
= \gamma' \left( \frac{d}{dt} f(\gamma(t)) \bigg|_{t=0} \right)
= \frac{d^2}{dt^2} f(\gamma(t)) \bigg|_{t=0}.
\]
Furthermore, since \( \gamma \) is a geodesic, \( \nabla_{\gamma'} \gamma' = 0 \), so we get
\[
\text{Hess}(f)_p(X,X) = X(X(f))(p) - (\nabla_X X)(f)(p) = X(X(f))(p),
\]
which proves our claim. \( \square \)
In local coordinates with respect to a chart, if we write
\[ df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i, \]
then it is easy to see that
\[ \nabla df = \sum_{i,j=1}^{n} \left( \frac{\partial^2 f}{\partial x_i \partial x_j} - \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} \Gamma_{ij}^k \right) dx_i \otimes dx_j, \]
where the \( \Gamma_{ij}^k \) are the Christoffel symbols of the connection in the chart.

A function \( f \in C^\infty(M) \) is convex (resp. strictly convex) iff its Hessian \( \text{Hess}(f) \) is positive semi-definite (resp. positive definite).

The computation of the gradient of a function defined either on the Stiefel manifold or on the Grassmannian manifold is instructive. Let us first consider the Stiefel manifold \( S(k,n) \). Recall from Section 19.4 that for every \( n \times k \) matrix \( Y \in S(k,n) \), tangent vectors to \( S(k,n) \) at \( Y \) are of the form
\[ X = YS + Y_\perp A, \]
where \( S \) is any \( k \times k \) skew-symmetric matrix, \( A \) is any \( (n-k) \times k \) matrix, and \( [Y Y_\perp] \) is an orthogonal matrix. Given any differentiable function \( F: S(k,n) \to \mathbb{R} \), if we let \( F_Y \) be the \( n \times k \) matrix of partial derivatives
\[ F_Y = \left( \frac{\partial F}{\partial Y_{ij}} \right), \]
then we have
\[ df_Y(X) = \text{tr}(F_Y^T X). \]
The gradient \( \text{grad}(F)_Y \) of \( F \) at \( Y \) is the uniquely defined tangent vector to \( S(k,n) \) at \( Y \) such that
\[ \langle \text{grad}(F)_Y, X \rangle = df_Y(X) = \text{tr}(F_Y^T X), \quad \text{for all } X \in T_Y S(k,n). \]
For short, if write \( Z = \text{grad}(F)_Y \), then \( Z \) must satisfy the equation
\[ \text{tr}(F_Y^T X) = \text{tr} \left( Z^T \left( I - \frac{1}{2} YY^T \right) X \right), \]
and since \( Z \) is of the form \( Z = YS + Y_\perp A \), we get
\[ \text{tr}(F_Y^T X) = \text{tr} \left( (S^T Y^T + A^T Y_\perp^T) \left( I - \frac{1}{2} YY^T \right) X \right) = \text{tr} \left( \left( \frac{1}{2} S^T Y^T + A^T Y_\perp^T \right) X \right). \]
for all $X \in T_Y S_{k,n}$. The above equation implies that we must find $Z = YS + Y_\perp A$ such that

$$F_Y^T = \frac{1}{2} S^T Y^T + A^T Y_\perp^T,$$

which is equivalent to

$$F_Y = \frac{1}{2} YS + Y_\perp A.$$

From the above equation, we deduce that

$$Y_\perp^T F_Y = A,
Y^T F_Y = \frac{1}{2} S.$$Since $S$ is skew-symmetric, we get

$$F_Y^T Y = -\frac{1}{2} S,$$so

$$S = Y^T F_Y - F_Y^T Y,$$and thus,

$$Z = YS + Y_\perp A
= Y(Y^T F_Y - F_Y^T Y) + Y_\perp Y_\perp^T F_Y
= (YY^T + Y_\perp Y_\perp^T) F_Y - YF_Y^T Y
= F_Y - YF_Y^T Y.$$Therefore, we proved that the gradient of $F$ at $Y$ is given by

$$\text{grad}(F)_Y = F_Y - YF_Y^T Y.$$Let us now turn to the Grassmannian $G(k,n)$. In Section 19.9, we showed that tangent vectors to $G(k,n)$ at $[Y]$ are of the form

$$X = Y_\perp A,$$Where $A$ is any $(n-k) \times k$ matrix. We would like to compute the gradient at $[Y]$ of a function $F: G(k,n) \to \mathbb{R}$. Again, if write $Z = \text{grad}(F)_Y$, then $Z$ must satisfy the equation

$$\text{tr}(F_Y^T X) = (Z, X) = \text{tr}(Z^T X), \text{ for all } X \in T_{[Y]} G(k,n).$$Since $Z$ is of the form $Z = Y_\perp A$, we get

$$\text{tr}(F_Y^T X) = \text{tr}(A^T Y_\perp^T X), \text{ for all } X \in T_{[Y]} G(k,n),$$
27.1. THE GRADIENT AND HESSIAN OPERATORS

which implies that

\[ F_Y^\top = A^\top Y_\perp^\top; \]

that is,

\[ F_Y = Y_\perp A. \]

The above yields

\[ A = Y_\perp^\top F_Y, \]

so we have

\[ Z = Y_\perp Y_\perp^\top F_Y = (I - YY^\top)F_Y. \]

Therefore, the gradient of \( F \) at \([Y]\) is given by

\[ \text{grad}(F)_Y = F_Y - YY^\top F_Y. \]

Since the geodesics in the Stiefel manifold and in the Grassmannian can be determined explicitly, we can find the Hessian of a function using the formula

\[ \text{Hess}(f)_p(X, X) = d^2 \frac{dt}{dt}^2 f(\gamma(t)) \bigg|_{t=0}. \]

Let us do this for a function \( F \) defined on the Grassmannian, the computation on the Stiefel manifold being more complicated; see Edelman, Arias and Smith [65] for details.

For any two tangent vectors \( X_1, X_2 \in T_{[Y]} G(k, n) \) to \( G(k, n) \) at \([Y]\), define \( F_{YY}(X_1, X_2) \) by

\[ F_{YY}(X_1, X_2) = \sum_{ij,kl} (F_{YY})_{ij,kl}(X_1)_{ij}(X_2)_{kl}, \]

with

\[ (F_{YY})_{ij,kl} = \frac{\partial^2 F}{\partial Y_{ij} \partial Y_{kl}}. \]

Then, we find that

\[ \text{Hess}(F)(X_1, X_2) = F_{YY}(X_1, X_2) - \text{tr}(X_1^\top X_2 Y^\top F_Y), \]

where

\[ F_Y = \left( \frac{\partial F}{\partial Y_{ij}} \right), \]

as above, when we found a formula for the gradient of \( F \) at \( Y \).

If one wishes to use Newton’s method, then it is necessary to “invert” the Hessian, in the sense that, given \( G = \text{grad}(F) = F_Y - YY^\top F_Y \) at \([Y]\), we must determine a tangent vector \( X \) such that

\[ \text{Hess}(F)(X, Z) = \langle -G, Z \rangle = \text{tr}(-Z^\top G), \quad \text{for all } Z \in T_{[Y]} G(k, n). \]
Since
\[ \text{Hess}(F)(X, Z) = \text{Hess}(F)(Z, X) = F_{YY}(Z, X) - \text{tr}(Z^T XY^T F_Y), \]
we must have
\[ F_{YY}(Z, X) - \text{tr}(Z^T XY^T F_Y) = \text{tr}(-Z^T G), \]
for all \( Z \in T_{[Y]}G(k, n), \)
so if \( F_{YY}(X) \) denotes the unique tangent vector such that
\[ F_{YY}(Z, X) = \langle Z, F_{YY}(X) \rangle = \text{tr}(Z^T F_{YY}(X)), \]
for all \( Z \in T_{[Y]}G(k, n), \)
we must have
\[ F_{YY}(X) - X(Y^T F_Y) = -G, \quad Y^T X = 0. \tag{\ast} \]
Since \( F_{YY}(Z, X) \) is linear in \( X \) and in \( Z \), the expression \( F_{YY}(X) \) is linear in \( X \), so problem \((\ast)\) is linear in \( X \). Some explicit examples for various functions \( F \) are discussed in Edelman, Arias and Smith [65].

### 27.2 The Hodge \( \ast \) Operator on Riemannian Manifolds

By the results of Section 22.6, the inner product \( \langle -, - \rangle_p \) on \( T_pM \) induces an inner product on \( \bigwedge^k T^*_pM \). Therefore, for any two \( k \)-forms \( \omega, \eta \in \mathcal{A}^k(M) \), we get the smooth function \( \langle \omega, \eta \rangle \) given by
\[ \langle \omega, \eta \rangle(p) = \langle \omega_p, \eta_p \rangle_p. \]

Furthermore, if \( M \) is oriented, then we can apply the results of Section 22.6 so the vector bundle \( T^*M \) is oriented (by giving \( T^*_pM \) the orientation induced by the orientation of \( T_pM \), for every \( p \in M \)), and for every \( p \in M \), we get a Hodge \( \ast \)-operator
\[ \ast : \bigwedge^k T^*_pM \rightarrow \bigwedge^{n-k} T^*_pM. \]

Then, given any \( k \)-form \( \omega \in \mathcal{A}^k(M) \), we can define \( \ast \omega \) by
\[ \ast \omega = \ast(\omega_p), \quad p \in M. \]

We have to check that \( \ast \omega \) is indeed a smooth form in \( \mathcal{A}^{n-k}(M) \), but this is not hard to do in local coordinates (for help, see Morita [133], Chapter 4, Section 1). Therefore, if \( M \) is a Riemannian oriented manifold of dimension \( n \), we have Hodge \( \ast \)-operators
\[ \ast : \mathcal{A}^k(M) \rightarrow \mathcal{A}^{n-k}(M). \]

Observe that \( \ast 1 \) is just the volume form \( \text{Vol}_M \) induced by the metric. Indeed, we know from Section 21.2 that in local coordinates \( x_1, \ldots, x_n \) near \( p \), the metric on \( T^*_pM \) is given by
27.3. THE LAPLACE-BELTRAMI AND DIVERGENCE OPERATORS

the inverse \((g^{ij})\) of the metric \((g_{ij})\) on \(T_p M\), and by the results of Section 22.6,

\[ * (1) = \frac{1}{\sqrt{\det(g_{ij})}} \, dx_1 \wedge \cdots \wedge dx_n = \sqrt{\det(g_{ij})} \, dx_1 \wedge \cdots \wedge dx_n = \text{Vol}_M. \]

Proposition 22.14 yields the following:

**Proposition 27.4.** If \(M\) is a Riemannian oriented manifold of dimension \(n\), then we have the following properties:

(i) \(* (f \omega + g \eta) = f * \omega + g * \eta, \) for all \(\omega, \eta \in \mathcal{A}^k(M)\) and all \(f, g \in C^\infty(M)\).

(ii) \(* * = (-\text{id})^{k(n-k)}\).

(iii) \(\omega \wedge * \eta = \eta \wedge * \omega = (\omega, \eta) \text{Vol}_M, \) for all \(\omega, \eta \in \mathcal{A}^k(M)\).

(iv) \(* (\omega \wedge * \eta) = (\eta \wedge * \omega) = (\omega, \eta)\), for all \(\omega, \eta \in \mathcal{A}^k(M)\).

(v) \((* \omega, * \eta) = (\omega, \eta), \) for all \(\omega, \eta \in \mathcal{A}^k(M)\).

Recall that exterior differentiation \(d\) is a map \(d: \mathcal{A}^k(M) \to \mathcal{A}^{k+1}(M)\). Using the Hodge \(*\)-operator, we can define an operator \(\delta: \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)\) that will turn out to be adjoint to \(d\) with respect to an inner product on \(\mathcal{A}^\bullet(M)\).

**Definition 27.3.** Let \(M\) be an oriented Riemannian manifold of dimension \(n\). For any \(k\), with \(1 \leq k \leq n\), let

\[ \delta = (-1)^{(k+1)n+1} * d * . \]

Clearly, \(\delta\) is a map \(\delta: \mathcal{A}^k(M) \to \mathcal{A}^{k-1}(M)\), and \(\delta = 0\) on \(\mathcal{A}^0(M) = C^\infty(M)\). It is easy to see that

\[ * \delta = (-1)^k d *, \quad \delta * = (-1)^{k+1} * d, \quad \delta \circ \delta = 0. \]

27.3 The Laplace-Beltrami and Divergence Operators on Riemannian Manifolds

Using \(d\) and \(\delta\), we can generalize the Laplacian to an operator on differential forms.

**Definition 27.4.** Let \(M\) be an oriented Riemannian manifold of dimension \(n\). The Laplace-Beltrami operator, for short Laplacian, is the operator \(\Delta: \mathcal{A}^k(M) \to \mathcal{A}^k(M)\) defined by

\[ \Delta = d \delta + \delta d. \]

A form, \(\omega \in \mathcal{A}^k(M)\) such that \(\Delta \omega = 0\) is a harmonic form. In particular, a function \(f \in \mathcal{A}^0(M) = C^\infty(M)\) such that \(\Delta f = 0\) is called a harmonic function.
The Laplacian in Definition 27.4 is also called the Hodge Laplacian.

If $M = \mathbb{R}^n$ with the Euclidean metric and $f$ is a smooth function, a laborious computation yields

$$\Delta f = -\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2};$$

that is, the usual Laplacian with a negative sign in front (the computation can be found in Morita [133], Example 4.12, or Jost [99], Chapter 2, Section 2.1). It is also easy to see that $\Delta$ commutes with $*$; that is,

$$\Delta * = * \Delta.$$

**Definition 27.5.** Let $M$ be an oriented Riemannian manifold of dimension $n$. Given any vector field $X \in \mathfrak{X}(M)$, its divergence $\text{div} X$ is defined by

$$\text{div} X = \delta X^\flat.$$

Now, for a function $f \in C^\infty(M)$, we have $\delta f = 0$, so $\Delta f = \delta df$. However,

$$\text{div}(\text{grad} f) = \delta(\text{grad} f)^\flat = \delta((df)^\flat)^\flat = \delta df,$$

so

$$\Delta f = \text{div} \text{grad} f,$$

as in the case of $\mathbb{R}^n$.

**Remark:** Since the definition of $\delta$ involves two occurrences of the Hodge $*$-operator, $\delta$ also makes sense on non-orientable manifolds by using a local definition. Therefore, the Laplacian $\Delta$ and the divergence also makes sense on non-orientable manifolds.

In the rest of this section, we assume that $M$ is orientable.

The relationship between $\delta$ and $d$ can be made clearer by introducing an inner product on forms with compact support. Recall that $\mathcal{A}_k^c(M)$ denotes the space of $k$-forms with compact support (an infinite dimensional vector space). For any two $k$-forms with compact support $\omega, \eta \in \mathcal{A}_k^c(M)$, set

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle \, \text{Vol}_M = \int_M \langle \omega, \eta \rangle \ast (1).$$

Using Proposition 27.4, we have

$$(\omega, \eta) = \int_M \langle \omega, \eta \rangle \, \text{Vol}_M = \int_M \omega \wedge * \eta = \int_M \eta \wedge * \omega,$$

so it is easy to check that $(-, -)$ is indeed an inner product on $k$-forms with compact support. We can extend this inner product to forms with compact support in $\mathcal{A}^c(M) = \bigoplus_{k=0}^{n} \mathcal{A}_k^c(M)$ by making $\mathcal{A}_h^c(M)$ and $\mathcal{A}_k^c(M)$ orthogonal if $h \neq k$. 
Proposition 27.5. If $M$ is an orientable Riemannian manifold, then $\delta$ is (formally) adjoint to $d$; that is,
\[(d\omega, \eta) = (\omega, \delta\eta),\]
for all $k$-forms $\omega, \eta$ with compact support.

Proof. By linearity and orthogonality of the $\mathcal{A}_c^k(M)$, the proof reduces to the case where $\omega \in \mathcal{A}_c^{k-1}(M)$ and $\eta \in \mathcal{A}_c^k(M)$ (both with compact support). By definition of $\delta$ and the fact that
\[*\ast = (-\text{id})^{(k-1)(n-k+1)}\]
for $\ast: \mathcal{A}^{k-1}(M) \to \mathcal{A}^{n-k+1}(M)$, we have
\[\ast\delta = (-1)^kd\ast,\]
and since
\[d(\omega \wedge \ast\eta) = d\omega \wedge \ast\eta + (-1)^{k-1}\omega \wedge \ast d\eta\]
\[= d\omega \wedge \ast\eta - \omega \wedge \ast\delta\eta\]
we get
\[\int_M d(\omega \wedge \ast\eta) = \int_M d\omega \wedge \ast\eta - \int_M \omega \wedge \ast\delta\eta\]
\[= (d\omega, \eta) - (\omega, \delta\eta).\]

However, by Stokes Theorem (Theorem 24.16),
\[\int_M d(\omega \wedge \ast\eta) = 0,\]
so $(d\omega, \eta) - (\omega, \delta\eta) = 0$; that is, $(d\omega, \eta) = (\omega, \delta\eta)$, as claimed.

Corollary 27.6. If $M$ is an orientable Riemannian manifold, then the Laplacian $\Delta$ is self-adjoint; that is,
\[(\Delta\omega, \eta) = (\omega, \Delta\eta),\]
for all $k$-forms $\omega, \eta$ with compact support.

We also obtain the following useful fact:

Proposition 27.7. If $M$ is an orientable Riemannian manifold, then for every $k$-form $\omega$ with compact support, $\Delta\omega = 0$ iff $d\omega = 0$ and $\delta\omega = 0$.

Proof. Since $\Delta = d\delta + \delta d$, it is obvious that if $d\omega = 0$ and $\delta\omega = 0$, then $\Delta\omega = 0$. Conversely,
\[(\Delta\omega, \omega) = ((d\delta + \delta d)\omega, \omega) = (d\delta\omega, \omega) + (\delta d\omega, \omega) = (\delta\omega, \delta\omega) + (d\omega, d\omega).\]
Thus, if $\Delta\omega = 0$, then $(\delta\omega, \delta\omega) = (d\omega, d\omega) = 0$, which implies $d\omega = 0$ and $\delta\omega = 0$. 

\[\square\]
As a consequence of Proposition 27.7, if $M$ is a connected, orientable, compact Riemannian manifold, then every harmonic function on $M$ is a constant. Indeed, if $M$ is compact then $f$ is a 0-form of compact support, and if $\Delta f = 0$ then $df = 0$. Since $f$ is connected, $f$ is a constant function.

For practical reasons, we need a formula for the Laplacian of a function $f \in C^\infty(M)$, in local coordinates. If $(U, \varphi)$ is a chart near $p$, as usual, let
\[
\frac{\partial f}{\partial x_j}(p) = \frac{\partial(f \circ \varphi^{-1})}{\partial u_j}(\varphi(p)),
\]
where $(u_1, \ldots, u_n)$ are the coordinate functions in $\mathbb{R}^n$. Write $|g| = \det(g_{ij})$, where $(g_{ij})$ is the symmetric, positive definite matrix giving the metric in the chart $(U, \varphi)$.

**Proposition 27.8.** If $M$ is an orientable Riemannian manifold, then for every local chart $(U, \varphi)$, for every function $f \in C^\infty(M)$, we have
\[
\Delta f = -\frac{1}{\sqrt{|g|}} \sum_{i,j} \frac{\partial}{\partial x_i}\left(\sqrt{|g|} g^{ij} \frac{\partial f}{\partial x_j}\right).
\]

**Proof.** We follow Jost [99], Chapter 2, Section 1. Pick any function $h \in C^\infty(M)$ with compact support. We have
\[
\int_M (\Delta f) h \ast (1) = (\Delta f, h) = (\delta df, h) = (df, dh) = \int_M \left< df, dh \right> \ast (1) = \int_M \sum_{ij} g^{ij} \frac{\partial f}{\partial x_i} \frac{\partial h}{\partial x_j} \ast (1) = -\int_M \sum_{ij} \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_j}\left(\sqrt{|g|} g^{ij} \frac{\partial f}{\partial x_i}\right) h \ast (1),
\]
where we have used integration by parts in the last line. Since the above equation holds for all $h$, we get our result. \qed

It turns out that in a Riemannian manifold, the divergence of a vector field and the Laplacian of a function can be given a definition that uses the covariant derivative (see Chapter 29, Section 29.1) instead of the Hodge $\ast$-operator. For the sake of completeness, we present this alternate definition which is the one used in Gallot, Hulin and Lafontaine [73] (Chapter 4), and O’Neill [138] (Chapter 3).
Definition 27.6. Let $M$ be a Riemannian manifold. If $\nabla$ is the Levi-Civita connection induced by the Riemannian metric, then the connection divergence (for short divergence) of a vector field $X \in \mathfrak{X}(M)$ is the function $\text{div} X : M \to \mathbb{R}$ defined so that

$$(\text{div} X)(p) = \text{tr}(Y(p) \mapsto (-\nabla_Y X)_p);$$

namely, for every $p$, $(\text{div} X)(p)$ is the trace of the linear map $Y(p) \mapsto (-\nabla_Y X)_p$.

Of course, for any function $f \in C^\infty(M)$, we define $\Delta f$ by

$$\Delta f = \text{div} \text{grad} f.$$

Observe that the above definition of the divergence (and of the Laplacian) makes sense even if $M$ is non-orientable. For orientable manifolds, the equivalence of this new definition of the divergence with our definition is proved in Petersen [140] (see Chapter 7). The main reason is the following result proved in O’Neill [138] (Chapter 7, Lemma 21).

Proposition 27.9. Let $M$ be a Riemannian manifold. For any vector field $X \in \mathfrak{X}(M)$, we have

$$L_X \text{Vol}_M = -(\text{div} X)\text{Vol}_M,$$

where div $X$ is the connection divergence of $X$.

By Cartan’s Formula (Proposition 23.15), $L_X = i(X) \circ d + d \circ i(X)$; as $d\text{Vol}_M = 0$ (since $\text{Vol}_M$ is a top form) we get

$$(\text{div} X)\text{Vol}_M = -d(i(X)\text{Vol}_M).$$

The above formulae also holds for a local volume form (i.e. for a volume form on a local chart).

The operator $\delta : \mathcal{A}^1(M) \to \mathcal{A}^0(M)$ can also be defined in terms of the covariant derivative (see Gallot, Hulin and Lafontaine [73], Chapter 4). For any one-form $\omega \in \mathcal{A}^1(M)$, recall that

$$(\nabla_X \omega)(Y) = X(\omega(Y)) - \omega(\nabla_X Y).$$

Then, it turns out that

$$\delta \omega = -\text{tr} \nabla \omega,$$

where the trace should be interpreted as the trace of the $\mathbb{R}$-bilinear map $X,Y \mapsto (\nabla_X \omega)(Y)$, as in Chapter 21 (see Proposition 21.3). This means that in any chart $(U, \varphi)$,

$$\delta \omega = -\sum_{i=1}^n (\nabla_{E_i} \omega)(E_i),$$

for any orthonormal frame field $(E_1, \ldots, E_n)$ over $U$. It can be shown that

$$\delta (f df) = f \Delta f - \langle \text{grad} f, \text{grad} f \rangle,$$
and as a consequence,
\[(\Delta f, f) = \int_M \langle \text{grad } f, \text{grad } f \rangle \text{Vol}_M,\]
for any orientable, compact manifold \(M\).

Since the proof of the next proposition is quite technical, we omit the proof which can be found in Petersen [140] (Chapter 7, Proposition 31).

**Proposition 27.10.** If \(M\) is an orientable and compact Riemannian manifold, then for every vector field \(X \in \mathfrak{X}(M)\), the connection divergence is given by
\[
\text{div } X = \delta X^\flat.
\]
Consequently, for the Laplacian, we have
\[
\Delta f = \delta df = \text{div } \text{grad } f.
\]

**Remark:** Some authors omit the negative sign in the definition of the divergence; that is, they define
\[
(\text{div } X)(p) = \text{tr} (Y(p) \mapsto (\nabla_Y X)_p).
\]

Here is a frequently used corollary of Proposition 27.10:

**Proposition 27.11.** (Green’s Formula) If \(M\) is an orientable and compact Riemannian manifold without boundary, then for every vector field \(X \in \mathfrak{X}(M)\), we have
\[
\int_M (\text{div } X) \text{Vol}_M = 0.
\]

**Proof.** Proofs of proposition 27.11 can be found in Gallot, Hulin and Lafontaine [73] (Chapter 4, Proposition 4.9) and Helgason [87] (Chapter 2, Section 2.4). As we explained earlier, as a consequence of Proposition 27.9 we have
\[
(\text{div } X) \text{Vol}_M = -d(i(X) \text{Vol}_M).
\]
Thus,
\[
\int_M (\text{div } X) \text{Vol}_M = -\int_M d(i(X) \text{Vol}_M) = -\int_{\partial M} i(X) \text{Vol}_M = 0
\]
by Stokes’ Theorem, since \(\partial M = 0\).

There is a generalization of the formula expressing \(\delta \omega\) over an orthonormal frame \(E_1, \ldots, E_n\) for a one-form \(\omega\) that applies to any differential form. In fact, there are formulae expressing both \(d\) and \(\delta\) over an orthonormal frame and its coframe, and these are often handy in proofs. Recall that for every vector field \(X \in \mathfrak{X}(M)\), the interior product \(i(X) : \mathcal{A}^{k+1}(M) \to \mathcal{A}^k(M)\) is defined by
\[
(i(X)\omega)(Y_1, \ldots, Y_k) = \omega(X, Y_1, \ldots, Y_k),
\]
for all \(Y_1, \ldots, Y_k \in \mathfrak{X}(M)\).
Proposition 27.12. Let $M$ be a compact, orientable, Riemannian manifold. For every $p \in M$, for every local chart $(U, \varphi)$ with $p \in M$, if $(E_1, \ldots, E_n)$ is an orthonormal frame over $U$ and $(\theta_1, \ldots, \theta_n)$ is its dual coframe, then for every $k$-form $\omega \in \mathcal{A}^k(M)$, we have:

$$d\omega = \sum_{i=1}^{n} \theta_i \wedge \nabla_{E_i} \omega$$

$$\delta \omega = -\sum_{i=1}^{n} i(E_i) \nabla_{E_i} \omega.$$

A proof of Proposition 27.12 can be found in Petersen [140] (Chapter 7, proposition 37) or Jost [99] (Chapter 3, Lemma 3.3.4). When $\omega$ is a one-form, $\delta \omega_p$ is just a number, and indeed

$$\delta \omega = -\sum_{i=1}^{n} i(E_i) \nabla_{E_i} \omega = -\sum_{i=1}^{n} (\nabla_{E_i} \omega)(E_i),$$

as stated earlier.

### 27.4 Harmonic Forms, the Hodge Theorem, Poincaré Duality

Let us now assume that $M$ is orientable and compact.

Definition 27.7. Let $M$ be an orientable and compact Riemannian manifold of dimension $n$. For every $k$, with $0 \leq k \leq n$, let

$$\mathbb{H}^k(M) = \{ \omega \in \mathcal{A}^k(M) \mid \Delta \omega = 0 \},$$

the space of harmonic $k$-forms.

The following proposition is left as an easy exercise:

Proposition 27.13. Let $M$ be an orientable and compact Riemannian manifold of dimension $n$. The Laplacian commutes with the Hodge $*$-operator, and we have a linear map

$$* : \mathbb{H}^k(M) \to \mathbb{H}^{n-k}(M).$$

One of the deepest and most important theorems about manifolds is the Hodge decomposition theorem, which we now state.

Theorem 27.14. (Hodge Decomposition Theorem) Let $M$ be an orientable and compact Riemannian manifold of dimension $n$. For every $k$, with $0 \leq k \leq n$, the space $\mathbb{H}^k(M)$ is finite dimensional, and we have the following orthogonal direct sum decomposition of the space of $k$-forms:

$$\mathcal{A}^k(M) = \mathbb{H}^k(M) \oplus d(\mathcal{A}^{k-1}(M)) \oplus \delta(\mathcal{A}^{k+1}(M)).$$
The proof of Theorem 27.14 involves a lot of analysis and it is long and complicated. A complete proof can be found in Warner [175] (Chapter 6). Other treatments of Hodge theory can be found in Morita [133] (Chapter 4) and Jost [99] (Chapter 2).

The Hodge Decomposition Theorem has a number of important corollaries, one of which is the Hodge Theorem:

**Theorem 27.15.** (Hodge Theorem) Let $M$ be an orientable and compact Riemannian manifold of dimension $n$. For every $k$, with $0 \leq k \leq n$, there is an isomorphism between $\mathbb{H}^k(M)$ and the de Rham cohomology vector space $H^k_{\text{DR}}(M)$:

$$H^k_{\text{DR}}(M) \cong \mathbb{H}^k(M).$$

**Proof.** Since by Proposition 27.7, every harmonic form $\omega \in \mathbb{H}^k(M)$ is closed, we get a linear map from $\mathbb{H}^k(M)$ to $H^k_{\text{DR}}(M)$ by assigning its cohomology class $[\omega]$ to $\omega$. This map is injective. Indeed, if $[\omega] = 0$ for some $\omega \in \mathbb{H}^k(M)$, then $\omega = d\eta$ for some $\eta \in \mathcal{A}^{k-1}(M)$ so

$$(\omega, \omega) = (d\eta, \omega) = (\eta, \delta \omega).$$

But, as $\omega \in \mathbb{H}^k(M)$ we have $\delta \omega = 0$ by Proposition 27.7, so $(\omega, \omega) = 0$; that is, $\omega = 0$.

Our map is also surjective. This is the hard part of Hodge Theorem. By the Hodge Decomposition Theorem, for every closed form $\omega \in \mathcal{A}^k(M)$, we can write

$$\omega = \omega_H + d\eta + \delta \theta,$$

with $\omega_H \in \mathbb{H}^k(M)$, $\eta \in \mathcal{A}^{k-1}(M)$, and $\theta \in \mathcal{A}^{k+1}(M)$. Since $\omega$ is closed and $\omega_H \in \mathbb{H}^k(M)$, we have $d\omega = 0$ and $d\omega_H = 0$, thus

$$d\delta \theta = 0$$

and so

$$0 = (d\delta \theta, \theta) = (\delta \theta, \delta \theta);$$

that is, $\delta \theta = 0$. Therefore, $\omega = \omega_H + d\eta$, which implies $[\omega] = [\omega_H]$ with $\omega_H \in \mathbb{H}^k(M)$, proving the surjectivity of our map. \hfill $\square$

The Hodge Theorem also implies the Poincaré Duality Theorem. If $M$ is a compact, orientable, $n$-dimensional smooth manifold, for each $k$, with $0 \leq k \leq n$, we define a bilinear map

$$((-,-)) : H^k_{\text{DR}}(M) \times H^{n-k}_{\text{DR}}(M) \longrightarrow \mathbb{R}$$

by setting

$$((-,[\omega],[\eta])) = \int_M \omega \wedge \eta.$$

We need to check that this definition does not depend on the choice of closed forms in the cohomology classes $[\omega]$ and $[\eta]$. However, as $d\omega = d\eta = 0$, we have

$$d(\alpha \wedge \eta + (-1)^k \omega \wedge \beta + \alpha \wedge d\beta) = d\alpha \wedge \eta + \omega \wedge d\beta + d\alpha \wedge d\beta,$$

This completes the proof of the Poincaré Duality Theorem.
so by Stokes’ Theorem,
\[
\int_M (\omega + d\alpha) \wedge (\eta + d\beta) = \int_M \omega \wedge \eta + \int_M d(\alpha \wedge \eta + (-1)^k \omega \wedge \beta + \alpha \wedge d\beta) = \int_M \omega \wedge \eta.
\]

**Theorem 27.16.** (Poincaré Duality) If \( M \) is a compact, orientable, smooth manifold of dimension \( n \), then the bilinear map
\[
((\cdot, \cdot)) : H^k_{\text{DR}}(M) \times H^{n-k}_{\text{DR}}(M) \to \mathbb{R}
\]
defined above is a nondegenerate pairing, and hence yields an isomorphism
\[
H^k_{\text{DR}}(M) \cong (H^{n-k}_{\text{DR}}(M))^*. 
\]

**Proof.** Pick any Riemannian metric on \( M \). It is enough to show that for every nonzero cohomology class \([\omega] \in H^k_{\text{DR}}(M)\), there is some \([\eta] \in H^{n-k}_{\text{DR}}(M)\) such that
\[
(([\omega],[\eta])) = \int_M \omega \wedge \eta \neq 0.
\]

By Hodge Theorem, we may assume that \( \omega \) is a nonzero harmonic form. By Proposition 27.13, \( \eta = \ast \omega \) is also harmonic and \( \eta \in H^{n-k}_{\text{DR}}(M) \). Then, we get
\[
(\omega, \omega) = \int_M \omega \wedge \ast \omega = (([\omega],[\eta])),
\]
and indeed, \((([\omega],[\eta])) \neq 0\), since \( \omega \neq 0 \).

\[\square\]

### 27.5 The Connection Laplacian, Weitzenböck Formula and the Bochner Technique

If \( M \) is compact, orientable, Riemannian manifold, then the inner product \( \langle \cdot, \cdot \rangle_p \) on \( T_p M \) (with \( p \in M \)) induces an inner product on differential forms, as we explained in Section 27.3. We also get an inner product on vector fields if, for any two vector field \( X, Y \in \mathfrak{X}(M) \), we define \( (X, Y) \) by
\[
(X, Y) = \int_M \langle X, Y \rangle \text{Vol}_M,
\]
where \( \langle X, Y \rangle \) is the function defined pointwise by
\[
\langle X, Y \rangle(p) = \langle X(p), Y(p) \rangle_p.
\]
Using Proposition 29.5, we can define the covariant derivative $\nabla_X\omega$ of any $k$-form $\omega \in \mathcal{A}^k(M)$ as the $k$-form given by

$$(\nabla_X\omega)(Y_1, \ldots, Y_k) = X(\omega(Y_1, \ldots, Y_k)) - \sum_{j=1}^k \omega(Y_1, \ldots, \nabla_XY_j, \ldots, Y_k).$$

We can view $\nabla$ as a linear map

$$\nabla: \mathcal{A}^k(M) \to \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)),$$

where $\nabla\omega$ is the $C^\infty(M)$-linear map $X \mapsto \nabla_X\omega$. The inner product on $\mathcal{A}^k(M)$ allows us to define the (formal) adjoint $\nabla^*$ of $\nabla$ as a linear map

$$\nabla^*: \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)) \to \mathcal{A}^k(M).$$

For any linear map $A \in \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$, let $A^*$ be the adjoint of $A$ defined by

$$(AX, \theta) = (X, A^*\theta),$$

for all vector fields $X \in \mathfrak{X}(M)$ and all $k$-forms $\theta \in \mathcal{A}^k(M)$. It can be verified that $A^* \in \text{Hom}_{C^\infty(M)}(\mathcal{A}^k(M), \mathfrak{X}(M))$. Then, given $A, B \in \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$, the expression $\text{tr}(A^*B)$ is a smooth function on $M$, and it can be verified that

$$\langle A, B \rangle = \text{tr}(A^*B)$$

defines a non-degenerate pairing on $\text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$. Using this pairing, we obtain the ($\mathbb{R}$-valued) inner product on $\text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ given by

$$(A, B) = \int_M \text{tr}(A^*B) \text{Vol}_M.$$ 

Using all this, the (formal) adjoint $\nabla^*$ of $\nabla: \mathcal{A}^k(M) \to \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ is the linear map $\nabla^*: \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M)) \to \mathcal{A}^k(M)$ defined implicitly by

$$(\nabla^*A, \omega) = (A, \nabla\omega);$$

that is,

$$\int_M \langle \nabla^*A, \omega \rangle \text{Vol}_M = \int_M \langle A, \nabla\omega \rangle \text{Vol}_M,$$

for all $A \in \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^k(M))$ and all $\omega \in \mathcal{A}^k(M)$. 

The notation $\nabla^*$ for the adjoint of $\nabla$ should not be confused with the dual connection on $T^*M$ of a connection $\nabla$ on $TM$! Here, $\nabla$ denotes the connection on $\mathcal{A}^*(M)$ induced by the orginal connection $\nabla$ on $TM$. The argument type (differential form or vector field) should make it clear which $\nabla$ is intended, but it might have been better to use a notation such as $\nabla^\top$ instead of $\nabla^*$. 


27.5. THE CONNECTION LAPLACIAN AND THE BOCHNER TECHNIQUE

What we just did also applies to $\mathcal{A}^*(M) = \bigoplus_{k=0}^n \mathcal{A}^k(M)$ (where $\dim(M) = n$), and so we can view the connection $\nabla$ as a linear map $\nabla: \mathcal{A}^*(M) \to \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$, and its adjoint as a linear map $\nabla^* : \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M)) \to \mathcal{A}^*(M)$.

**Definition 27.8.** Given a compact, orientable, Riemannian manifold, $M$, the connection Laplacian (or Bochner Laplacian) $\nabla^* \nabla$ is defined as the composition of the connection $\nabla: \mathcal{A}^*(M) \to \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M))$ with its adjoint $\nabla^* : \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathcal{A}^*(M)) \to \mathcal{A}^*(M)$, as defined above.

Observe that $(\nabla^* \nabla \omega, \omega) = (\nabla \omega, \nabla \omega) = \int_M \langle \nabla \omega, \nabla \omega \rangle \text{Vol}_M$, for all $\omega \in \mathcal{A}^k(M)$. Consequently, the “harmonic forms” $\omega$ with respect to $\nabla^* \nabla$ must satisfy

$$\nabla \omega = 0,$$

but this condition is not equivalent to the harmonicity of $\omega$ with respect to the Hodge Laplacian.

Thus, in general, $\nabla^* \nabla$ and $\Delta$ are different operators. The relationship between the two is given by formulae involving contractions of the curvature tensor, and are known as Weitzenböck formulae. We will state such a formula in case of one-forms later on. But first, we can give another definition of the connection Laplacian using second covariant derivatives of forms. Given any $k$-form $\omega \in \mathcal{A}^k(M)$, for any two vector fields $X, Y \in \mathfrak{X}(M)$, we define $\nabla^2_{X,Y} \omega$ by

$$\nabla^2_{X,Y} \omega = \nabla_X(\nabla_Y \omega) - \nabla_{\nabla_X Y} \omega.$$

Given any local chart $(U, \varphi)$ and given any orthonormal frame $(E_1, \ldots, E_n)$ over $U$, we can take the trace $\text{tr}(\nabla^2 \omega)$ of $\nabla^2_{X,Y} \omega$ defined by

$$\text{tr}(\nabla^2 \omega) = \sum_{i=1}^n \nabla^2_{E_i,E_i} \omega.$$

It is easily seen that $\text{tr}(\nabla^2 \omega)$ does not depend on the choice of local chart and orthonormal frame.

**Proposition 27.17.** If is $M$ a compact, orientable, Riemannian manifold, then the connection Laplacian $\nabla^* \nabla$ is given by

$$\nabla^* \nabla \omega = -\text{tr}(\nabla^2 \omega),$$

for all differential forms $\omega \in \mathcal{A}^*(M)$. 
The proof of Proposition 27.17, which is quite technical, can be found in Petersen [140] (Chapter 7, Proposition 34).

If \( \omega \in A^1(M) \) is a one-form, then the covariant derivative of \( \omega \) defines a \((0,2)\)-tensor \( T \) given by

\[
T(Y,Z) = \nabla_Y \omega(Z).
\]

Thus, we can define the second covariant derivative \( \nabla^2 \omega \) of \( \omega \) as the covariant derivative of \( T \) (see Proposition 29.5); that is,

\[
(\nabla^2 Y)(Z) = X((\nabla Y \omega)(Z)) - (\nabla^2 Y \omega)(Z) - (\nabla_Y \omega)(\nabla X Z).
\]

and so

\[
(\nabla^2_{X,Y} \omega)(Z) = X((\nabla Y \omega)(Z)) - (\nabla^2_{X,Y} \omega)(Z) - (\nabla_Y \omega)(\nabla X Z).
\]

Therefore,

\[
\nabla^2_{X,Y} \omega = \nabla_X(\nabla Y \omega) - \nabla_X(\nabla Y \omega);
\]

that is, \( \nabla^2_{X,Y} \omega \) is formally the same as \( \nabla^2_{X,Y} Z \). Then, it is natural to ask what is

\[
\nabla^2_{X,Y} \omega - \nabla^2_{Y,X} \omega.
\]

The answer is given by the following proposition which plays a crucial role in the proof of a version of Bochner’s formula:

**Proposition 27.18.** For any vector fields \( X,Y,Z \in \mathfrak{X}(M) \) and any one-form \( \omega \in A^1(M) \) on a Riemannian manifold \( M \), we have

\[
((\nabla^2_{X,Y} - \nabla^2_{Y,X}) \omega)(Z) = \omega(R(X,Y)Z).
\]

**Proof.** Recall that we proved in Section 29.5 that

\[
(\nabla_X \omega)^\sharp = \nabla \omega^\sharp.
\]

We claim that we also have

\[
(\nabla^2_{X,Y} \omega)^\sharp = \nabla^2_{X,Y} \omega^\sharp.
\]

This is because

\[
(\nabla^2_{X,Y} \omega)^\sharp = \nabla_X(\nabla_Y \omega)^\sharp - \nabla_X(\nabla Y \omega)^\sharp
\]

\[
= \nabla_X(\nabla_Y \omega)^\sharp - \nabla_X(\nabla Y \omega)^\sharp
\]

\[
= \nabla_X(\nabla_Y \omega)^\sharp - \nabla_X(\nabla Y \omega)^\sharp
\]

\[
= \nabla^2_{X,Y} \omega^\sharp.
\]

Thus, we deduce that

\[
((\nabla^2_{X,Y} - \nabla^2_{Y,X}) \omega)^\sharp = (\nabla^2_{X,Y} - \nabla^2_{Y,X}) \omega^\sharp = R(Y,X) \omega^\sharp.
\]
Consequently,

\[
((\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega)(Z) = \langle(\nabla^2_{X,Y} - \nabla^2_{Y,X})\omega, Z\rangle \\
= \langle R(Y, X)\omega^\sharp, Z\rangle \\
= R(Y, X, \omega^\sharp, Z) \\
= \langle R(X, Y)Z, \omega^\sharp\rangle \\
= \omega(R(X, Y)Z),
\]

where we used properties (3) and (4) of Proposition 13.2.

We are now ready to prove the Weitzenböck formulae for one-forms.

**Theorem 27.19.** (Weitzenböck–Bochner Formula) If is \(M\) a compact, orientable, Riemannian manifold, then for every one-form \(\omega \in \mathcal{A}^1(M)\), we have

\[
\Delta \omega = \nabla^* \nabla \omega + \text{Ric}(\omega),
\]

where \(\text{Ric}(\omega)\) is the one-form given by

\[
\text{Ric}(\omega)(X) = \omega(\text{Ric}^\sharp(X)),
\]

and where \(\text{Ric}^\sharp\) is the Ricci curvature viewed as a \((1, 1)\)-tensor (that is, \(\langle \text{Ric}^\sharp(u), v \rangle_p = \text{Ric}(u, v)\), for all \(u, v \in T_pM\) and all \(p \in M\)).

**Proof.** For any \(p \in M\), pick any normal local chart \((U, \varphi)\) with \(p \in U\), and pick any orthonormal frame \((E_1, \ldots, E_n)\) over \(U\). Because \((U, \varphi)\) is a normal chart at \(p\), we have \(\langle \nabla_{E_i}E_j \rangle_p = 0\) for all \(i, j\). Recall from the discussion at the end of Section 27.3 that for every one-form \(\omega\), we have

\[
\delta \omega = -\sum_i \nabla_{E_i}\omega(E_i),
\]

and so

\[
d\delta \omega = -\sum_i \nabla_X \nabla_{E_i}\omega(E_i).
\]

Also recall that

\[
d\omega(X, Y) = \nabla_X \omega(Y) - \nabla_Y \omega(X),
\]

and using Proposition 27.12, we can show that

\[
\delta d \omega(X) = -\sum_i \nabla_{E_i} \nabla_{E_i} \omega(X) + \sum_i \nabla_{E_i} \nabla_X \omega(E_i).
\]
Thus, we get

\[ \Delta \omega(X) = -\sum_i \nabla_{E_i} \nabla_{E_i} \omega(X) + \sum_i (\nabla_{E_i} \nabla_X - \nabla_X \nabla_{E_i}) \omega(E_i) \]

\[ = -\sum_i \nabla^2_{E_i, E_i} \omega(X) + \sum_i (\nabla^2_{E_i, X} - \nabla^2_{X, E_i}) \omega(E_i) \]

\[ = \nabla^* \nabla \omega(X) + \sum_i \omega(R(E_i, X)E_i) \]

\[ = \nabla^* \nabla \omega(X) + \omega(\text{Ric}^\sharp(X)), \]

using the fact that \((\nabla_{E_i} E_j)_p = 0\) for all \(i, j\), and using Proposition 27.18 and Proposition 27.17.

For simplicity of notation, we will write \(\text{Ric}(u)\) for \(\text{Ric}^\sharp(u)\). There should be no confusion, since \(\text{Ric}(u, v)\) denotes the Ricci curvature, a \((0, 2)\)-tensor. There is another way to express \(\text{Ric}(\omega)\) which will be useful in the proof of the next theorem. Observe that

\[ \text{Ric}(\omega)(Z) = \omega(\text{Ric}(Z)) = \langle \omega^\sharp, \text{Ric}(Z) \rangle = \langle \text{Ric}(Z), \omega^\sharp \rangle = \text{Ric}(Z, \omega^\sharp) = \text{Ric}(\omega^\sharp, Z) = \langle \text{Ric}(\omega^\sharp), Z \rangle = (\text{Ric}(\omega^\sharp))^\flat(Z), \]

and thus,

\[ \text{Ric}(\omega)(Z) = (\text{Ric}(\omega^\sharp))^\flat(Z). \]

Consequently the Weitzenböck formula can be written as

\[ \Delta \omega = \nabla^* \nabla \omega + (\text{Ric}(\omega^\sharp))^\flat. \]

The Weitzenböck–Bochner Formula implies the following theorem due to Bochner:

**Theorem 27.20.** *(Bochner)* If \(M\) is a compact, orientable, connected Riemannian manifold, then the following properties hold:

(i) If the Ricci curvature is non-negative, that is \(\text{Ric}(u, u) \geq 0\) for all \(p \in M\) and all \(u \in T_p M\), and if \(\text{Ric}(u, u) > 0\) for some \(p \in M\) and all \(u \in T_p M\), then \(H^1_{\text{DR}} M = (0)\).

(ii) If the Ricci curvature is non-negative, then \(\nabla \omega = 0\) for all \(\omega \in A^1(M)\), and \(\dim H^1_{\text{DR}} M \leq \dim M\).
27.5. THE CONNECTION LAPLACIAN AND THE BOCHNER TECHNIQUE

Proof. (i) Assume $H^1_{\text{DR}} M \neq (0)$. Then, by the Hodge Theorem there is some nonzero harmonic one-form $\omega$. The Weitzenböck–Bochner Formula implies that

$$ (\Delta \omega, \omega) = (\nabla^* \nabla \omega, \omega) + ((\text{Ric}(\omega^\sharp))^\flat, \omega). $$

Since $\Delta \omega = 0$, we get

$$ 0 = (\nabla^* \nabla \omega, \omega) + \int_M \langle (\text{Ric}(\omega^\sharp))^\flat, \omega \rangle \text{Vol}_M $$

$$ = (\nabla \omega, \nabla \omega) + \int_M \langle \text{Ric}(\omega^\sharp), \omega^\sharp \rangle \text{Vol}_M $$

$$ = (\nabla \omega, \nabla \omega) + \int_M \text{Ric}(\omega^\sharp, \omega^\sharp) \text{Vol}_M. $$

However, $(\nabla \omega, \nabla \omega) \geq 0$, and by the assumption on the Ricci curvature, the integrand is nonnegative and strictly positive at some point, so the integral is strictly positive, a contradiction.

(ii) Again, for any one-form $\omega$, we have

$$ (\Delta \omega, \omega) = (\nabla \omega, \nabla \omega) + \int_M \text{Ric}(\omega^\sharp, \omega^\sharp) \text{Vol}_M, $$

so if the Ricci curvature is non-negative, $\Delta \omega = 0$ iff $\nabla \omega = 0$. This means that $\omega$ is invariant by parallel transport (see Section 29.3), and thus $\omega$ is completely determined by its value $\omega_p$ at some point $p \in M$, so there is an injection $\mathbb{H}^1(M) \hookrightarrow T^*_p M$, which implies that $\dim H^1_{\text{DR}} M = \dim \mathbb{H}^1(M) \leq \dim M$. \hfill $\square$

There is a version of the Weitzenböck formula for $p$-forms, but it involves a more complicated curvature term and its proof is also more complicated. The Bochner technique can also be generalized in various ways, in particular, to spin manifolds, but these considerations are beyond the scope of these notes. Let me just say that Weitzenböck formulae involving the Dirac operator play an important role in physics and 4-manifold geometry. We refer the interested reader to Gallot, Hulin and Lafontaine [73] (Chapter 4) Petersen [140] (Chapter 7), Jost [99] (Chapter 3), and Berger [19] (Section 15.6), for more details on Weitzenböck formulae and the Bochner technique.
Chapter 28

Bundles, Metrics on Bundles, and Homogeneous Spaces, II

28.1 Fibre Bundles

We saw in Section 5.2 that a transitive action \( \cdot : G \times X \to X \) of a group \( G \) on a set \( X \) yields a description of \( X \) as a quotient \( G/G_x \), where \( G_x \) is the stabilizer of any element, \( x \in X \). In Theorem 5.14, we saw that if \( X \) is a “well-behaved” topological space, \( G \) is a “well-behaved” topological group, and the action is continuous, then \( G/G_x \) is homeomorphic to \( X \). In particular the conditions of Theorem 5.14 are satisfied if \( G \) is a Lie group and \( X \) is a manifold. Intuitively, the above theorem says that \( G \) can be viewed as a family of “fibres” \( G_x \), all isomorphic to \( G \), these fibres being parametrized by the “base space” \( X \), and varying smoothly when \( x \) moves in \( X \). We have an example of what is called a fibre bundle, in fact, a principal fibre bundle. Now that we know about manifolds and Lie groups, we can be more precise about this situation.

If \( G \) is a Lie group and if \( M \) is a manifold, an action \( \varphi : G \times M \to M \) is smooth if \( \varphi \) is smooth. Then, for every \( g \in G \), the map \( \varphi_g : M \to M \) is a diffeomorphism. The same definition applies to right actions.

Although we will not make extensive use of it, we begin by reviewing the definition of a fibre bundle because we believe that it clarifies the notions of vector bundles and principal fibre bundles, the concepts that are our primary concern. The following definition is not the most general, but it is sufficient for our needs:

Definition 28.1. A fibre bundle with (typical) fibre \( F \) and structure group \( G \) is a tuple \( \xi = (E, \pi, B, F, G) \), where \( E, B, F \) are smooth manifolds, \( \pi : E \to B \) is a smooth surjective map, \( G \) is a Lie group of diffeomorphisms of \( F \), and there is some open cover \( \mathcal{U} = (U_a)_{a \in I} \) of \( B \) and a family \( \varphi = (\varphi_a)_{a \in I} \) of diffeomorphisms

\[
\varphi_a : \pi^{-1}(U_a) \to U_a \times F.
\]
The space $B$ is called the base space, $E$ is called the total space, $F$ is called the (typical) fibre, and each $\varphi_\alpha$ is called a (local) trivialization. The pair $(U_\alpha, \varphi_\alpha)$ is called a bundle chart, and the family $\{(U_\alpha, \varphi_\alpha)\}$ is a trivializing cover. For each $b \in B$, the space $\pi^{-1}(b)$ is called the fibre above $b$; it is also denoted by $E_b$, and $\pi^{-1}(U_\alpha)$ is also denoted by $E \upharpoonright U_\alpha$. Furthermore, the following properties hold:

(a) The diagram

$$
\begin{array}{ccc}
\pi^{-1}(U_\alpha) & \xrightarrow{\varphi_\alpha} & U_\alpha \times F \\
\downarrow{\pi} & & \downarrow{p_1} \\
U_\alpha & & \end{array}
$$

commutes for all $\alpha \in I$, where $p_1: U_\alpha \times F \to U_\alpha$ is the first projection. Equivalently, for all $(b, y) \in U_\alpha \times F$,

$$\pi \circ \varphi^{-1}_\alpha(b, y) = b.$$ 

For every $(U_\alpha, \varphi_\alpha)$ and every $b \in U_\alpha$, because $p_1 \circ \varphi_\alpha = \pi$, by (a) the restriction of $\varphi_\alpha$ to $E_b = \pi^{-1}(b)$ is a diffeomorphism between $E_b$ and $\{b\} \times F$, so we have the diffeomorphism

$$\varphi_{\alpha,b}: E_b \to F$$

given by

$$\varphi_{\alpha,b}(Z) = (p_2 \circ \varphi_\alpha)(Z), \quad \text{for all } Z \in E_b.$$ 

Furthermore, for all $U_\alpha, U_\beta$ in $\mathcal{U}$ such that $U_\alpha \cap U_\beta \neq \emptyset$, for every $b \in U_\alpha \cap U_\beta$, there is a relationship between $\varphi_{\alpha,b}$ and $\varphi_{\beta,b}$ which gives the twisting of the bundle:

(b) The diffeomorphism

$$\varphi_{\alpha,b} \circ \varphi_{\beta,b}^{-1}: F \to F$$

is an element of the group $G$.

(c) The map $g_{\alpha\beta}: U_\alpha \cap U_\beta \to G$ defined by

$$g_{\alpha\beta}(b) = \varphi_{\alpha,b} \circ \varphi_{\beta,b}^{-1},$$

is smooth. The maps $g_{\alpha\beta}$ are called the transition maps of the fibre bundle.

A fibre bundle $\xi = (E, \pi, B, F, G)$ is also referred to, somewhat loosely, as a fibre bundle over $B$ or a $G$-bundle, and it is customary to use the notation

$$F \to E \to B,$$

or

$$F \to E \to B,$$
28.1. FIBRE BUNDLES

even though it is imprecise (the group $G$ is missing!), and it clashes with the notation for short exact sequences. Observe that the bundle charts $(U_\alpha, \varphi_\alpha)$ are similar to the charts of a manifold.

Actually, Definition 28.1 is too restrictive because it does not allow for the addition of compatible bundle charts, for example when considering a refinement of the cover $\mathcal{U}$. This problem can easily be fixed using a notion of equivalence of trivializing covers analogous to the equivalence of atlases for manifolds (see Remark (2) below). Also, observe that (b) and (c) imply that the isomorphism $\varphi_\alpha \circ \varphi_\beta^{-1} : (U_\alpha \cap U_\beta) \times F \to (U_\alpha \cap U_\beta) \times F$ is related to the smooth map $g_{\alpha\beta} : U_\alpha \cap U_\beta \to G$ by the identity

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x)),$$

for all $b \in U_\alpha \cap U_\beta$ and all $x \in F$.

Note that the isomorphism $\varphi_\alpha \circ \varphi_\beta^{-1} : (U_\alpha \cap U_\beta) \times F \to (U_\alpha \cap U_\beta) \times F$ describes how the fibres viewed over $U_\beta$ are viewed over $U_\alpha$. Thus, it might have been better to denote $g_{\alpha\beta}$ by $g_{\beta\alpha}$, so that

$$g_{\beta\alpha} = \varphi_{\beta,b} \circ \varphi_{\alpha,b}^{-1},$$

where the subscript $\alpha$ indicates the source, and the superscript $\beta$ indicates the target.

Intuitively, a fibre bundle over $B$ is a family $E = (E_b)_{b \in B}$ of spaces $E_b$ (fibres) indexed by $B$ and varying smoothly as $b$ moves in $B$, such that every $E_b$ is diffeomorphic to $F$. The bundle $E = B \times F$, where $\pi$ is the first projection, is called the trivial bundle (over $B$). The trivial bundle $B \times F$ is often denoted $\epsilon^F$. The local triviality condition (a) says that locally, that is over every subset $U_\alpha$ from some open cover of the base space $B$, the bundle $\xi \mid U_\alpha$ is trivial. Note that if $G$ is the trivial one-element group, then the fibre bundle is trivial. In fact, the purpose of the group $G$ is to specify the “twisting” of the bundle; that is, how the fibre $E_b$ gets twisted as $b$ moves in the base space $B$.

A Möbius strip is an example of a nontrivial fibre bundle where the base space $B$ is the circle $S^1$, the fibre space $F$ is the closed interval $[-1, 1]$, and the structural group is $G = \{1, -1\}$, where $-1$ is the reflection of the interval $[-1, 1]$ about its midpoint, 0. The total space $E$ is the strip obtained by rotating the line segment $[-1, 1]$ around the circle, keeping its midpoint in contact with the circle, and gradually twisting the line segment so that after a full revolution, the segment has been tilted by $\pi$. The reader should work out the transition functions for an open cover consisting of two open intervals on the circle.

A Klein bottle is also a fibre bundle for which both the base space and the fibre are the circle, $S^1$. Again, the reader should work out the details for this example.

Other examples of fibre bundles are:

(1) $\text{SO}(n+1)$, an $\text{SO}(n)$-bundle over the sphere $S^n$ with fibre $\text{SO}(n)$. (for $n \geq 0$).

(2) $\text{SU}(n+1)$, an $\text{SU}(n)$-bundle over the sphere $S^{2n+1}$ with fibre $\text{SU}(n)$ (for $n \geq 0$).
(3) $\text{SL}(2, \mathbb{R})$, an $\text{SO}(2)$-bundle over the upper-half space $H$, with fibre $\text{SO}(2)$.

(4) $\text{GL}(n, \mathbb{R})$, an $\text{O}(n)$-bundle over the space $\text{SPD}(n)$ of symmetric, positive definite matrices, with fibre $\text{O}(n)$.

(5) $\text{GL}^+(n, \mathbb{R})$, an $\text{SO}(n)$-bundle over the space, $\text{SPD}(n)$ of symmetric, positive definite matrices, with fibre $\text{SO}(n)$.

(6) $\text{SO}(n+1)$, an $\text{O}(n)$-bundle over the real projective space $\mathbb{R}P^n$, with fibre $\text{O}(n)$ (for $n \geq 0$).

(7) $\text{SU}(n+1)$, an $\text{U}(n)$-bundle over the complex projective space $\mathbb{C}P^n$, with fibre $\text{U}(n)$ (for $n \geq 0$).

(8) $\text{O}(n)$, an $\text{O}(k) \times \text{O}(n-k)$-bundle over the Grassmannian $G(k, n)$, with fibre $\text{O}(k) \times \text{O}(n-k)$.

(9) $\text{SO}(n)$ an $S(\text{O}(k) \times \text{O}(n-k))$-bundle over the Grassmannian $G(k, n)$, with fibre $S(\text{O}(k) \times \text{O}(n-k))$.

(10) $\text{SO}(n)$ is a principal $\text{SO}(n-k)$-bundle over the Stiefel manifold $S(k, n)$, with $1 \leq k \leq n-1$.

(11) From Section 5.4, we see that the Lorentz group, $\text{SO}_0(n, 1)$, is an $\text{SO}(n)$-bundle over the space $\mathcal{H}^+_n(1)$ consisting of one sheet of the hyperbolic paraboloid $\mathcal{H}_n(1)$, with fibre $\text{SO}(n)$.

Observe that in all the examples above, $F = G$; that is, the typical fibre is identical to the group $G$. Special bundles of this kind are called principal fibre bundles.

Remarks:

(1) The above definition is slightly different (but equivalent) to the definition given in Bott and Tu [24], page 47-48. Definition 28.1 is closer to the one given in Hirzebruch [92]. Bott and Tu and Hirzebruch assume that $G$ acts effectively on the left on the fibre $F$. This means that there is a smooth action $\cdot : G \times F \to F$, and recall that $G$ acts effectively on $F$ iff for every $g \in G$,

$$\text{if } g \cdot x = x \text{ for all } x \in F, \text{ then } g = 1.$$ 

Every $g \in G$ induces a diffeomorphism $\varphi_g : F \to F$, defined by

$$\varphi_g(x) = g \cdot x, \text{ for all } x \in F.$$ 

The fact that $G$ acts effectively on $F$ means that the map $g \mapsto \varphi_g$ is injective. This justifies viewing $G$ as a group of diffeomorphisms of $F$, and from now on we will denote $\varphi_g(x)$ by $g(x)$. 
(2) We observed that Definition 28.1 is too restrictive because it does not allow for the addition of compatible bundle charts. We can fix this problem as follows: Given a trivializing cover \( \{(U_\alpha, \varphi_\alpha)\} \), for any open \( U \) of \( B \) and any diffeomorphism

\[
\varphi : \pi^{-1}(U) \to U \times F,
\]

we say that \((U, \varphi)\) is compatible with the trivializing cover \( \{(U_\alpha, \varphi_\alpha)\} \) iff whenever \( U \cap U_\alpha \neq \emptyset \), there is some smooth map \( g_\alpha : U \cap U_\alpha \to G \), so that

\[
\varphi \circ \varphi_\alpha^{-1}(b, x) = (b, g_\alpha(b)(x)),
\]

for all \( b \in U \cap U_\alpha \) and all \( x \in F \). Two trivializing covers are equivalent iff every bundle chart of one cover is compatible with the other cover. This is equivalent to saying that the union of two trivializing covers is a trivializing cover. Then, we can define a fibre bundle as a tuple \((E, \pi, B, F, G, \{(U_\alpha, \varphi_\alpha)\})\), where \( \{(U_\alpha, \varphi_\alpha)\} \) is an equivalence class of trivializing covers. As for manifolds, given a trivializing cover \( \{(U_\alpha, \varphi_\alpha)\} \), the set of all bundle charts compatible with \( \{(U_\alpha, \varphi_\alpha)\} \) is a maximal trivializing cover equivalent to \( \{(U_\alpha, \varphi_\alpha)\} \).

A special case of the above occurs when we have a trivializing cover \( \{(U_\alpha, \varphi_\alpha)\} \) with \( U = \{U_\alpha\} \) an open cover of \( B \), and another open cover \( V = (V_\beta)_{\beta \in I} \) of \( B \) which is a refinement of \( U \). This means that there is a map \( \tau : I \to J \), such that \( V_\beta \subseteq U_{\tau(\beta)} \) for all \( \beta \in J \). Then, for every \( V_\beta \in V \), since \( V_\beta \subseteq U_{\tau(\beta)} \), the restriction of \( \varphi_{\tau(\beta)} \) to \( V_\beta \) is a trivialization

\[
\varphi'_\beta : \pi^{-1}(V_\beta) \to V_\beta \times F,
\]

and conditions (b) and (c) are still satisfied, so \((V_\beta, \varphi'_\beta)\) is compatible with \( \{(U_\alpha, \varphi_\alpha)\} \).

(3) (For readers familiar with sheaves) Hirzebruch defines the sheaf \( G_\infty \), where \( \Gamma(U, G_\infty) \) is the group of smooth functions \( g : U \to G \), where \( U \) is some open subset of \( B \) and \( G \) is a Lie group acting effectively (on the left) on the fibre \( F \). The group operation on \( \Gamma(U, G_\infty) \) is induced by multiplication in \( G \); that is, given two (smooth) functions \( g : U \to G \) and \( h : U \to G \),

\[
gh(b) = g(b)h(b), \quad \text{for all} \ b \in U.
\]

Beware that \( gh \) is not function composition, unless \( G \) itself is a group of functions, which is the case for vector bundles.

Our conditions (b) and (c) are then replaced by the following equivalent condition: For all \( U_\alpha, U_\beta \) in \( U \) such that \( U_\alpha \cap U_\beta \neq \emptyset \), there is some \( g_{\alpha\beta} \in \Gamma(U_\alpha \cap U_\beta, G_\infty) \) such that

\[
\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x)),
\]

for all \( b \in U_\alpha \cap U_\beta \) and all \( x \in F \).
(4) The family of transition functions \((g_{\alpha\beta})\) satisfies the \textit{cocycle condition}

\[ g_{\alpha\beta}(b)g_{\beta\gamma}(b) = g_{\alpha\gamma}(b), \]

for all \(\alpha, \beta, \gamma\) such that \(U_\alpha \cap U_\beta \cap U_\gamma \neq \emptyset\) and all \(b \in U_\alpha \cap U_\beta \cap U_\gamma\).

Setting \(\alpha = \beta = \gamma\), we get

\[ g_{\alpha\alpha} = \text{id}, \]

and setting \(\gamma = \alpha\), we get

\[ g_{\beta\alpha} = g_{\alpha\beta}^{-1}. \]

Again, beware that this means that \(g_{\beta\alpha}(b) = g_{\alpha\beta}^{-1}(b)\), where \(g_{\alpha\beta}^{-1}(b)\) is the inverse of \(g_{\beta\alpha}(b)\) in \(G\). In general, \(g_{\alpha\beta}^{-1}\) is \textbf{not} the functional inverse of \(g_{\beta\alpha}\).

The classic source on fibre bundles is Steenrod [164]. The most comprehensive treatment of fibre bundles and vector bundles is probably given in Husemoller [97]. However, we can hardly recommend this book. We find the presentation overly formal and intuitions are absent. A more extensive list of references is given at the end of Section 28.5.

\textbf{Remark:} (The following paragraph is intended for readers familiar with Čech cohomology.) The cocycle condition makes it possible to view a fibre bundle over \(B\) as a member of a certain (Čech) cohomology set \(\tilde{H}^1(B, G)\), where \(G\) denotes a certain sheaf of functions from the manifold \(B\) into the Lie group \(G\), as explained in Hirzebruch [92], Section 3.2. However, this requires defining a noncommutative version of Čech cohomology (at least, for \(\tilde{H}^1\)), and clarifying when two open covers and two trivializations define the same fibre bundle over \(B\), or equivalently, defining when two fibre bundles over \(B\) are equivalent. If the bundles under considerations are line bundles (see Definition 28.6), then \(\tilde{H}^1(B, G)\) is actually a group. In this case, \(G = \text{GL}(1, \mathbb{R}) \cong \mathbb{R}^*\) in the real case, and \(G = \text{GL}(1, \mathbb{C}) \cong \mathbb{C}^*\) in the complex case (where \(\mathbb{R}^* = \mathbb{R} - \{0\}\) and \(\mathbb{C}^* = \mathbb{C} - \{0\}\)), and the sheaf \(G\) is the sheaf of smooth (real-valued or complex-valued) functions vanishing nowhere. The group \(\tilde{H}^1(B, G)\) plays an important role, especially when the bundle is a holomorphic line bundle over a complex manifold. In the latter case, it is called the \textit{Picard group} of \(B\).

The notion of a map between fibre bundles is more subtle than one might think because of the structure group \(G\). Let us begin with the simpler case where \(G = \text{Diff}(F)\), the group of all smooth diffeomorphisms of \(F\).

\textbf{Definition 28.2.} If \(\xi_1 = (E_1, \pi_1, B_1, F, \text{Diff}(F))\) and \(\xi_2 = (E_2, \pi_2, B_2, F, \text{Diff}(F))\) are two fibre bundles with the same typical fibre \(F\) and the same structure group \(G = \text{Diff}(F)\), a \textit{bundle map} (or \textit{bundle morphism}) \(f: \xi_1 \to \xi_2\) is a pair \(f = (f_E, f_B)\) of smooth maps \(f_E: E_1 \to E_2\) and \(f_B: B_1 \to B_2\), such that:
(a) The following diagram commutes:

\[
\begin{array}{ccc}
E_1 & \xrightarrow{f_E} & E_2 \\
\downarrow{\pi_1} & & \downarrow{\pi_2} \\
B_1 & \xrightarrow{f_B} & B_2
\end{array}
\]

(b) For every \( b \in B_1 \), the map of fibres

\[ f_E \upharpoonright \pi_1^{-1}(b) : \pi_1^{-1}(b) \to \pi_2^{-1}(f_B(b)) \]

is a diffeomorphism (preservation of the fibre).

A bundle map \( f : \xi_1 \to \xi_2 \) is an isomorphism if there is some bundle map \( g : \xi_2 \to \xi_1 \), called the inverse of \( f \), such that

\[ g_E \circ f_E = \text{id} \quad \text{and} \quad f_E \circ g_E = \text{id}. \]

The bundles \( \xi_1 \) and \( \xi_2 \) are called isomorphic. Given two fibre bundles \( \xi_1 = (E_1, \pi_1, B, F, G) \) and \( \xi_2 = (E_2, \pi_2, B, F, G) \) over the same base space \( B \), a bundle map (or bundle morphism) \( f : \xi_1 \to \xi_2 \) is a pair \( f = (f_E, f_B) \), where \( f_B = \text{id} \) (the identity map). Such a bundle map is an isomorphism if it has an inverse as defined above. In this case, we say that the bundles \( \xi_1 \) and \( \xi_2 \) over \( B \) are isomorphic.

Observe that the commutativity of the diagram in Definition 28.2 implies that \( f_B \) is actually determined by \( f_E \). Also, when \( f \) is an isomorphism, the surjectivity of \( \pi_1 \) and \( \pi_2 \) implies that

\[ g_B \circ f_B = \text{id} \quad \text{and} \quad f_B \circ g_B = \text{id}. \]

Thus, when \( f = (f_E, f_B) \) is an isomorphism, both \( f_E \) and \( f_B \) are diffeomorphisms.

Remark: Some authors do not require the “preservation” of fibres. However, it is automatic for bundle isomorphisms.

When we have a bundle map \( f : \xi_1 \to \xi_2 \) as above, for every \( b \in B \), for any trivializations \( \varphi_\alpha : \pi_1^{-1}(U_\alpha) \to U_\alpha \times F \) of \( \xi_1 \) and \( \varphi'_\beta : \pi_2^{-1}(V_\beta) \to V_\beta \times F \) of \( \xi_2 \), with \( b \in U_\alpha \) and \( f_B(b) \in V_\beta \), we have the map

\[ \varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1} : (U_\alpha \cap f_B^{-1}(V_\beta)) \times F \to V_\beta \times F. \]

Consequently, as \( \varphi_\alpha \) and \( \varphi'_\beta \) are diffeomorphisms and as \( f \) is a diffeomorphism on fibres, we have a map \( \rho_{\alpha,\beta} : U_\alpha \cap f_B^{-1}(V_\beta) \to \text{Diff}(F) \), such that

\[ \varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}(b, x) = (f_B(b), \rho_{\alpha,\beta}(b)(x)), \]

for all \( b \in U_\alpha \cap f_B^{-1}(V_\beta) \) and all \( x \in F \). Unfortunately, in general, there is no guarantee that \( \rho_{\alpha,\beta}(b) \in G \) or that it be smooth. However, this will be the case when \( \xi \) is a vector bundle or a principal bundle.
Since we may always pick \( U_\alpha \) and \( V_\beta \) so that \( f_B(U_\alpha) \subseteq V_\beta \), we may also write \( \rho_\alpha \) instead of \( \rho_{\alpha,\beta} \), with \( \rho_\alpha: U_\alpha \to G \). Then, observe that locally, \( f_E \) is given as the composition

\[
\pi_1^{-1}(U_\alpha) \xrightarrow{\varphi} U_\alpha \times F \xrightarrow{\tilde{f}_\alpha} V_\beta \times F \xrightarrow{\varphi_\beta^{-1}} \pi_2^{-1}(V_\beta)
\]

\[
z \xrightarrow{} (b, x) \xrightarrow{} (f_B(b), \rho_\alpha(b)(x)) \xrightarrow{\varphi_\beta^{-1}(f_B(b), \rho_\alpha(b)(x))),
\]

with \( \tilde{f}_\alpha(b, x) = (f_B(b), \rho_\alpha(b)(x)) \), that is,

\[
f_E(z) = \varphi_\beta^{-1}(f_B(b), \rho_\alpha(b)(x)), \quad \text{with } z \in \pi_1^{-1}(U_\alpha) \text{ and } (b, x) = \varphi_\alpha(z).
\]

Conversely, if \( (f_E, f_B) \) is a pair of smooth maps satisfying the commutative diagram of Definition 28.2 and the above conditions hold locally, then \( \varphi_\alpha, \varphi_\beta^{-1}, \) and \( \rho_\alpha(b) \) are diffeomorphisms on fibres, we see that \( f_E \) is a diffeomorphism on fibres.

In the general case where the structure group \( G \) is not the whole group of diffeomorphisms \( \text{Diff}(F) \), following Hirzebruch [92], we use the local conditions above to define the “right notion” of bundle map, namely Definition 28.3. Another advantage of this definition is that two bundles (with the same fibre, structure group, and base) are isomorphic iff they are equivalent (see Proposition 28.1 and Proposition 28.2).

**Definition 28.3.** Given two fibre bundles \( \xi_1 = (E_1, \pi_1, B_1, F, G) \) and \( \xi_2 = (E_2, \pi_2, B_2, F, G) \), a bundle map \( f: \xi_1 \to \xi_2 \) is a pair \( f = (f_E, f_B) \) of smooth maps \( f_E: E_1 \to E_2 \) and \( f_B: B_1 \to B_2 \), such that:

(a) The diagram

\[
\begin{array}{ccc}
E_1 & \xrightarrow{f_E} & E_2 \\
\pi_1 \downarrow & & \downarrow \pi_2 \\
B_1 & \xrightarrow{f_B} & B_2
\end{array}
\]

commutes.

(b) There is an open cover \( U = (U_\alpha)_{\alpha \in I} \) for \( B_1 \), an open cover \( V = (V_\beta)_{\beta \in J} \) for \( B_2 \), a family \( \varphi = (\varphi_\alpha)_{\alpha \in I} \) of trivializations \( \varphi_\alpha: \pi_1^{-1}(U_\alpha) \to U_\alpha \times F \) for \( \xi_1 \), a family \( \varphi' = (\varphi'_\beta)_{\beta \in J} \) of trivializations \( \varphi'_\beta: \pi_2^{-1}(V_\beta) \to V_\beta \times F \) for \( \xi_2 \), such that for every \( b \in B \), there are some trivializations \( \varphi_\alpha: \pi_1^{-1}(U_\alpha) \to U_\alpha \times F \) and \( \varphi'_\beta: \pi_2^{-1}(V_\beta) \to V_\beta \times F \), with \( f_B(U_\alpha) \subseteq V_\beta \), \( b \in U_\alpha \) and some smooth map \( \rho_\alpha: U_\alpha \to G \), such that \( \varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}: U_\alpha \times F \to V_\alpha \times F \) is given by

\[
\varphi'_\beta \circ f_E \circ \varphi_\alpha^{-1}(b, x) = (f_B(b), \rho_\alpha(b)(x)),
\]

for all \( b \in U_\alpha \) and all \( x \in F \).
A bundle map is an \textit{isomorphism} if it has an inverse as in Definition 28.2. If the bundles $\xi_1$ and $\xi_2$ are over the same base $B$, then we also require $f_B = \text{id}$.

As we remarked in the discussion before Definition 28.3, condition (b) insures that the maps of fibres

$$f_E \upharpoonright \pi_1^{-1}(b): \pi_1^{-1}(b) \to \pi_2^{-1}(f_B(b))$$

are diffeomorphisms. In the special case where $\xi_1$ and $\xi_2$ have the same base, $B_1 = B_2 = B$, we require $f_B = \text{id}$, and we can use the same cover (\textit{i.e.,} $\mathcal{U} = \mathcal{V}$), in which case condition (b) becomes: There is some smooth map $\rho: U_\alpha \to G$, such that

$$\varphi'_\alpha \circ f \circ \varphi_\alpha^{-1}(b, x) = (b, \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$.

We say that a bundle $\xi$ with base $B$ and structure group $G$ is trivial iff $\xi$ is isomorphic to the product bundle $B \times F$, according to the notion of isomorphism of Definition 28.3.

We can also define the notion of equivalence for fibre bundles over the same base space $B$ (see Hirzebruch [92], Section 3.2, Chern [40], Section 5, and Husemoller [97], Chapter 5). We will see shortly that two bundles over the same base are equivalent iff they are isomorphic.

\textbf{Definition 28.4.} Given two fibre bundles $\xi_1 = (E_1, \pi_1, B, F, G)$ and $\xi_2 = (E_2, \pi_2, B, F, G)$ over the same base space $B$, we say that $\xi_1$ and $\xi_2$ are \textit{equivalent} if there is an open cover $\mathcal{U} = (U_\alpha)_{\alpha \in I}$ for $B$, a family $\varphi = (\varphi_\alpha)_{\alpha \in I}$ of trivializations $\varphi_\alpha: \pi_1^{-1}(U_\alpha) \to U_\alpha \times F$ for $\xi_1$, a family $\varphi' = (\varphi'_\alpha)_{\alpha \in I}$ of trivializations $\varphi'_\alpha: \pi_2^{-1}(U_\alpha) \to U_\alpha \times F$ for $\xi_2$, and a family $(\rho_\alpha)_{\alpha \in I}$ of smooth maps $\rho_\alpha: U_\alpha \to G$, such that

$$g'_{\alpha\beta}(b) = \rho_\alpha(b)g_{\alpha\beta}(b)\rho_\beta(b)^{-1}, \quad \text{for all } b \in U_\alpha \cap U_\beta.$$

Since the trivializations are bijections, the family $(\rho_\alpha)_{\alpha \in I}$ is unique. The following proposition shows that isomorphic fibre bundles are equivalent:

\textbf{Proposition 28.1.} If two fibre bundles $\xi_1 = (E_1, \pi_1, B, F, G)$ and $\xi_2 = (E_2, \pi_2, B, F, G)$ over the same base space $B$ are isomorphic, then they are equivalent.

\textit{Proof.} Let $f: \xi_1 \to \xi_2$ be a bundle isomorphism. Then, we know that for some suitable open cover of the base $B$, and some trivializing families $(\varphi_\alpha)$ for $\xi_1$ and $(\varphi'_\alpha)$ for $\xi_2$, there is a family of maps $\rho_\alpha: U_\alpha \to G$, so that

$$\varphi'_\alpha \circ f \circ \varphi_\alpha^{-1}(b, x) = (b, \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$. Recall that

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x)),$$

for all $b \in U_\alpha \cap U_\beta$ and all $x \in F$. This is equivalent to

$$\varphi_\alpha^{-1}(b, x) = \varphi_\beta^{-1}(b, g_{\alpha\beta}(b)(x)),$$
so it is notationally advantageous to introduce \( \psi_\alpha \) such that \( \psi_\alpha = \varphi_\alpha^{-1} \). Then, we have

\[
\psi_\beta(b, x) = \psi_\alpha(b, g_{\alpha\beta}(b)(x)),
\]

and

\[
\varphi'_\alpha \circ f \circ \varphi_\alpha^{-1}(b, x) = (b, \rho_\alpha(b)(x))
\]

becomes

\[
\psi_\alpha(b, x) = f^{-1} \circ \psi'_\alpha(b, \rho_\alpha(b)(x)).
\]

We have

\[
\psi_\beta(b, x) = \psi_\alpha(b, g_{\alpha\beta}(b)(x)) = f^{-1} \circ \psi'_\alpha(b, \rho_\alpha(b)(g_{\alpha\beta}(b)(x))),
\]

and also

\[
\psi_\beta(b, x) = f^{-1} \circ \psi'_\beta(b, \rho_\beta(b)(x)) = f^{-1} \circ \psi'_\alpha(b, g'_{\alpha\beta}(b)(\rho_\beta(b)(x)));
\]

from which we deduce

\[
\rho_\alpha(b)(g_{\alpha\beta}(b)(x)) = g'_{\alpha\beta}(b)(\rho_\beta(b)(x)),
\]

that is

\[
g'_{\alpha\beta}(b) = \rho_\alpha(b)g_{\alpha\beta}(b)\rho_\beta(b)^{-1}, \quad \text{for all } b \in U_\alpha \cap U_\beta,
\]

as claimed.

\[ \square \]

Remark: If \( \xi_1 = (E_1, \pi_1, B_1, F, G) \) and \( \xi_2 = (E_2, \pi_2, B_2, F, G) \) are two bundles over different bases and \( f: \xi_1 \to \xi_2 \) is a bundle isomorphism, with \( f = (f_B, f_E) \), then \( f_E \) and \( f_B \) are diffeomorphisms, and it is easy to see that we get the conditions

\[
g'_{\alpha\beta}(f_B(b)) = \rho_\alpha(b)g_{\alpha\beta}(b)\rho_\beta(b)^{-1}, \quad \text{for all } b \in U_\alpha \cap U_\beta.
\]

The converse of Proposition 28.1 also holds.

**Proposition 28.2.** If two fibre bundles \( \xi_1 = (E_1, \pi_1, B, F, G) \) and \( \xi_2 = (E_2, \pi_2, B, F, G) \) over the same base space \( B \) are equivalent, then they are isomorphic.

**Proof.** Assume that \( \xi_1 \) and \( \xi_2 \) are equivalent. Then, for some suitable open cover of the base \( B \) and some trivializing families \( (\varphi_\alpha) \) for \( \xi_1 \) and \( (\varphi'_\alpha) \) for \( \xi_2 \), there is a family of maps \( \rho_\alpha: U_\alpha \to G \), so that

\[
g'_{\alpha\beta}(b) = \rho_\alpha(b)g_{\alpha\beta}(b)\rho_\beta(b)^{-1}, \quad \text{for all } b \in U_\alpha \cap U_\beta,
\]

which can be written as

\[
g'_{\alpha\beta}(b)\rho_\beta(b) = \rho_\alpha(b)g_{\alpha\beta}(b).
\]
28.1. FIBRE BUNDLES

For every $U_\alpha$, define $f_\alpha$ as the composition

$$\pi_1^{-1}(U_\alpha) \xrightarrow{\varphi_\alpha} U_\alpha \times F \xrightarrow{\tilde{f}_\alpha} U_\alpha \times F \xrightarrow{\varphi_\alpha^{-1}} \pi_2^{-1}(U_\alpha)$$

that is,

$$f_\alpha(z) = \varphi_\alpha^{-1}(b, \rho_\alpha(b)(x)),$$

with $z \in \pi_1^{-1}(U_\alpha)$ and $(b, x) = \varphi_\alpha(z)$.

Clearly, the definition of $f_\alpha$ implies that

$$\varphi'_\alpha \circ f_\alpha \circ \varphi_\alpha^{-1}(b, x) = (b, \rho_\alpha(b)(x)),$$

for all $b \in U_\alpha$ and all $x \in F$, and locally $f_\alpha$ is a bundle isomorphism with respect to $\rho_\alpha$. If we can prove that any two $f_\alpha$ and $f_\beta$ agree on the overlap $U_\alpha \cap U_\beta$, then the $f_\alpha$’s patch and yield a bundle map between $\xi_1$ and $\xi_2$.

Now, on $U_\alpha \cap U_\beta$,

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha\beta}(b)(x))$$

yields

$$\varphi_\beta^{-1}(b, x) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)(x)).$$

We need to show that for every $z \in U_\alpha \cap U_\beta$,

$$f_\alpha(z) = \varphi_\alpha^{-1}(b, \rho_\alpha(b)(x)) = \varphi_\beta^{-1}(b, \rho_\beta(b)(x')) = f_\beta(z),$$

where $\varphi_\alpha(z) = (b, x)$ and $\varphi_\beta(z) = (b, x')$.

From $z = \varphi_\beta^{-1}(b, x') = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)(x'))$, we get

$$x = g_{\alpha\beta}(b)(x').$$

We also have

$$\varphi_\beta^{-1}(b, \rho_\beta(b)(x')) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}'(b)(\rho_\beta(b)(x'))),$$

and since $g_{\alpha\beta}'(b)(\rho_\beta(b)) = \rho_\alpha(b)g_{\alpha\beta}(b)$ and $x = g_{\alpha\beta}(b)(x')$, we get

$$\varphi_\beta^{-1}(b, \rho_\beta(b)(x')) = \varphi_\alpha^{-1}(b, \rho_\alpha(b)(g_{\alpha\beta}(b))(x')) = \varphi_\alpha^{-1}(b, \rho_\alpha(b)(x)).$$

as desired. Therefore, the $f_\alpha$’s patch to yield a bundle map $f$, with respect to the family of maps $\rho_\alpha: U_\alpha \to G$.

The map $f$ is bijective because it is an isomorphism on fibres, but it remains to show that it is a diffeomorphism. This is a local matter, and as the $\varphi_\alpha$ and $\varphi_\alpha'$ are diffeomorphisms, it suffices to show that the map $\tilde{f}_\alpha: U_\alpha \times F \to U_\alpha \times F$ given by

$$(b, x) \mapsto (b, \rho_\alpha(b)(x))$$
is a diffeomorphism. For this, observe that in local coordinates, the Jacobian matrix of this map is of the form

\[ J = \begin{pmatrix} I & 0 \\ C & J(\rho_\alpha(b)) \end{pmatrix}, \]

where \( I \) is the identity matrix and \( J(\rho_\alpha(b)) \) is the Jacobian matrix of \( \rho_\alpha(b) \). Since \( \rho_\alpha(b) \) is a diffeomorphism, \( \det(J) \neq 0 \), and by the Inverse Function Theorem, the map \( \tilde{f}_\alpha \) is a diffeomorphism, as desired.

Remark: If in Proposition 28.2, \( \xi_1 = (E_1, \pi_1, B_1, F, G) \) and \( \xi_2 = (E_2, \pi_2, B_2, F, G) \) are two bundles over different bases and if we have a diffeomorphism \( f_B : B_1 \rightarrow B_2 \), and the conditions

\[ g'_\alpha\beta(f_B(b)) = \rho_\alpha(b)g_\alpha\beta(b)\rho_\beta(b)^{-1}, \quad \text{for all } b \in U_\alpha \cap U_\beta \]

hold, then there is a bundle isomorphism \((f_B, f_E)\) between \( \xi_1 \) and \( \xi_2 \).

It follows from Proposition 28.1 and Proposition 28.2 that two bundles over the same base are equivalent iff they are isomorphic, a very useful fact. Actually, we can use the proof of Proposition 28.2 to show that any bundle morphism \( f : \xi_1 \rightarrow \xi_2 \) between two fibre bundles over the same base \( B \) is a bundle isomorphism. Because a bundle morphism \( f \) as above is fibre preserving, \( f \) is bijective, but it is not obvious that its inverse is smooth.

**Proposition 28.3.** Any bundle morphism \( f : \xi_1 \rightarrow \xi_2 \) between two fibre bundles over the same base \( B \) is an isomorphism.

**Proof.** Since \( f \) is bijective this is a local matter, and it is enough to prove that each \( \tilde{f}_\alpha : U_\alpha \times F \rightarrow U_\alpha \times F \) is a diffeomorphism, since \( f \) can be written as

\[ f = \varphi'_\alpha \circ \tilde{f}_\alpha \circ \varphi_\alpha, \]

with

\[ \tilde{f}_\alpha(b, x) = (b, \rho_\alpha(b)(x)). \]

However, the end of the proof of Proposition 28.2 shows that \( \tilde{f}_\alpha \) is a diffeomorphism.

Given a fibre bundle \( \xi = (E, \pi, B, F, G) \), we observed that the family \( g = (g_{\alpha\beta}) \) of transition maps \( g_{\alpha\beta} : U_\alpha \cap U_\beta \rightarrow G \) induced by a trivializing family \( \varphi = (\varphi_\alpha)_{\alpha \in \mathcal{I}} \) relative to the open cover \( \mathcal{U} = (U_\alpha)_{\alpha \in \mathcal{I}} \) for \( B \) satisfies the cocycle condition

\[ g_{\alpha\beta}(b)g_{\beta\gamma}(b) = g_{\alpha\gamma}(b), \]

for all \( \alpha, \beta, \gamma \) such that \( U_\alpha \cap U_\beta \cap U_\gamma \neq \emptyset \) and all \( b \in U_\alpha \cap U_\beta \cap U_\gamma \).

Without altering anything, we may assume that \( g_{\alpha\beta} \) is the (unique) function from \( \emptyset \) to \( G \), when \( U_\alpha \cap U_\beta = \emptyset \). Then, we call a family \( g = (g_{\alpha\beta})_{(\alpha, \beta) \in \mathcal{I} \times \mathcal{I}} \) as above a \( \mathcal{U} \)-cocycle, or simply a cocycle.
Remarkably, given such a cocycle $g$ relative to $U$, a fibre bundle $\xi_g$ over $B$ with fibre $F$ and structure group $G$ having $g$ as family of transition functions can be constructed.

In view of Proposition 28.1, we say that two cocycles $g = (g_{\alpha\beta})_{(\alpha,\beta)\in I \times I}$ and $g' = (g'_{\alpha\beta})_{(\alpha,\beta)\in I \times I}$ are equivalent if there is a family $(\rho_{\alpha})_{\alpha \in I}$ of smooth maps $\rho_{\alpha}: U_{\alpha} \to G$, such that
\[
g'_{\alpha\beta}(b) = \rho_{\alpha}(b)g_{\alpha\beta}(b)\rho_{\beta}(b)^{-1}, \quad \text{for all } b \in U_{\alpha} \cap U_{\beta}.
\]

**Theorem 28.4.** Given two smooth manifolds $B$ and $F$, a Lie group $G$ acting effectively on $F$, an open cover $U = (U_{\alpha})_{\alpha \in I}$ of $B$, and a cocycle $g = (g_{\alpha\beta})_{(\alpha,\beta)\in I \times I}$, there is a fibre bundle $\xi_g = (E,\pi,B,F,G)$ whose transition maps are the maps in the cocycle $g$. Furthermore, if $g$ and $g'$ are equivalent cocycles, then $\xi_g$ and $\xi_{g'}$ are isomorphic.

**Proof sketch.** First, we define the space $Z$ as the disjoint sum
\[
Z = \coprod_{\alpha \in I} U_{\alpha} \times F.
\]

We define the relation $\simeq$ on $Z \times Z$ as follows: For all $(b,x) \in U_{\beta} \times F$ and $(b,y) \in U_{\alpha} \times F$, if $U_{\alpha} \cap U_{\beta} \neq \emptyset$,
\[
(b,x) \simeq (b,y) \quad \text{iff} \quad y = g_{\alpha\beta}(b)(x).
\]

We let $E = Z/\simeq$, and we give $E$ the largest topology such that the injections $\eta_{\alpha}: U_{\alpha} \times F \to Z$ are smooth. The cocycle condition insures that $\simeq$ is indeed an equivalence relation. We define $\pi: E \to B$ by $\pi([b,x]) = b$. If $p: Z \to E$ is the the quotient map, observe that the maps $p \circ \eta_{\alpha}: U_{\alpha} \times F \to E$ are injective, and that
\[
\pi \circ p \circ \eta_{\alpha}(b,x) = b.
\]

Thus,
\[
p \circ \eta_{\alpha}: U_{\alpha} \times F \to \pi^{-1}(U_{\alpha})
\]
is a bijection, and we define the trivializing maps by setting\[
\varphi_{\alpha} = (p \circ \eta_{\alpha})^{-1}.
\]

It is easily verified that the corresponding transition functions are the original $g_{\alpha\beta}$. There are some details to check. A complete proof (the only one we could find!) is given in Steenrod [164], Part I, Section 3, Theorem 3.2. The fact that $\xi_g$ and $\xi_{g'}$ are equivalent when $g$ and $g'$ are equivalent follows from Proposition 28.2 (see Steenrod [164], Part I, Section 2, Lemma 2.10). \hfill \Box

**Remark:** (The following paragraph is intended for readers familiar with Čech cohomology.) Obviously, if we start with a fibre bundle $\xi = (E,\pi,B,F,G)$ whose transition maps are the cocycle $g = (g_{\alpha\beta})$, and form the fibre bundle $\xi_g$, the bundles $\xi$ and $\xi_g$ are equivalent. This
leads to a characterization of the set of equivalence classes of fibre bundles over a base space $B$ as the cohomology set $\check{H}^1(B, G)$.

In the present case, the sheaf $G$ is defined such that $\Gamma(U, G)$ is the group of smooth maps from the open subset $U$ of $B$ to the Lie group $G$. Since $G$ is not abelian, the coboundary maps have to be interpreted multiplicatively. If we define the sets of cochains $C^k(U, G)$, so that

$$C^0(U, G) = \prod_{\alpha} G(U_{\alpha}), \quad C^1(U, G) = \prod_{\alpha<\beta} G(U_{\alpha} \cap U_{\beta}), \quad C^2(U, G) = \prod_{\alpha<\beta<\gamma} G(U_{\alpha} \cap U_{\beta} \cap U_{\gamma}),$$

etc., then it is natural to define

$$\delta_0: C^0(U, G) \rightarrow C^1(U, G)$$

by

$$(\delta_0 g)_{\alpha\beta} = g_{\alpha}^{-1} g_{\beta},$$

for any $g = (g_{\alpha})$, with $g_{\alpha} \in \Gamma(U_{\alpha}, G)$. As to

$$\delta_1: C^1(U, G) \rightarrow C^2(U, G),$$

since the cocycle condition in the usual case is

$$g_{\alpha\beta} + g_{\beta\gamma} = g_{\alpha\gamma},$$

we set

$$(\delta_1 g)_{\alpha\beta\gamma} = g_{\alpha\beta} g_{\beta\gamma} g_{\alpha\gamma}^{-1},$$

for any $g = (g_{\alpha\beta})$, with $g_{\alpha\beta} \in \Gamma(U_{\alpha} \cap U_{\beta}, G)$. Note that a cocycle $g = (g_{\alpha\beta})$ is indeed an element of $Z^1(U, G)$, and the condition for being in the kernel of

$$\delta_1: C^1(U, G) \rightarrow C^2(U, G)$$

is the cocycle condition

$$g_{\alpha\beta}(b) g_{\beta\gamma}(b) = g_{\alpha\gamma}(b),$$

for all $b \in U_{\alpha} \cap U_{\beta} \cap U_{\gamma}$. In the commutative case, two cocycles $g$ and $g'$ are equivalent if their difference is a boundary, which can be stated as

$$g'_{\alpha\beta} + \rho_{\beta} = g_{\alpha\beta} + \rho_{\alpha} = \rho_{\alpha} + g_{\alpha\beta},$$

where $\rho_{\alpha} \in \Gamma(U_{\alpha}, G)$, for all $\alpha \in I$. In the present case, two cocycles $g$ and $g'$ are equivalent iff there is a family $(\rho_{\alpha})_{\alpha \in I}$, with $\rho_{\alpha} \in \Gamma(U_{\alpha}, G)$, such that

$$g'_{\alpha\beta}(b) = \rho_{\alpha}(b) g_{\alpha\beta}(b) \rho_{\beta}(b)^{-1},$$
for all \( b \in U_\alpha \cap U_\beta \). This is the same condition of equivalence defined earlier. Thus, it is easily seen that if \( g, h \in Z^1(\mathcal{U}, G) \), then \( \xi_g \) and \( \xi_h \) are equivalent iff \( g \) and \( h \) correspond to the same element of the cohomology set \( \check{H}^1(\mathcal{U}, G) \).

As usual, \( \check{H}^1(B, G) \) is defined as the direct limit of the directed system of sets \( \check{H}^1(\mathcal{U}, G) \) over the preordered directed family of open covers. For details, see Hirzebruch [92], Section 3.1. In summary, there is a bijection between the equivalence classes of fibre bundles over \( B \) (with fibre \( F \) and structure group \( G \)) and the cohomology set \( \check{H}^1(B, G) \). In the case of line bundles, it turns out that \( \check{H}^1(B, G) \) is in fact a group.

As an application of Theorem 28.4, we define the notion of pullback (or induced) bundle. Say \( \xi = (E, \pi, B, F, G) \) is a fibre bundle and assume we have a smooth map \( f : N \to B \). We seek a bundle \( f^*\xi \) over \( N \), together with a bundle map \((f^*, f) : f^*\xi \to \xi\),

\[
\begin{array}{ccc}
  f^*E & \xrightarrow{f^*} & E \\
  \pi^* & \downarrow & \pi \\
  N & \xrightarrow{f} & B,
\end{array}
\]

where \( f^*E \) is a pullback in the categorical sense. This means that for any other bundle \( \xi' \) over \( N \) and any bundle map

\[
\begin{array}{ccc}
  E' & \xrightarrow{f'} & E \\
  \pi' & \downarrow & \pi \\
  N & \xrightarrow{f} & B,
\end{array}
\]

there is a unique bundle map \((\tilde{f}', \text{id}) : \xi' \to f^*\xi\), so that \((f', f) = (f^*, f) \circ (\tilde{f}', \text{id})\). Thus, there is an isomorphism (natural)

\[
\text{Hom}(\xi', \xi) \cong \text{Hom}(\xi', f^*\xi).
\]

As a consequence, by Proposition 28.3, for any bundle map between \( \xi' \) and \( \xi \),

\[
\begin{array}{ccc}
  E' & \xrightarrow{f'} & E \\
  \pi' & \downarrow & \pi \\
  N & \xrightarrow{f} & B,
\end{array}
\]

there is an isomorphism, \( \xi' \cong f^*\xi \).

The bundle \( f^*\xi \) can be constructed as follows: Pick any open cover \((U_\alpha)\) of \( B \), then \((f^{-1}(U_\alpha))\) is an open cover of \( N \), and check that if \((g_{\alpha\beta})\) is a cocycle for \( \xi \), then the maps \( g_{\alpha\beta} \circ f : f^{-1}(U_\alpha) \cap f^{-1}(U_\beta) \to G \) satisfy the cocycle conditions. Then, \( f^*\xi \) is the bundle
defined by the cocycle \((g_{\alpha \beta} \circ f)\). We leave as an exercise to show that the pullback bundle \(f^*\xi\) can be defined explicitly if we set
\[
f^*E = \{(n, e) \in N \times E \mid f(n) = \pi(e)\},
\]
\(\pi^* = \text{pr}_1\) and \(f^* = \text{pr}_2\). For any trivialization \(\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times F\) of \(\xi\), we have
\[
(\pi^*)^{-1}(f^{-1}(U_\alpha)) = \{(n, e) \in N \times E \mid n \in f^{-1}(U_\alpha), e \in \pi^{-1}(f(n))\},
\]
and so we have a bijection \(\tilde{\varphi}_\alpha: (\pi^*)^{-1}(f^{-1}(U_\alpha)) \to f^{-1}(U_\alpha) \times F\), given by
\[
\tilde{\varphi}_\alpha(n, e) = (n, \text{pr}_2(\varphi_\alpha(e))).
\]
By giving \(f^*E\) the smallest topology that makes each \(\tilde{\varphi}_\alpha\) a diffeomorphism, \(\tilde{\varphi}_\alpha\) is a trivialization of \(f^*\xi\) over \(f^{-1}(U_\alpha)\), and \(f^*\xi\) is a smooth bundle. Note that the fibre of \(f^*\xi\) over a point \(n \in N\) is isomorphic to the fibre \(\pi^{-1}(f(n))\) of \(\xi\) over \(f(n)\). If \(g: M \to N\) is another smooth map of manifolds, it is easy to check that
\[
(f \circ g)^*\xi = g^*(f^*\xi).
\]

Given a bundle \(\xi = (E, \pi, B, F, G)\) and a submanifold \(N\) of \(B\), we define the restriction of \(\xi\) to \(N\) as the bundle \(\xi \upharpoonright N = (\pi^{-1}(N), \pi \upharpoonright \pi^{-1}(N), B, F, G)\).

Experience shows that most objects of interest in geometry (vector fields, differential forms, etc.) arise as sections of certain bundles. Furthermore, deciding whether or not a bundle is trivial often reduces to the existence of a (global) section. Thus, we define the important concept of a section right away.

**Definition 28.5.** Given a fibre bundle \(\xi = (E, \pi, B, F, G)\), a *smooth section* of \(\xi\) is a smooth map \(s: B \to E\), so that \(\pi \circ s = \text{id}_B\). Given an open subset \(U\) of \(B\), a (smooth) section of \(\xi\) over \(U\) is a smooth map \(s: U \to E\), so that \(\pi \circ s(b) = b\), for all \(b \in U\); we say that \(s\) is a *local section* of \(\xi\). The set of all sections over \(U\) is denoted \(\Gamma(U, \xi)\), and \(\Gamma(B, \xi)\) (for short, \(\Gamma(\xi)\)) is the set of global sections of \(\xi\).

Here is an observation that proves useful for constructing global sections. Let \(s: B \to E\) be a global section of a bundle \(\xi\). For every trivialization \(\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times F\), let \(s_\alpha: U_\alpha \to E\) and \(\sigma_\alpha: U_\alpha \to F\) be given by
\[
s_\alpha = s \upharpoonright U_\alpha \quad \text{and} \quad \sigma_\alpha = \text{pr}_2 \circ \varphi_\alpha \circ s_\alpha,
\]
so that
\[
s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)).
\]
Obviously, \(\pi \circ s_\alpha = \text{id}\), so \(s_\alpha\) is a local section of \(\xi\), and \(\sigma_\alpha\) is a function \(\sigma_\alpha: U_\alpha \to F\). We claim that on overlaps, we have
\[
\sigma_\alpha(b) = g_{\alpha \beta}(b)\sigma_\beta(b).
\]
Proof. Indeed, recall that
\[ \varphi_\alpha \circ \varphi_\beta^{-1}(b, x) = (b, g_{\alpha \beta}(b)x), \]
for all \( b \in U_\alpha \cap U_\beta \) and all \( x \in F \), and as \( s_\alpha = s \upharpoonright U_\alpha \) and \( s_\beta = s \upharpoonright U_\beta \), \( s_\alpha \) and \( s_\beta \) agree on \( U_\alpha \cap U_\beta \). Consequently, from
\[ s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)) \quad \text{and} \quad s_\beta(b) = \varphi_\beta^{-1}(b, \sigma_\beta(b)), \]
we get
\[ \varphi_\alpha^{-1}(b, \sigma_\alpha(b)) = s_\alpha(b) = s_\beta(b) = \varphi_\beta^{-1}(b, \sigma_\beta(b)) = \varphi_\alpha^{-1}(b, g_{\alpha \beta}(b)\sigma_\beta(b)), \]
which implies \( \sigma_\alpha(b) = g_{\alpha \beta}(b)\sigma_\beta(b) \), as claimed.

Conversely, assume that we have a collection of functions \( \sigma_\alpha : U_\alpha \to F \), satisfying
\[ \sigma_\alpha(b) = g_{\alpha \beta}(b)\sigma_\beta(b) \]
on overlaps. Let \( s_\alpha : U_\alpha \to E \) be given by
\[ s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)). \]
Each \( s_\alpha \) is a local section and we claim that these sections agree on overlaps, so they patch and define a global section \( s \). We need to show that
\[ s_\alpha(b) = \varphi_\alpha^{-1}(b, \sigma_\alpha(b)) = \varphi_\beta^{-1}(b, \sigma_\beta(b)) = s_\beta(b), \]
for \( b \in U_\alpha \cap U_\beta \); that is,
\[ (b, \sigma_\alpha(b)) = \varphi_\alpha \circ \varphi_\beta^{-1}(b, \sigma_\beta(b)), \]
and since \( \varphi_\alpha \circ \varphi_\beta^{-1}(b, \sigma_\beta(b)) = (b, g_{\alpha \beta}(b)\sigma_\beta(b)) \), and by hypothesis, \( \sigma_\alpha(b) = g_{\alpha \beta}(b)\sigma_\beta(b) \), our equation \( s_\alpha(b) = s_\beta(b) \) is verified. \( \square \)

There are two particularly interesting special cases of fibre bundles:

1. **Vector bundles**, which are fibre bundles for which the typical fibre is a finite-dimensional vector space \( V \), and the structure group is a subgroup of the group of linear isomorphisms \( (\text{GL}(n, \mathbb{R}) \text{ or } \text{GL}(n, \mathbb{C})) \), where \( n = \dim V \).

2. **Principal fibre bundles**, which are fibre bundles for which the fibre \( F \) is equal to the structure group \( G \), with \( G \) acting on itself by left translation.

First, we discuss vector bundles.
892 CHAPTER 28. BUNDLES, METRICS ON BUNDLES, HOMOGENEOUS SPACES

28.2 Vector Bundles

Given a real vector space \( V \), we denote by \( \text{GL}(V) \) (or \( \text{Aut}(V) \)) the vector space of linear invertible maps from \( V \) to \( V \). If \( V \) has dimension \( n \), then \( \text{GL}(V) \) has dimension \( n^2 \). Obviously, \( \text{GL}(V) \) is isomorphic to \( \text{GL}(n, \mathbb{R}) \), so we often write \( \text{GL}(n, \mathbb{R}) \) instead of \( \text{GL}(V) \), but this may be slightly confusing if \( V \) is the dual space \( W^* \) of some other space \( W \). If \( V \) is a complex vector space, we also denote by \( \text{GL}(V) \) (or \( \text{Aut}(V) \)) the vector space of linear invertible maps from \( V \) to \( V \), but this time \( \text{GL}(V) \) is isomorphic to \( \text{GL}(n, \mathbb{C}) \), so we often write \( \text{GL}(n, \mathbb{C}) \) instead of \( \text{GL}(V) \).

**Definition 28.6.** A rank \( n \) real smooth vector bundle with fibre \( V \) is a tuple \( \xi = (E, \pi, B, V) \), such that \( (E, \pi, B, V, \text{GL}(V)) \) is a smooth fibre bundle, the fibre \( V \) is a real vector space of dimension \( n \), and the following conditions hold:

(a) For every \( b \in B \), the fibre \( \pi^{-1}(b) \) is an \( n \)-dimensional (real) vector space.

(b) For every trivialization \( \varphi_\alpha : \pi^{-1}(U_\alpha) \rightarrow U_\alpha \times V \), for every \( b \in U_\alpha \), the restriction of \( \varphi_\alpha \) to the fibre \( \pi^{-1}(b) \) is a linear isomorphism \( \pi^{-1}(b) \rightarrow V \).

A rank \( n \) complex smooth vector bundle with fibre \( V \) is a tuple \( \xi = (E, \pi, B, V) \), where \( (E, \pi, B, V, \text{GL}(V)) \) is a smooth fibre bundle such that the fibre \( V \) is an \( n \)-dimensional complex vector space (viewed as a real smooth manifold), and conditions (a) and (b) above hold (for complex vector spaces). When \( n = 1 \), a vector bundle is called a line bundle.

The trivial vector bundle \( E = B \times V \) is often denoted \( \mathcal{E}^V \). When \( V = \mathbb{R}^k \), we also use the notation \( \mathcal{E}^k \). Given a (smooth) manifold \( M \) of dimension \( n \), the tangent bundle \( TM \) and the cotangent bundle \( T^*M \) are rank \( n \) vector bundles. Indeed, in Section 8.1, we defined trivialization maps (denoted \( \tau_U \)) for \( TM \). Let us compute the transition functions for the tangent bundle \( TM \), where \( M \) is a smooth manifold of dimension \( n \).

Recall from Definition 7.8 that for every \( p \in M \), the tangent space \( T_pM \) consists of all equivalence classes of triples \((U, \varphi, x)\), where \((U, \varphi)\) is a chart with \( p \in U \), \( x \in \mathbb{R}^n \), and the equivalence relation on triples is given by

\[
(U, \varphi, x) \equiv (V, \psi, y) \quad \text{iff} \quad (\psi \circ \varphi^{-1})'(\varphi(p))(x) = y.
\]

We have a natural isomorphism \( \theta_{U,\varphi,p} : \mathbb{R}^n \rightarrow T_pM \) between \( \mathbb{R}^n \) and \( T_pM \) given by

\[
\theta_{U,\varphi,p} : x \mapsto [(U, \varphi, x)], \quad x \in \mathbb{R}^n.
\]

Observe that for any two overlapping charts \((U, \varphi)\) and \((V, \psi)\),

\[
\theta_{V,\psi,p}^{-1} \circ \theta_{U,\varphi,p} = (\psi \circ \varphi^{-1})'_z
\]

for all \( p \in U \cap V \), with \( z = \varphi(p) = \psi(p) \). We let \( TM \) be the disjoint union

\[
TM = \bigcup_{p \in M} T_pM,
\]
28.2. VECTOR BUNDLES

define the projection $\pi : TM \to M$ so that $\pi(v) = p$ if $v \in T_p M$, and we give $TM$ the weakest topology that makes the functions $\tilde{\varphi} : \pi^{-1}(U) \to \mathbb{R}^{2n}$ given by

$$\tilde{\varphi}(v) = (\varphi \circ \pi(v), \theta_{U,\varphi,\pi(v)}^{-1}(v))$$

continuous, where $(U, \varphi)$ is any chart of $M$. Each function $\tilde{\varphi} : \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n$ is a homeomorphism, and given any two overlapping charts $(U, \varphi)$ and $(V, \psi)$, since $\theta_{V,\psi,p}^{-1} \circ \theta_{U,\varphi,p} = (\psi \circ \varphi^{-1})_\varphi'(z)$, with $z = \varphi(p) = \psi(p)$, the transition map

$$\tilde{\psi} \circ \tilde{\varphi}^{-1} : \varphi(U \cap V) \times \mathbb{R}^n \to \psi(U \cap V) \times \mathbb{R}^n$$

is given by

$$\tilde{\psi} \circ \tilde{\varphi}^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), (\psi \circ \varphi^{-1})_\varphi'(x)), \quad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.$$

It is clear that $\tilde{\psi} \circ \tilde{\varphi}^{-1}$ is smooth. Moreover, the bijection

$$\tau_U : \pi^{-1}(U) \to U \times \mathbb{R}^n$$

given by

$$\tau_U(v) = (\pi(v), \theta_{U,\varphi,\pi(v)}^{-1}(v))$$

satisfies $pr_1 \circ \tau_U = \pi$ on $\pi^{-1}(U)$ and is a linear isomorphism restricted to fibres, so it is a trivialization for $TM$. For any two overlapping charts $(U_\alpha, \varphi_\alpha)$ and $(U_\beta, \varphi_\beta)$, the transition function, $g_{\alpha\beta} : U_\alpha \cap U_\beta \to GL(n, \mathbb{R})$ is given by

$$g_{\alpha\beta}(p) = (\varphi_\alpha \circ \varphi_\beta^{-1})_\varphi'(p).$$

We can also compute trivialization maps for $T^*M$. This time, $T^*M$ is the disjoint union

$$T^*M = \bigcup_{p \in M} T^*_p M,$$

and $\pi : T^*M \to M$ is given by $\pi(\omega) = p$ if $\omega \in T^*_p M$, where $T^*_p M$ is the dual of the tangent space $T_p M$. For each chart $(U, \varphi)$, by dualizing the map $\theta_{U,\varphi,p} : \mathbb{R}^n \to T_p M$, we obtain an isomorphism $\theta_{U,\varphi,p}^* : T^*_p M \to (\mathbb{R}^n)^*$. Composing $\theta_{U,\varphi,p}^*$ with the isomorphism $\iota : (\mathbb{R}^n)^* \to \mathbb{R}^n$ (induced by the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$ and its dual basis), we get an isomorphism $\theta_{U,\varphi,p}^* = \iota \circ \theta_{U,\varphi,p}^T : T^*_p M \to \mathbb{R}^n$. Then, define the bijection

$$\tilde{\varphi}^* : \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n \subseteq \mathbb{R}^{2n}$$

by

$$\tilde{\varphi}^*(\omega) = (\varphi \circ \pi(\omega), \theta_{U,\varphi,\pi(\omega)}^*(\omega)),$$

with $\omega \in \pi^{-1}(U)$. We give $T^*M$ the weakest topology that makes the functions $\tilde{\varphi}^*$ continuous, and then each function $\tilde{\varphi}^*$ is a homeomorphism. Given any two overlapping charts $(U, \varphi)$ and $(V, \psi)$, as

$$\theta_{V,\psi,p}^{-1} \circ \theta_{U,\varphi,p} = (\psi \circ \varphi^{-1})_\varphi'(p),$$
by dualization we get
\[
\theta_{U,\varphi,\psi}^\top \circ (\theta_{U,\psi,\varphi}^\top)^{-1} = \theta_{U,\varphi,\psi}^\top \circ (\theta_{V,\psi,\varphi}^\top)^{-1} = (((\psi \circ \varphi^{-1})_{\varphi(p)})^\top, \!
\]
then
\[
\theta_{V,\psi,\varphi}^\top \circ (\theta_{U,\varphi,\psi}^\top)^{-1} = (((\psi \circ \varphi^{-1})_{\varphi(p)})^\top)^{-1}, \!
\]
and so
\[
\iota \circ \theta_{V,\psi,\varphi}^\top \circ (\theta_{U,\varphi,\psi}^\top)^{-1} = \iota \circ (((\psi \circ \varphi^{-1})_{\varphi(p)})^\top)^{-1} \circ \iota^{-1}; \!
\]
that is,
\[
\theta_{V,\psi,\varphi}^* \circ (\theta_{U,\varphi,\psi}^*)^{-1} = \iota \circ (((\psi \circ \varphi^{-1})_{\varphi(p)})^\top)^{-1} \circ \iota^{-1}. \!
\]
Consequently, the transition map
\[
\tilde{\psi}^* \circ (\tilde{\varphi}^*)^{-1} : \varphi(U \cap V) \times \mathbb{R}^n \longrightarrow \psi(U \cap V) \times \mathbb{R}^n
\]
is given by
\[
\tilde{\psi}^* \circ (\tilde{\varphi}^*)^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), \iota \circ (((\psi \circ \varphi^{-1})_{\varphi(p)})^\top)^{-1} \circ \iota^{-1}(x)), \quad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n. \!
\]
If we view \((\psi \circ \varphi^{-1})_{\varphi(p)}') as a matrix, then we can forget \(\iota\) and the second component of \(\tilde{\psi}^* \circ (\tilde{\varphi}^*)^{-1}(z, x)\) is \(((\psi \circ \varphi^{-1})_{\varphi(p)})^\top)^{-1} x.\!

We also have trivialization maps \(\tau_U^* : \pi^{-1}(U) \rightarrow U \times (\mathbb{R}^n)^*\) for \(T^* M\), given by
\[
\tau_U^*(\omega) = (\pi(\omega), \theta_{U,\varphi,\psi,\pi(\omega)}^\top(\omega)), \!
\]
for all \(\omega \in \pi^{-1}(U)\). The transition function \(g_{\alpha\beta}^* : U_\alpha \cap U_\beta \rightarrow \text{GL}(n, \mathbb{R})\) is given by
\[
g_{\alpha\beta}^*(p)(\eta) = \tau_{U,\varphi,\psi,\pi(\eta)}^* \circ (\tau_{U,\psi,\varphi,\pi(\eta)}^*)^{-1}(\eta)
\]
\[
= \theta_{U,\varphi,\pi(\eta)}^\top \circ (\theta_{U,\varphi,\psi,\pi(\eta)}^\top)^{-1}(\eta)
\]
\[
= (((\theta_{U,\varphi,\pi(\eta)}\circ \theta_{U,\varphi,\psi,\pi(\eta)})^\top)^{-1}(\eta)
\]
\[
= (((\varphi^{-1}_{\psi(p)}\circ \varphi^{-1}_{\varphi(p)})^\top)^{-1}(\eta), \!
\]
with \(\eta \in (\mathbb{R}^n)^*\). Also note that \(\text{GL}(n, \mathbb{R})\) should really be \(\text{GL}((\mathbb{R}^n)^*)\), but \(\text{GL}((\mathbb{R}^n)^*)\) is isomorphic to \(\text{GL}(n, \mathbb{R})\). We conclude that
\[
g_{\alpha\beta}^*(p) = (g_{\alpha\beta}(p)^\top)^{-1}, \quad \text{for every } p \in M. \!
\]
This is a general property of dual bundles; see Property (f) in Section 28.3.

Maps of vector bundles are maps of fibre bundles such that the isomorphisms between fibres are linear.

**Definition 28.7.** Given two vector bundles \(\xi_1 = (E_1, \pi_1, B_1, V)\) and \(\xi_2 = (E_2, \pi_2, B_2, V)\) with the same typical fibre \(V\), a **bundle map** (or bundle morphism) \(f : \xi_1 \rightarrow \xi_2\) is a pair \(f = (f_E, f_B)\) of smooth maps \(f_E : E_1 \rightarrow E_2\) and \(f_B : B_1 \rightarrow B_2\), such that:
(a) The following diagram commutes:

\[
\begin{array}{c}
E_1 \xrightarrow{f_E} E_2 \\
\downarrow \pi_1 \quad \downarrow \pi_2 \\
B_1 \xrightarrow{f_B} B_2
\end{array}
\]

(b) For every \( b \in B_1 \), the map of fibres

\[ f_E | \pi_1^{-1}(b): \pi_1^{-1}(b) \rightarrow \pi_2^{-1}(f_B(b)) \]

is a bijective linear map.

A bundle map \textit{isomorphism} \( f: \xi_1 \rightarrow \xi_2 \) is defined as in Definition 28.2. Given two vector bundles \( \xi_1 = (E_1, \pi_1, B, V) \) and \( \xi_2 = (E_2, \pi_2, B, V) \) over the same base space \( B \), we require \( f_B = \text{id} \).

\textbf{Remark:} Some authors do not require the preservation of fibres; that is, the map

\[ f_E | \pi_1^{-1}(b): \pi_1^{-1}(b) \rightarrow \pi_2^{-1}(f_B(b)) \]

is simply a linear map. It is automatically bijective for bundle isomorphisms.

Note that Definition 28.7 does not include condition (b) of Definition 28.3. However, because the restrictions of the maps \( \varphi_\alpha, \varphi'_\beta \), and \( f \) to the fibres are linear isomorphisms, it turns out that condition (b) (of Definition 28.3) does hold.

Indeed, if \( f_B(U_\alpha) \subseteq V_\beta \), then

\[ \varphi'_\beta \circ f \circ \varphi_\alpha^{-1}: U_\alpha \times V \longrightarrow V_\beta \times V \]

is a smooth map and, for every \( b \in B \), its restriction to \( \{b\} \times V \) is a linear isomorphism between \( \{b\} \times V \) and \( \{f_B(b)\} \times V \). Therefore, there is a smooth map \( \rho_\alpha: U_\alpha \rightarrow \text{GL}(n, \mathbb{R}) \) so that

\[ \varphi'_\beta \circ f \circ \varphi_\alpha^{-1}(b, x) = (f_B(b), \rho_\alpha(b)(x)), \]

and a vector bundle map is a fibre bundle map.

A \textit{holomorphic vector bundle} is a fibre bundle where \( E, B \) are complex manifolds, \( V \) is a complex vector space of dimension \( n \), the map \( \pi \) is holomorphic, the \( \varphi_\alpha \) are biholomorphic, and the transition functions \( g_{\alpha\beta} \) are holomorphic. When \( n = 1 \), a holomorphic vector bundle is called a \textit{holomorphic line bundle}.

Definition 28.4 (equivalence of bundles) also applies to vector bundles (just replace \( G \) by \( \text{GL}(n, \mathbb{R}) \) or \( \text{GL}(n, \mathbb{C}) \)) and defines the notion of equivalence of vector bundles over \( B \). Since vector bundle maps are fibre bundle maps, Propositions 28.1 and 28.2 immediately yield
Proposition 28.5. Two vector bundles $\xi_1 = (E_1, \pi_1, B, V)$ and $\xi_2 = (E_2, \pi_2, B, V)$ over the same base space $B$ are equivalent iff they are isomorphic.

Since a vector bundle map is a fibre bundle map, Proposition 28.3 also yields the useful fact:

Proposition 28.6. Any vector bundle map $f : \xi_1 \to \xi_2$ between two vector bundles over the same base $B$ is an isomorphism.

Theorem 28.4 also holds for vector bundles and yields a technique for constructing new vector bundles over some base, $B$.

Theorem 28.7. Given a smooth manifold $B$, an $n$-dimensional (real, resp. complex) vector space $V$, an open cover $U = (U_\alpha)_{\alpha \in I}$ of $B$, and a cocycle $g = (g_{\alpha \beta})_{(\alpha, \beta) \in I \times I}$ (with $g_{\alpha \beta} : U_\alpha \cap U_\beta \to \text{GL}(n, \mathbb{R})$, resp. $g_{\alpha \beta} : U_\alpha \cap U_\beta \to \text{GL}(n, \mathbb{C})$), there is a vector bundle $\xi_g = (E, \pi, B, V)$ whose transition maps are the maps in the cocycle $g$. Furthermore, if $g$ and $g'$ are equivalent cocycles, then $\xi_g$ and $\xi_{g'}$ are equivalent.

Observe that a cocycle $g = (g_{\alpha \beta})_{(\alpha, \beta) \in I \times I}$ is given by a family of matrices in $\text{GL}(n, \mathbb{R})$ (resp. $\text{GL}(n, \mathbb{C})$).

A vector bundle $\xi$ always has a global section, namely the zero section, which assigns the element $0 \in \pi^{-1}(b)$ to every $b \in B$. A global section $s$ is a non-zero section iff $s(b) \neq 0$ for all $b \in B$.

It is usually difficult to decide whether a bundle has a nonzero section. This question is related to the nontriviality of the bundle, and there is a useful test for triviality.

Assume $\xi$ is a trivial rank $n$ vector bundle. Then, there is a bundle isomorphism $f : B \times V \to \xi$. For every $b \in B$, we know that $f(b, -)$ is a linear isomorphism, so for any choice of a basis $(e_1, \ldots, e_n)$ of $V$, we get a basis $(f(b, e_1), \ldots, f(b, e_n))$ of the fibre $\pi^{-1}(b)$. Thus, we have $n$ global sections $s_1 = f(-, e_1), \ldots, s_n = f(-, e_n)$ such that $(s_1(b), \ldots, s_n(b))$ forms a basis of the fibre $\pi^{-1}(b)$, for every $b \in B$.

Definition 28.8. Let $\xi = (E, \pi, B, V)$ be a rank $n$ vector bundle. For any open subset $U \subseteq B$, an $n$-tuple of local sections $(s_1, \ldots, s_n)$ over $U$ is called a frame over $U$ iff $(s_1(b), \ldots, s_n(b))$ is a basis of the fibre $\pi^{-1}(b)$, for every $b \in U$. If $U = B$, then the $s_i$ are global sections and $(s_1, \ldots, s_n)$ is called a frame (of $\xi$).

The notion of a frame is due to Élie Cartan who (after Darboux) made extensive use of them under the name of moving frame (and the moving frame method). Cartan’s terminology is intuitively clear: As a point $b$ moves in $U$, the frame $(s_1(b), \ldots, s_n(b))$ moves from fibre to fibre. Physicists refer to a frame as a choice of local gauge.

The converse of the property established just before Definition 28.8 is also true.
Proposition 28.8. A rank $n$ vector bundle $\xi$ is trivial iff it possesses a frame of global sections.

Proof. We only need to prove that if $\xi$ has a frame $(s_1,\ldots,s_n)$, then it is trivial. Pick a basis $(e_1,\ldots,e_n)$ of $V$, and define the map $f: B \times V \to \xi$ as follows:

$$f(b,v) = \sum_{i=1}^{n} v_is_i(b),$$

where $v = \sum_{i=1}^{n} v_ie_i$. Clearly, $f$ is bijective on fibres, smooth, and a map of vector bundles. By Proposition 28.6, the bundle map, $f$, is an isomorphism. 

The above considerations show that if $\xi$ is any rank $n$ vector bundle, not necessarily trivial, then for any local trivialization $\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times V$, there are always frames over $U_\alpha$. Indeed, for every choice of a basis $(e_1,\ldots,e_n)$ of the typical fibre $V$, if we set $s_\alpha^i(b) = \varphi_\alpha^{-1}(b,e_i)$, $b \in U_\alpha$, $1 \leq i \leq n$, then $(s_\alpha^1,\ldots,s_\alpha^n)$ is a frame over $U_\alpha$.

Given any two vector spaces $V$ and $W$, both of dimension $n$, we denote by $\text{Iso}(V,W)$ the space of all linear isomorphisms between $V$ and $W$. The space of $n$-frames $F(V)$ is the set of bases of $V$. Since every basis $(v_1,\ldots,v_n)$ of $V$ is in one-to-one correspondence with the map from $\mathbb{R}^n$ to $V$ given by $e_i \mapsto v_i$, where $(e_1,\ldots,e_n)$ is the canonical basis of $\mathbb{R}^n$ (so, $e_i = (0,\ldots,1,\ldots,0)$ with the 1 in the $i$th slot), we have an isomorphism,

$$F(V) \cong \text{Iso}(\mathbb{R}^n,V).$$

(The choice of a basis in $V$ also yields an isomorphism $\text{Iso}(\mathbb{R}^n,V) \cong \text{GL}(n,\mathbb{R})$, so $F(V) \cong \text{GL}(n,\mathbb{R})$.)

For any rank $n$ vector bundle $\xi$, we can form the frame bundle $F(\xi)$, by replacing the fibre $\pi^{-1}(b)$ over any $b \in B$ by $F(\pi^{-1}(b))$. In fact, $F(\xi)$ can be constructed using Theorem 28.4. Indeed, identifying $F(V)$ with $\text{Iso}(\mathbb{R}^n,V)$, the group $\text{GL}(n,\mathbb{R})$ acts on $F(V)$ effectively on the left via

$$A \cdot v = v \circ A^{-1}.$$

(The only reason for using $A^{-1}$ instead of $A$ is that we want a left action.) The resulting bundle has typical fibre $F(V) \cong \text{GL}(n,\mathbb{R})$, and turns out to be a principal bundle. We will take a closer look at principal bundles in Section 28.5.

We conclude this section with an example of a bundle that plays an important role in algebraic geometry, the canonical line bundle on $\mathbb{P}^n$. Let $H^\mathbb{R}_n \subseteq \mathbb{R}^n \times \mathbb{R}^{n+1}$ be the subset

$$H^\mathbb{R}_n = \{(L,v) \in \mathbb{R}^n \times \mathbb{R}^{n+1} \mid v \in L\},$$
where \( \mathbb{R}P^n \) is viewed as the set of lines \( L \) in \( \mathbb{R}^{n+1} \) through 0, or more explicitly,

\[
H_n^\mathbb{R} = \{((x_0: \cdots: x_n), \lambda(x_0, \ldots, x_n)) \mid (x_0: \cdots: x_n) \in \mathbb{R}P^n, \lambda \in \mathbb{R}\}.
\]

Geometrically, \( H_n^\mathbb{R} \) consists of the set of lines \([(x_0, \ldots, x_n)]\) associated with points \((x_0: \cdots: x_n)\) of \( \mathbb{R}P^n \). If we consider the projection \( \pi: H_n^\mathbb{R} \to \mathbb{R}P^n \) of \( H_n^\mathbb{R} \) onto \( \mathbb{R}P^n \), we see that each fibre is isomorphic to \( \mathbb{R} \). We claim that \( H_n^\mathbb{R} \) is a line bundle. For this, we exhibit trivializations, leaving as an exercise the fact that \( H_n^\mathbb{R} \) is a manifold.

Recall the open cover \( U_0, \ldots, U_n \) of \( \mathbb{R}P^n \), where

\[
U_i = \{(x_0: \cdots: x_n) \in \mathbb{R}P^n \mid x_i \neq 0\}.
\]

Then, the maps \( \varphi_i: \pi^{-1}(U_i) \to U_i \times \mathbb{R} \) given by

\[
\varphi_i((x_0: \cdots: x_n), \lambda(x_0, \ldots, x_n)) = ((x_0: \cdots: x_n), \lambda x_i)
\]

are trivializations. The transition function \( g_{ij}: U_i \cap U_j \to \text{GL}(1, \mathbb{R}) \) is given by

\[
g_{ij}(x_0: \cdots: x_n)(u) = \frac{x_i}{x_j} u,
\]

where we identify \( \text{GL}(1, \mathbb{R}) \) and \( \mathbb{R}^* = \mathbb{R} - \{0\} \).

Interestingly, the bundle \( H_n^\mathbb{R} \) is nontrivial for all \( n \geq 1 \). For this, by Proposition 28.8 and since \( H_n^\mathbb{R} \) is a line bundle, it suffices to prove that every global section vanishes at some point. So, let \( \sigma \) be any section of \( H_n^\mathbb{R} \). Composing the projection, \( p: S^n \to \mathbb{R}P^n \), with \( \sigma \), we get a smooth function, \( s = \sigma \circ p: S^n \to H_n^\mathbb{R} \), and we have

\[
s(x) = (p(x), f(x)x),
\]

for every \( x \in S^n \), where \( f: S^n \to \mathbb{R} \) is a smooth function. Moreover, \( f \) satisfies

\[
f(-x) = -f(x),
\]

since \( s(-x) = s(x) \). As \( S^n \) is connected and \( f \) is continuous, by the intermediate value theorem, there is some \( x \) such that \( f(x) = 0 \), and thus, \( \sigma \) vanishes, as desired.

The reader should look for a geometric representation of \( H_1^\mathbb{R} \). It turns out that \( H_1^\mathbb{R} \) is an open Möbius strip; that is, a Möbius strip with its boundary deleted (see Milnor and Stasheff [129], Chapter 2). There is also a complex version of the canonical line bundle on \( \mathbb{C}P^n \), with

\[
H_n = \{(L, v) \in \mathbb{C}P^n \times \mathbb{C}^{n+1} \mid v \in L\},
\]

where \( \mathbb{C}P^n \) is viewed as the set of lines \( L \) in \( \mathbb{C}^{n+1} \) through 0. These bundles are also nontrivial. Furthermore, unlike the real case, the dual bundle \( H_n^* \) is not isomorphic to \( H_n \). Indeed, \( H_n^* \) turns out to have nonzero global holomorphic sections!
28.3 Operations on Vector Bundles

Because the fibres of a vector bundle are vector spaces all isomorphic to some given space $V$, we can perform operations on vector bundles that extend familiar operations on vector spaces, such as: direct sum, tensor product, (linear) function space, and dual space. Basically, the same operation is applied on fibres. It is usually more convenient to define operations on vector bundles in terms of operations on cocycles, using Theorem 28.7.

(a) (Whitney Sum or Direct Sum)

If $\xi = (E, \pi, B, V)$ is a rank $m$ vector bundle and $\xi' = (E', \pi', B, W)$ is a rank $n$ vector bundle, both over the same base $B$, then their Whitney sum $\xi \oplus \xi'$ is the rank $(m + n)$ vector bundle whose fibre over any $b \in B$ is the direct sum $E_b \oplus E'_b$; that is, the vector bundle with typical fibre $V \oplus W$ (given by Theorem 28.7) specified by the cocycle whose matrices are

$$
\begin{pmatrix}
g_{\alpha\beta}(b) & 0 \\
0 & g'_{\alpha\beta}(b)
\end{pmatrix}, \quad b \in U_\alpha \cap U_\beta.
$$

(b) (Tensor Product)

If $\xi = (E, \pi, B, V)$ is a rank $m$ vector bundle and $\xi' = (E', \pi', B, W)$ is a rank $n$ vector bundle, both over the same base $B$, then their tensor product $\xi \otimes \xi'$ is the rank $mn$ vector bundle whose fibre over any $b \in B$ is the tensor product $E_b \otimes E'_b$; that is, the vector bundle with typical fibre $V \otimes W$ (given by Theorem 28.7) specified by the cocycle whose matrices are

$$
g_{\alpha\beta}(b) \otimes g'_{\alpha\beta}(b), \quad b \in U_\alpha \cap U_\beta.
$$

(Here, we identify a matrix with the corresponding linear map.)

(c) (Tensor Power)

If $\xi = (E, \pi, B, V)$ is a rank $m$ vector bundle, then for any $k \geq 0$, we can define the tensor power bundle $\xi \otimes^k$, whose fibre over any $b \in \xi$ is the tensor power $E_b \otimes^k$, and with typical fibre $V \otimes^k$. (When $k = 0$, the fibre is $\mathbb{R}$ or $\mathbb{C}$). The bundle $\xi \otimes^k$ is determined by the cocycle

$$
g_{\alpha\beta}^k(b), \quad b \in U_\alpha \cap U_\beta.
$$

(d) (Exterior Power)

If $\xi = (E, \pi, B, V)$ is a rank $m$ vector bundle, then for any $k \geq 0$, we can define the exterior power bundle $\bigwedge^k \xi$, whose fibre over any $b \in \xi$ is the exterior power $\bigwedge^k E_b$, and with typical fibre $\bigwedge^k V$. The bundle $\bigwedge^k \xi$ is determined by the cocycle

$$
\bigwedge^k g_{\alpha\beta}(b), \quad b \in U_\alpha \cap U_\beta.
$$

Using (a), we also have the exterior algebra bundle $\bigwedge \xi = \bigoplus_{k=0}^m \bigwedge^k \xi$. (When $k = 0$, the fibre is $\mathbb{R}$ or $\mathbb{C}$).
(e) \((\text{Symmetric Power})\) If \(\xi = (E, \pi, B, V)\) is a rank \(m\) vector bundle, then for any \(k \geq 0\), we can define the symmetric power bundle \(S^k \xi\), whose fibre over any \(b \in \xi\) is the exterior power \(S^k E_b\), and with typical fibre \(S^k V\). (When \(k = 0\), the fibre is \(\mathbb{R}\) or \(\mathbb{C}\).) The bundle \(S^k \xi\) is determined by the cocycle
\[
S^k g_{\alpha \beta}(b), \quad b \in U_{\alpha} \cap U_{\beta}.
\]

(f) \((\text{Dual Bundle})\) If \(\xi = (E, \pi, B, V)\) is a rank \(m\) vector bundle, then its dual bundle \(\xi^*\) is the rank \(m\) vector bundle whose fibre over any \(b \in B\) is the dual space \(E_b^*\); that is, the vector bundle with typical fibre \(V^*\) (given by Theorem 28.7) specified by the cocycle whose matrices are
\[
(g_{\alpha \beta}(b)^\top)^{-1}, \quad b \in U_{\alpha} \cap U_{\beta}.
\]
The reason for this seemingly complicated formula is this: For any trivialization \(\varphi_{\alpha}: \pi^{-1}(U_{\alpha}) \to U_{\alpha} \times V\), for any \(b \in B\), recall that the restriction \(\varphi_{\alpha, b}: \pi^{-1}(b) \to V\) of \(\varphi_{\alpha}\) to \(\pi^{-1}(b)\) is a linear isomorphism. By dualization we get a map \(\varphi_{\alpha, b}^*: V^* \to (\pi^{-1}(b))^*\), and thus \(\varphi_{\alpha, b}^*\) for \(\xi^*\) is given by
\[
\varphi_{\alpha, b}^* = (\varphi_{\alpha, b}^\top)^{-1}: (\pi^{-1}(b))^* \to V^*.
\]
As \(g_{\alpha \beta}^*(b) = \varphi_{\alpha, b}^* \circ (\varphi_{\beta, b}^*)^{-1}\), we get
\[
g_{\alpha \beta}^*(b) = (\varphi_{\alpha, b}^\top)^{-1} \circ \varphi_{\beta, b}^\top \\
= ((\varphi_{\beta, b}^\top)^{-1} \circ \varphi_{\alpha, b}^\top)^{-1} \\
= (\varphi_{\beta, b}^\top)^{-1} \circ \varphi_{\alpha, b}^\top \\
= ((\varphi_{\alpha, b} \circ \varphi_{\beta, b}^\top)^{-1})^{-1} \\
= (g_{\alpha \beta}(b)^\top)^{-1},
\]
as claimed.

(g) \((\text{Hom Bundle})\)

If \(\xi = (E, \pi, B, V)\) is a rank \(m\) vector bundle and \(\xi' = (E', \pi', B, W)\) is a rank \(n\) vector bundle, both over the same base \(B\), then their \(\text{Hom}\) bundle \(\text{Hom}(\xi, \xi')\) is the rank \(mn\) vector bundle whose fibre over any \(b \in B\) is \(\text{Hom}(E_b, E_b')\); that is, the vector bundle with typical fibre \(\text{Hom}(V, W)\). The transition functions of this bundle are obtained as follows: For any trivializations \(\varphi_{\alpha}: \pi^{-1}(U_{\alpha}) \to U_{\alpha} \times V\) and \(\varphi_{\alpha}': (\pi')^{-1}(U_{\alpha}) \to U_{\alpha} \times W\), for any \(b \in B\), recall that the restrictions \(\varphi_{\alpha, b}: \pi^{-1}(b) \to V\) and \(\varphi_{\alpha, b}': (\pi')^{-1}(b) \to W\) are linear isomorphisms. We have a linear isomorphism \(\varphi_{\alpha, b}^\text{Hom}: \text{Hom}(\pi^{-1}(b), (\pi')^{-1}(b)) \to \text{Hom}(V, W)\) given by
\[
\varphi_{\alpha, b}^\text{Hom}(f) = \varphi_{\alpha, b}' \circ f \circ \varphi_{\alpha, b}^{-1}, \quad f \in \text{Hom}(\pi^{-1}(b), (\pi')^{-1}(b)).
\]
Then, \(g_{\alpha \beta}^\text{Hom}(b) = \varphi_{\alpha, b}^\text{Hom} \circ (\varphi_{\beta, b}^\text{Hom})^{-1}\).
28.3. OPERATIONS ON VECTOR BUNDLES

(h) (Tensor Bundle of type \((r, s)\))

If \(\xi = (E, \pi, B, V)\) is a rank \(m\) vector bundle, then for any \(r, s \geq 0\), we can define the bundle \(T^{r,s}\xi\) whose fibre over any \(b \in \xi\) is the tensor space \(T^{r,s}E_b\), and with typical fibre \(T^{r,s}V\). The bundle \(T^{r,s}\xi\) is determined by the cocycle

\[ g_{\alpha\beta}^{r,s}(b) \otimes ((g_{\alpha\beta}(b)^T)^{-1})^{\otimes s}(b), \quad b \in U_{\alpha} \cap U_{\beta}. \]

In view of the canonical isomorphism \(\text{Hom}(V, W) \cong V^* \otimes W\), it is easy to show that \(\text{Hom}(\xi, \xi')\), is isomorphic to \(\xi^* \otimes \xi'\). Similarly, \(\xi^{**}\) is isomorphic to \(\xi\). We also have the isomorphism

\[ T^{r,s}\xi \cong \xi^{\otimes r} \otimes (\xi^*)^{\otimes s}. \]

Do not confuse the space of bundle morphisms \(\text{Hom}(\xi, \xi')\) with the bundle \(\mathcal{H}om(\xi, \xi')\). However, observe that \(\text{Hom}(\xi, \xi')\) is the set of global sections of \(\mathcal{H}om(\xi, \xi')\).

As an illustration of (d), consider the exterior power \(\bigwedge^r T^*M\), where \(M\) is a manifold of dimension \(n\). We have trivialization maps \(\tau_U^r: \pi^{-1}(U) \to U \times \bigwedge^r (\mathbb{R}^n)\) for \(\bigwedge^r T^*M\), given by

\[ \tau_U^r(\omega) = (\pi(\omega), \bigwedge^r \theta_{U,\varphi_\pi(\omega)}(\omega)), \]

for all \(\omega \in \pi^{-1}(U)\). The transition function \(g_{\alpha\beta}^r: U_{\alpha} \cap U_{\beta} \to \text{GL}(n, \mathbb{R})\) is given by

\[ g_{\alpha\beta}^r(p)(\omega) = \bigwedge^r (((\varphi_\alpha \circ \varphi_\beta^{-1})'_{\varphi(p)})^T)^{-1}(\omega), \]

for all \(\omega \in \pi^{-1}(U)\). Consequently,

\[ g_{\alpha\beta}^r(p) = \bigwedge^r (g_{\alpha\beta}(p)^T)^{-1}, \]

for every \(p \in M\), a special case of (h).

For rank 1 vector bundles, namely line bundles, it is easy to show that the set of equivalence classes of line bundles over a base \(B\) forms a group, where the group operation is \(\otimes\), the inverse is \(*\) (dual), and the identity element is the trivial bundle. This is the Picard group of \(B\).

In general, the dual \(E^*\) of a bundle is not isomorphic to the original bundle \(E\). This is because \(V^*\) is not canonically isomorphic to \(V\), and to get a bundle isomorphism between \(\xi\) and \(\xi^*\), we need canonical isomorphisms between the fibres. However, if \(\xi\) is real, then (using a partition of unity) \(\xi\) can be given a Euclidean metric and so, \(\xi\) and \(\xi^*\) are isomorphic. It is not true in general that a complex vector bundle is isomorphic to its dual because a Hermitian metric only induces a canonical isomorphism between \(E^*\) and \(\overline{E}\), where \(\overline{E}\) is the conjugate of \(E\), with scalar multiplication in \(\overline{E}\) given by \((z, w) \mapsto \overline{w}z\).
Remark: Given a real vector bundle, \( \xi \), the *complexification* \( \xi_C \) of \( \xi \) is the complex vector bundle defined by

\[
\xi_C = \xi \otimes_{\mathbb{R}} \mathbb{C},
\]

where \( \epsilon_C = B \times \mathbb{C} \) is the trivial complex line bundle. Given a complex vector bundle \( \xi \), by viewing its fibre as a real vector space we obtain the real vector bundle \( \xi_{\mathbb{R}} \). The following facts can be shown:

1. For every real vector bundle \( \xi \),
\[
(\xi_C)_{\mathbb{R}} \cong \xi \oplus \xi.
\]
2. For every complex vector bundle \( \xi \),
\[
(\xi_{\mathbb{R}})_C \cong \xi \oplus \xi^*.
\]

The notion of subbundle is defined as follows:

**Definition 28.9.** Given two vector bundles \( \xi = (E, \pi, B, V) \) and \( \xi' = (E', \pi', B, V') \) over the same base \( B \), we say that \( \xi \) is a *subbundle* of \( \xi' \) iff \( E \) is a submanifold of \( E' \), \( V \) is a subspace of \( V' \), and for every \( b \in B \), the fibre \( \pi^{-1}(b) \) is a subspace of the fibre \( (\pi')^{-1}(b) \).

If \( \xi \) is a subbundle of \( \xi' \), we can form the *quotient bundle* \( \xi'/\xi \) as the bundle over \( B \) whose fibre at \( b \in B \) is the quotient space \( (\pi')^{-1}(b)/\pi^{-1}(b) \). We leave it as an exercise to define trivializations for \( \xi'/\xi \). In particular, if \( N \) is a submanifold of \( M \), then \( TN \) is a subbundle of \( TM \upharpoonright N \) and the quotient bundle \( (TM \upharpoonright N)/TN \) is called the *normal bundle* of \( N \) in \( M \).

### 28.4 Metrics on Vector Bundles, Reduction of Structure Groups, Orientation

Because the fibres of a vector bundle are vector spaces, the definition of a Riemannian metric on a manifold can be lifted to vector bundles.

**Definition 28.10.** Given a (real) rank \( n \) vector bundle \( \xi = (E, \pi, B, V) \), we say that \( \xi \) is *Euclidean* iff there is a family \( \langle -, - \rangle_b \) \( b \in B \) of inner products on each fibre \( \pi^{-1}(b) \), such that \( \langle -, - \rangle_b \) depends smoothly on \( b \), which means that for every trivializing map \( \varphi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times V \), for every frame, \( (s_1, \ldots, s_n) \), on \( U_\alpha \), the maps

\[
b \mapsto \langle s_i(b), s_j(b) \rangle_b, \quad b \in U_\alpha, \quad 1 \leq i, j \leq n
\]

are smooth. We say that \( \langle -, - \rangle \) is a *Euclidean metric* (or *Riemannian metric*) on \( \xi \). If \( \xi \) is a complex rank \( n \) vector bundle \( \xi = (E, \pi, B, V) \), we say that \( \xi \) is *Hermitian* iff there is a family \( \langle -, - \rangle_b \) \( b \in B \) of Hermitian inner products on each fibre \( \pi^{-1}(b) \), such that \( \langle -, - \rangle_b \) depends smoothly on \( b \). We say that \( \langle -, - \rangle \) is a *Hermitian metric* on \( \xi \). For any smooth manifold \( M \), if \( TM \) is a Euclidean vector bundle, then we say that \( M \) is a *Riemannian manifold*. 
28.4. METRICS ON BUNDLES, REDUCTION, ORIENTATION

Now, given a real (resp. complex) vector bundle ξ, provided that B is a sufficiently nice topological space, namely that B is paracompact (see Section 9.1), a Euclidean metric (resp. Hermitian metric) exists on ξ. This is a consequence of the existence of partitions of unity (see Theorem 9.4).

**Theorem 28.9.** Every real (resp. complex) vector bundle admits a Euclidean (resp. Hermitian) metric. In particular, every smooth manifold admits a Riemannian metric.

**Proof.** Let \((U_α)\) be a trivializing open cover for ξ and pick any frame \((s_1^α, \ldots, s_n^α)\) over \(U_α\). For every \(b \in U_α\), the basis \((s_1^α(b), \ldots, s_n^α(b))\) defines a Euclidean (resp. Hermitian) inner product \(\langle -, - \rangle_b\) on the fibre \(π^{-1}(b)\), by declaring \((s_1^α(b), \ldots, s_n^α(b))\) orthonormal w.r.t. this inner product. (For \(x = \sum_{i=1}^n x_i s_i^α(b)\) and \(y = \sum_{i=1}^n y_i s_i^α(b)\), let \(\langle x, y \rangle_b = \sum_{i=1}^n x_i y_i\), resp. \(\langle x, y \rangle_b = \sum_{i=1}^n x_i \bar{y}_i\), in the complex case.) The \(\langle -, - \rangle_b\) (with \(b \in U_α\)) define a metric on \(π^{-1}(U_α)\), denote it \(\langle -, - \rangle_α\). Now, using Theorem 9.4, glue these inner products using a partition of unity \((f_α)\) subordinate to \((U_α)\), by setting

\[
\langle x, y \rangle = \sum_α f_α \langle x, y \rangle_α.
\]

We verify immediately that \(\langle -, - \rangle\) is a Euclidean (resp. Hermitian) metric on ξ. □

The existence of metrics on vector bundles allows the so-called reduction of structure group. Recall that the transition maps of a real (resp. complex) vector bundle ξ are functions \(g_{αβ}: U_α \cap U_β \to GL(n, \mathbb{R})\) (resp. \(GL(n, \mathbb{C})\)). Let \(GL^+(n, \mathbb{R})\) be the subgroup of \(GL(n, \mathbb{R})\) consisting of those matrices of positive determinant (resp. \(GL^+(n, \mathbb{C})\) be the subgroup of \(GL(n, \mathbb{C})\) consisting of those matrices of positive determinant).

**Definition 28.11.** For every real (resp. complex) vector bundle ξ, if it is possible to find a cocycle \(g = (g_{αβ})\) for ξ with values in a subgroup \(H\) of \(GL(n, \mathbb{R})\) (resp. of \(GL(n, \mathbb{C})\)), then we say that the structure group of ξ can be reduced to \(H\). We say that ξ is orientable if its structure group can be reduced to \(GL^+(n, \mathbb{R})\) (resp. \(GL^+(n, \mathbb{C})\)).

**Proposition 28.10.** (a) The structure group of a rank \(n\) real vector bundle ξ can be reduced to \(O(n)\); it can be reduced to \(SO(n)\) iff ξ is orientable.

(b) The structure group of a rank \(n\) complex vector bundle ξ can be reduced to \(U(n)\); it can be reduced to \(SU(n)\) iff ξ is orientable.

**Proof.** We prove (a), the proof of (b) being similar. Using Theorem 28.9, put a metric on ξ. For every \(U_α\) in a trivializing cover for ξ and every \(b \in B\), by Gram-Schmidt, orthonormal bases for \(π^{-1}(b)\) exit. Consider the family of trivializing maps \(\tilde{φ}_α: π^{-1}(U_α) \to U_α \times V\) such that \(\tilde{φ}_{α,b}: π^{-1}(b) \to V\) maps orthonormal bases of the fibre to orthonormal bases of \(V\). Then, it is easy to check that the corresponding cocycle takes values in \(O(n)\) and if ξ is orientable, the determinants being positive, these values are actually in \(SO(n)\). □
Remark: If $\xi$ is a Euclidean rank $n$ vector bundle, then by Proposition 28.10, we may assume that $\xi$ is given by some cocycle $(g_{\alpha\beta})$, where $g_{\alpha\beta}(b) \in O(n)$, for all $b \in U_\alpha \cap U_\beta$. We saw in Section 28.3 (f) that the dual bundle $\xi^*$ is given by the cocycle

$$(g_{\alpha\beta}(b))^{-1}, \quad b \in U_\alpha \cap U_\beta.$$  

As $g_{\alpha\beta}(b)$ is an orthogonal matrix, $(g_{\alpha\beta}(b)^{\top})^{-1} = g_{\alpha\beta}(b)$, and thus, any Euclidean bundle is isomorphic to its dual. As we noted earlier, this is false for Hermitian bundles.

Let $\xi = (E, \pi, B, V)$ be a rank $n$ vector bundle and assume $\xi$ is orientable. A family of trivializing maps $\varphi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times V$ is oriented iff for all $\alpha, \beta$, the transition function $g_{\alpha\beta}(b)$ has positive determinant for all $b \in U_\alpha \cap U_\beta$. Two oriented families of trivializing maps $\varphi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\psi_\beta : \pi^{-1}(W_\beta) \to W_\alpha \times V$ are equivalent iff for every $b \in U_\alpha \cap W_\beta$, the map $pr_2 \circ \varphi_\alpha \circ \psi_\beta^{-1} \mid \{b\} \times V : V \to V$ has positive determinant.

It is easily checked that this is an equivalence relation and that it partitions all the oriented families of trivializations of $\xi$ into two equivalence classes. Either equivalence class is called an orientation of $\xi$.

If $M$ is a manifold and $\xi = TM$, the tangent bundle of $\xi$, we know from Section 28.2 that the transition functions of $TM$ are of the form

$$g_{\alpha\beta}(p)(u) = (\varphi_\alpha \circ \varphi_\beta^{-1})_{\varphi(p)}(u),$$

where each $\varphi_\alpha : U_\alpha \to \mathbb{R}^n$ is a chart of $M$. Consequently, $TM$ is orientable iff the Jacobian of $(\varphi_\alpha \circ \varphi_\beta^{-1})_{\varphi(p)}$ is positive, for every $p \in M$. This is equivalent to the condition of Definition 24.1 for $M$ to be orientable. Therefore, the tangent bundle $TM$ of a manifold $M$ is orientable iff $M$ is orientable.

The notion of orientability of a vector bundle $\xi = (E, \pi, B, V)$ is not equivalent to the orientability of its total space $E$. Indeed, if we look at the transition functions of the total space of $TM$ given in Section 28.2, we see that $TM$, as a manifold, is always orientable, even if $M$ is not orientable. Indeed, the transition functions of the tangent bundle $TM$ are of the form

$$\tilde{\psi} \circ \tilde{\varphi}^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), (\psi \circ \varphi^{-1})'_{z}(x)), \quad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.$$

Since $(\psi \circ \varphi^{-1})'_{z}$ is a linear map, its derivative at any point is equal to itself, and it follows that the derivative of $\tilde{\psi} \circ \tilde{\varphi}^{-1}$ at $(z, x)$ is given by

$$(\tilde{\psi} \circ \tilde{\varphi}^{-1})'_{(z,x)}(u, v) = ((\psi \circ \varphi^{-1})'_{z}(u), (\psi \circ \varphi^{-1})'_{z}(v)), \quad (u, v) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Then, the Jacobian matrix of this map is of the form

$$J = \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix}$$

where $A$ is an $n \times n$ matrix, since $(\psi \circ \varphi^{-1})'_{z}(u)$ does not involve the variables in $v$ and $(\psi \circ \varphi^{-1})'_{z}(v)$ does not involve the variables in $u$. Therefore $\det(J) = \det(A)^2$, which shows that the transition functions have positive Jacobian determinant, and thus that $TM$ is orientable.
Yet, as a bundle, $TM$ is orientable iff $M$.

On the positive side, if $\xi = (E, \pi, B, V)$ is an orientable vector bundle and its base $B$ is an orientable manifold, then $E$ is orientable too.

To see this, assume that $B$ is a manifold of dimension $m$, $\xi$ is a rank $n$ vector bundle with fibre $V$, let $((U_\alpha, \psi_\alpha))_\alpha$ be an atlas for $B$, let $\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times V$ be a collection of trivializing maps for $\xi$, and pick any isomorphism, $\iota: V \to \mathbb{R}^n$. Then, we get maps

$$(\psi_\alpha \times \iota) \circ \varphi_\alpha: \pi^{-1}(U_\alpha) \to \mathbb{R}^m \times \mathbb{R}^n.$$ 

It is clear that these maps form an atlas for $E$. Check that the corresponding transition maps for $E$ are of the form

$$(x, y) \mapsto (\psi_\beta \circ \psi^{-1}_\alpha(x), g_{\alpha\beta}(\psi^{-1}_\alpha(x))y).$$

Moreover, if $B$ and $\xi$ are orientable, check that these transition maps have positive Jacobian.

The fact that every bundle admits a metric allows us to define the notion of orthogonal complement of a subbundle. We state the following theorem without proof. The reader is invited to consult Milnor and Stasheff [129] for a proof (Chapter 3).

**Proposition 28.11.** Let $\xi$ and $\eta$ be two vector bundles with $\xi$ a subbundle of $\eta$. Then, there exists a subbundle $\xi^\perp$ of $\eta$, such that every fibre of $\xi^\perp$ is the orthogonal complement of the fibre of $\xi$ in the fibre of $\eta$ over every $b \in B$, and

$$\eta \approx \xi \oplus \xi^\perp.$$ 

In particular, if $N$ is a submanifold of a Riemannian manifold $M$, then the orthogonal complement of $TN$ in $TM \upharpoonright N$ is isomorphic to the normal bundle $(TM \upharpoonright N)/TN$.

**Remark:** It can be shown (see Madsen and Tornehave [119], Chapter 15) that for every real smooth vector bundle $\xi$, there is some integer $k$ such that $\xi$ has a complement $\eta$ in $\mathbb{E}^k$, where $\mathbb{E}^k = B \times \mathbb{R}^k$ is the trivial rank $k$ vector bundle, so that

$$\xi \oplus \eta = \mathbb{E}^k.$$ 

This fact can be used to prove an interesting property of the space of global sections $\Gamma(\xi)$.

First, observe that $\Gamma(\xi)$ is not just a real vector space, but also a $C^\infty(B)$-module (see Section 21.12). Indeed, for every smooth function $f: B \to \mathbb{R}$ and every smooth section $s: B \to E$, the map $fs: B \to E$ given by

$$(fs)(b) = f(b)s(b), \quad b \in B,$$

is a smooth section of $\xi$. 
In general, $\Gamma(\xi)$ is not a free $C^\infty(B)$-module unless $\xi$ is trivial. However, the above remark implies that
\[\Gamma(\xi) \oplus \Gamma(\eta) = \Gamma(\epsilon^k),\]
where $\Gamma(\epsilon^k)$ is a free $C^\infty(B)$-module of dimension $\dim(\xi) + \dim(\eta)$.

This proves that $\Gamma(\xi)$ is a finitely generated $C^\infty(B)$-module which is a summand of a free $C^\infty(B)$-module. Such modules are projective modules; see Definition 21.11 in Section 21.12. Therefore, $\Gamma(\xi)$ is a finitely generated projective $C^\infty(B)$-module. The following isomorphisms can be shown (see Madsen and Tornehave [119], Chapter 16):

**Proposition 28.12.** The following isomorphisms hold for vector bundles:
\[
\Gamma(\mathcal{H}om(\xi, \eta)) \cong \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\eta)) \\
\Gamma(\xi \otimes \eta) \cong \Gamma(\xi) \otimes_{C^\infty(B)} \Gamma(\eta) \\
\Gamma(\xi^*) \cong \text{Hom}_{C^\infty(B)}(\Gamma(\xi), C^\infty(B)) = (\Gamma(\xi))^* \\
\Gamma(\bigwedge^k \xi) \cong \bigwedge^k_{C^\infty(B)} (\Gamma(\xi)).
\]

Using the operations on vector bundles described in Section 28.3, we can define the set of vector valued differential forms $A^k(M; F)$ defined in Section 23.4 as the set of smooth sections of the vector bundle $\bigwedge^k T^*M \otimes \epsilon_F$; that is, as
\[A^k(M; F) = \Gamma\left(\bigwedge^k T^*M \otimes \epsilon_F\right),\]
where $\epsilon_F$ is the trivial vector bundle $\epsilon_F = M \times F$. In view of Proposition 28.12 and since $\Gamma(\epsilon_F) \cong C^\infty(M; F)$ and $A^k(M) = \Gamma\left(\bigwedge^k T^*M\right)$, we have
\[A^k(M; F) = \Gamma\left(\bigwedge^k T^*M \otimes \epsilon_F\right) \cong \Gamma\left(\bigwedge^k T^*M \otimes_{C^\infty(M)} \Gamma(\epsilon_F)\right) = A^k(M) \otimes_{C^\infty(M)} C^\infty(M; F) \cong \bigwedge^k_{C^\infty(M)} (\mathcal{X}(M); C^\infty(M; F)).\]

with all of the spaces viewed as $C^\infty(M)$-modules. Therefore,
\[A^k(M; F) \cong A^k(M) \otimes_{C^\infty(M)} C^\infty(M; F) \cong \text{Alt}^k_{C^\infty(M)}(\mathcal{X}(M); C^\infty(M; F)),\]
which reduces to Proposition 23.12 when $F = \mathbb{R}$.

In Section 29.1, we will consider a generalization of the above situation where the trivial vector bundle $\epsilon_F$ is replaced by any vector bundle $\xi = (E, \pi, B, V)$, and where $M = B$. 
28.5 Principal Fibre Bundles

We now consider principal bundles. Such bundles arise in terms of Lie groups acting on manifolds.

Definition 28.12. Let \( G \) be a Lie group. A principal fibre bundle, for short a principal bundle, is a fibre bundle \( \xi = (E, \pi, B, G, G) \) in which the fibre is \( G \) and the structure group is also \( G \), viewed as its group of left translations (ie., \( G \) acts on itself by multiplication on the left). This means that every transition function, \( g_{\alpha\beta}: U_\alpha \cap U_\beta \to G \), satisfies

\[
g_{\alpha\beta}(b)(h) = g(b)h, \quad \text{for some } g(b) \in G,
\]

for all \( b \in U_\alpha \cap U_\beta \) and all \( h \in G \). A principal \( G \)-bundle is denoted \( \xi = (E, \pi, B, G) \).

Note that \( G \) in \( g_{\alpha\beta}: U_\alpha \cap U_\beta \to G \) is viewed as its group of left translations under the isomorphism \( g \mapsto L_g \), and so \( g_{\alpha\beta}(b) \) is some left translation \( L_g(b) \). The inverse of the above isomorphism is given by \( L \mapsto L(1) \), so \( g(b) = g_{\alpha\beta}(b)(1) \). In view of these isomorphisms, we allow ourself the (convenient) abuse of notation

\[
g_{\alpha\beta}(b)(h) = g_{\alpha\beta}(b)h,
\]

where on the left, \( g_{\alpha\beta}(b) \) is viewed as a left translation of \( G \), and on the right as an element of \( G \).

When we want to emphasize that a principal bundle has structure group \( G \), we use the locution principal \( G \)-bundle.

It turns out that if \( \xi = (E, \pi, B, G) \) is a principal bundle, then \( G \) acts on the total space \( E \), on the right. For the next proposition, recall that a right action \( \cdot: X \times G \to X \) is free iff for every \( g \in G \), if \( g \neq 1 \), then \( x \cdot g \neq x \) for all \( x \in X \).

Proposition 28.13. If \( \xi = (E, \pi, B, G) \) is a principal bundle, then there is a right action of \( G \) on \( E \). This action takes each fibre to itself and is free. Moreover, \( E/G \) is diffeomorphic to \( B \).

Proof. We show how to define the right action and leave the rest as an exercise. Let \( \{(U_\alpha, \varphi_\alpha)\} \) be some trivializing cover defining \( \xi \). For every \( z \in E \), pick some \( U_\alpha \) so that \( \pi(z) \in U_\alpha \), and let \( \varphi_\alpha(z) = (b, h) \), where \( b = \pi(z) \) and \( h \in G \). For any \( g \in G \), we set

\[
z \cdot g = \varphi_\alpha^{-1}(b, hg).
\]

If we can show that this action does not depend on the choice of \( U_\alpha \), then it is clear that it is a free action. Suppose that we also have \( b = \pi(z) \in U_\beta \) and that \( \varphi_\beta(z) = (b, h') \). By definition of the transition functions, we have

\[
h' = g_{\beta\alpha}(b)h \quad \text{and} \quad \varphi_\beta(z \cdot g) = (b, g_{\beta\alpha}(b)(hg)).
\]


However,

\[ g_{\beta\alpha}(b)(hg) = (g_{\beta\alpha}(b)h)g = h'g, \]

hence

\[ z \cdot g = \varphi^{-1}_\beta(b,h'g), \]

which proves that our action does not depend on the choice of \( U_\alpha \).

Observe that the action of Proposition 28.13 is defined by

\[ z \cdot g = \varphi^{-1}_\alpha(b,\varphi_{\alpha,b}(z)g), \quad \text{with} \quad b = \pi(z), \]

for all \( z \in E \) and all \( g \in G \). It is clear that this action satisfies the following two properties:

For every \((U_\alpha, \varphi_\alpha)\),

(1) \( \pi(z \cdot g) = \pi(z) \), and

(2) \( \varphi_\alpha(z \cdot g) = \varphi_\alpha(z) \cdot g \), for all \( z \in E \) and all \( g \in G \),

where we define the right action of \( G \) on \( U_\alpha \times G \) so that \((b, h) \cdot g = (b, hg)\). We say that \( \varphi_\alpha \) is \( G \)-equivariant (or equivariant).

The following proposition shows that it is possible to define a principal \( G \)-bundle using a suitable right action and equivariant trivializations:

**Proposition 28.14.** Let \( E \) be a smooth manifold, \( G \) be a Lie group, and let \( \cdot : E \times G \to E \) be a smooth right action of \( G \) on \( E \) satisfying the following properties:

(a) The right action of \( G \) on \( E \) is free;

(b) The orbit space \( B = E/G \) is a smooth manifold under the quotient topology, and the projection \( \pi : E \to E/G \) is smooth;

(c) There is a family of local trivializations \( \{(U_\alpha, \varphi_\alpha)\} \), where \( \{U_\alpha\} \) is an open cover for \( B = E/G \), and each

\[ \varphi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times G \]

is an equivariant diffeomorphism, which means that

\[ \varphi_\alpha(z \cdot g) = \varphi_\alpha(z) \cdot g, \]

for all \( z \in \pi^{-1}(U_\alpha) \) and all \( g \in G \), where the right action of \( G \) on \( U_\alpha \times G \) is \((b, h) \cdot g = (b, hg)\).

Then, \( \xi = (E, \pi, E/G, G) \) is a principal \( G \)-bundle.
Proof. Since the action of $G$ on $E$ is free, every orbit $b = z \cdot G$ is isomorphic to $G$, and so every fibre $\pi^{-1}(b)$ is isomorphic to $G$. Thus, given that we have trivializing maps, we just have to prove that $G$ acts by left translation on itself. Pick any $(b, h)$ in $U_\beta \times G$ and let $z \in \pi^{-1}(U_\beta)$ be the unique element such that $\varphi_\beta(z) = (b, h)$. Then, as

$$\varphi_\beta(z \cdot g) = \varphi_\beta(z) \cdot g, \quad \text{for all } g \in G,$$

we have

$$\varphi_\beta(\varphi_\beta^{-1}(b, h) \cdot g) = \varphi_\beta(z \cdot g) = \varphi_\beta(z) \cdot g = (b, h) \cdot g,$$

which implies that

$$\varphi_\beta^{-1}(b, h) \cdot g = \varphi_\beta^{-1}((b, h) \cdot g).$$

Consequently,

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, h) = \varphi_\alpha \circ \varphi_\beta^{-1}((b, 1) \cdot h) = \varphi_\alpha(\varphi_\beta^{-1}(b, 1) \cdot h) = \varphi_\alpha \circ \varphi_\beta^{-1}(b, 1) \cdot h,$$

and since

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, h) = (b, g_{\alpha \beta}(b)(h)) \quad \text{and} \quad \varphi_\alpha \circ \varphi_\beta^{-1}(b, 1) = (b, g_{\alpha \beta}(b)(1))$$

we get

$$g_{\alpha \beta}(b)(h) = g_{\alpha \beta}(b)(1)h.$$

The above shows that $g_{\alpha \beta}(b): G \to G$ is the left translation by $g_{\alpha \beta}(b)(1)$, and thus the transition functions $g_{\alpha \beta}(b)$ constitute the group of left translations of $G$, and $\xi$ is indeed a principal $G$-bundle.

Bröcker and tom Dieck [31] (Chapter I, Section 4) and Duistermaat and Kolk [64] (Appendix A) define principal bundles using the conditions of Proposition 28.14. Propositions 28.13 and 28.14 show that this alternate definition is equivalent to ours (Definition 28.12).

It turns out that when we use the definition of a principal bundle in terms of the conditions of Proposition 28.14, it is convenient to define bundle maps in terms of equivariant maps. As we will see shortly, a map of principal bundles is a fibre bundle map.

**Definition 28.13.** If $\xi_1 = (E_1, \pi_1, B_1, G)$ and $\xi_2 = (E_2, \pi_2, B_1, G)$ are two principal bundles, a bundle map (or bundle morphism) $f: \xi_1 \to \xi_2$ is a pair, $f = (f_E, f_B)$ of smooth maps $f_E: E_1 \to E_2$ and $f_B: B_1 \to B_2$, such that:

(a) The following diagram commutes:

\[
\begin{array}{ccc}
E_1 & \xrightarrow{f_E} & E_2 \\
\pi_1 \downarrow & & \downarrow \pi_2 \\
B_1 & \xrightarrow{f_B} & B_2
\end{array}
\]
(b) The map \( f_E \) is \( G \)-equivariant; that is,
\[
f_E(a \cdot g) = f_E(a) \cdot g,
\]
for all \( a \in E_1 \) and all \( g \in G \).

A bundle map is an isomorphism if it has an inverse as in Definition 28.2. If the bundles \( \xi_1 \) and \( \xi_2 \) are over the same base \( B \), then we also require \( f_B = \text{id} \).

At first glance, it is not obvious that a map of principal bundles satisfies condition (b) of Definition 28.3. If we define \( \tilde{f}_\alpha : U_\alpha \times G \to V_\beta \times G \) by
\[
\tilde{f}_\alpha = \varphi_\beta' \circ f_E \circ \varphi_\alpha^{-1},
\]
then locally \( f_E \) is expressed as
\[
f_E = \varphi_\beta' \circ \tilde{f}_\alpha \circ \varphi_\alpha.
\]
Furthermore, it is trivial that if a map is equivariant and invertible, then its inverse is equivariant. Consequently, since
\[
\tilde{f}_\alpha = \varphi_\beta' \circ f_E \circ \varphi_\alpha^{-1},
\]
as \( \varphi_\alpha^{-1} \), \( \varphi_\beta' \) and \( f_E \) are equivariant, \( \tilde{f}_\alpha \) is also equivariant, and so \( \tilde{f}_\alpha \) is a map of (trivial) principal bundles. Thus, it is enough to prove that for every map of principal bundles \( \varphi : U_\alpha \times G \to V_\beta \times G \),
there is some smooth map \( \rho_\alpha : U_\alpha \to G \), so that
\[
\varphi(b, g) = (f_B(b), \rho_\alpha(b)(g)), \quad \text{for all } b \in U_\alpha \text{ and all } g \in G.
\]
Indeed, we have the following

**Proposition 28.15.** For every map of trivial principal bundles
\[
\varphi : U_\alpha \times G \to V_\beta \times G,
\]
there are smooth maps \( f_B : U_\alpha \to V_\beta \) and \( r_\alpha : U_\alpha \to G \), so that
\[
\varphi(b, g) = (f_B(b), r_\alpha(b)(g)), \quad \text{for all } b \in U_\alpha \text{ and all } g \in G.
\]
In particular, \( \varphi \) is a diffeomorphism on fibres.

**Proof.** As \( \varphi \) is a map of principal bundles
\[
\varphi(b, 1) = (f_B(b), r_\alpha(b)), \quad \text{for all } b \in U_\alpha,
\]
for some smooth maps \( f_B : U_\alpha \to V_\beta \) and \( r_\alpha : U_\alpha \to G \). Now, using equivariance, we get
\[
\varphi(b, g) = \varphi((b, 1)g) = \varphi(b, 1) \cdot g = (f_B(b), r_\alpha(b)) \cdot g = (f_B(b), r_\alpha(b)g),
\]
as claimed. \( \Box \)
Consequently, the map \( \rho_\alpha : U_\alpha \to G \) given by
\[
\rho_\alpha(b)(g) = r_\alpha(b)g \quad \text{for all } b \in U_\alpha \text{ and all } g \in G
\]
satisfies
\[
\varphi(b,g) = (f_B(b),\rho_\alpha(b)(g)), \quad \text{for all } b \in U_\alpha \text{ and all } g \in G,
\]
and a map of principal bundles is indeed a fibre bundle map (as in Definition 28.3). Since a principal bundle map is a fibre bundle map, Proposition 28.3 also yields the useful fact:

**Proposition 28.16.** Any map \( f : \xi_1 \to \xi_2 \) between two principal bundles over the same base \( B \) is an isomorphism.

Even though we are not aware of any practical applications in computer vision, robotics, or medical imaging, we wish to digress briefly on the issue of the triviality of bundles and the existence of sections.

A natural question is to ask whether a fibre bundle \( \xi \) is isomorphic to a trivial bundle. If so, we say that \( \xi \) is trivial. (By the way, the triviality of bundles comes up in physics, in particular, field theory.) Generally, this is a very difficult question, but a first step can be made by showing that it reduces to the question of triviality for principal bundles.

Indeed, if \( \xi = (E, \pi, B, F, G) \) is a fibre bundle with fibre \( F \), using Theorem 28.4, we can construct a principal fibre bundle \( P(\xi) \) using the transition functions \( \{g_{\alpha\beta}\} \) of \( \xi \), but using \( G \) itself as the fibre (acting on itself by left translation) instead of \( F \). We obtain the principal bundle \( P(\xi) \) associated to \( \xi \). For example, the principal bundle associated with a vector bundle is the frame bundle, discussed at the end of Section 28.3.

Then, given two fibre bundles \( \xi \) and \( \xi' \), we see that \( \xi \) and \( \xi' \) are isomorphic iff \( P(\xi) \) and \( P(\xi') \) are isomorphic (Steenrod [164], Part I, Section 8, Theorem 8.2). More is true: The fibre bundle \( \xi \) is trivial iff the principal fibre bundle \( P(\xi) \) is trivial (this is easy to prove, do it! Otherwise, see Steenrod [164], Part I, Section 8, Corollary 8.4). Moreover, there is a test for the triviality of a principal bundle, the existence of a (global) section.

The following proposition, although easy to prove, is crucial:

**Proposition 28.17.** If \( \xi \) is a principal bundle, then \( \xi \) is trivial iff it possesses some global section.

*Proof.* If \( f : B \times G \to \xi \) is an isomorphism of principal bundles over the same base \( B \), then for every \( g \in G \), the map \( b \mapsto f(b,g) \) is a section of \( \xi \).

Conversely, let \( s : B \to E \) be a section of \( \xi \). Then, observe that the map \( f : B \times G \to \xi \) given by
\[
f(b,g) = s(b)g
\]
is a map of principal bundles. By Proposition 28.16, it is an isomorphism, so \( \xi \) is trivial. \( \Box \)
Generally, in geometry, many objects of interest arise as global sections of some suitable bundle (or sheaf): vector fields, differential forms, tensor fields, etc.

Given a principal bundle $\xi = (E, \pi, B, G)$ and given a manifold $F$, if $G$ acts effectively on $F$ from the left, again, using Theorem 28.4, we can construct a fibre bundle $\xi[F]$ from $\xi$, with $F$ as typical fibre, and such that $\xi[F]$ has the same transitions functions as $\xi$.

In the case of a principal bundle, there is another slightly more direct construction that takes us from principal bundles to fibre bundles (see Duistermaat and Kolk [64], Chapter 2, and Davis and Kirk [47], Chapter 4, Definition 4.6, where it is called the Borel construction). This construction is of independent interest, so we describe it briefly (for an application of this construction, see Duistermaat and Kolk [64], Chapter 2).

As $\xi$ is a principal bundle, recall that $G$ acts on $E$ from the right, so we have a right action of $G$ on $E \times F$, via

$$(z, f) \cdot g = (z \cdot g, g^{-1} \cdot f).$$

Consequently, we obtain the orbit set $E \times F/\sim$, denoted $E \times_G F$, where $\sim$ is the equivalence relation

$$(z, f) \sim (z', f') \iff (\exists g \in G)(z' = z \cdot g, f' = g^{-1} \cdot f).$$

Note that the composed map

$$E \times F \xrightarrow{pr_1} E \xrightarrow{\pi} B$$

factors through $E \times_G F$, since

$$\pi(pr_1(z, f)) = \pi(z) = \pi(z \cdot g) = \pi(pr_1(z \cdot g, g^{-1} \cdot f)) \cdot 1.$$

Let $p: E \times_G F \to B$ be the corresponding map. The following proposition is not hard to show:

**Proposition 28.18.** If $\xi = (E, \pi, B, G)$ is a principal bundle and $F$ is any manifold such that $G$ acts effectively on $F$ from the left, then, $\xi[F] = (E \times_G F, p, B, F, G)$ is a fibre bundle with fibre $F$ and structure group $G$, and $\xi[F]$ and $\xi$ have the same transition functions.

Let us verify that the charts of $\xi$ yield charts for $\xi[F]$. For any $U_\alpha$ in an open cover for $B$, we have a diffeomorphism

$$\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times G.$$

Observe that we have an isomorphism

$$(U_\alpha \times G) \times_G F \simeq U_\alpha \times F,$$

where, as usual, $G$ acts on $U_\alpha \times G$ via $(z, h) \cdot g = (z, hg)$, an isomorphism

$$p^{-1}(U_\alpha) \simeq \pi^{-1}(U_\alpha) \times_G F,$$
and that \( \varphi_\alpha \) induces an isomorphism

\[
\pi^{-1}(U_\alpha) \times_G F \xrightarrow{\varphi_\alpha} (U_\alpha \times G) \times_G F.
\]

So, we get the commutative diagram

\[
\begin{array}{ccc}
p^{-1}(U_\alpha) & \xrightarrow{\sim} & U_\alpha \times F \\
p \downarrow & & \downarrow \text{pr}_1 \\
U_\alpha & \xrightarrow{\sim} & U_\alpha,
\end{array}
\]

which yields a local trivialization for \( \xi[F] \). It is easy to see that the transition functions of \( \xi[F] \) are the same as the transition functions of \( \xi \).

The fibre bundle \( \xi[F] \) is called the fibre bundle induced by \( \xi \). Now, if we start with a fibre bundle \( \xi \) with fibre \( F \) and structure group \( G \), if we make the associated principal bundle \( P(\xi) \), and then the induced fibre bundle \( P(\xi)[F] \), what is the relationship between \( \xi \) and \( P(\xi)[F] \)?

The answer is: \( \xi \) and \( P(\xi)[F] \) are equivalent (this is because the transition functions are the same.)

Now, if we start with a principal \( G \)-bundle \( \xi \), make the fibre bundle \( \xi[F] \) as above, and then the principal bundle \( P(\xi[F]) \), we get a principal bundle equivalent to \( \xi \). Therefore, the maps

\[
\xi \mapsto \xi[F] \quad \text{and} \quad \xi \mapsto P(\xi)
\]

are mutual inverses, and they set up a bijection between equivalence classes of principal \( G \)-bundles over \( B \) and equivalence classes of fibre bundles over \( B \) (with structure group \( G \)). Moreover, this map extends to morphisms, so it is functorial (see Steenrod [164], Part I, Section 2, Lemma 2.6–Lemma 2.10).

As a consequence, in order to “classify” equivalence classes of fibre bundles (assuming \( B \) and \( G \) fixed), it is enough to know how to classify principal \( G \)-bundles over \( B \). Given some reasonable conditions on the coverings of \( B \), Milnor solved this classification problem, but this is taking us way beyond the scope of these notes!

The classical reference on fibre bundles, vector bundles and principal bundles, is Steenrod [164]. More recent references include Bott and Tu [24], Madsen and Tornehave [119], Morita [133], Griffith and Harris [80], Wells [178], Hirzebruch [92], Milnor and Stasheff [129], Davis and Kirk [47], Atiyah [13], Chern [40], Choquet-Bruhat, DeWitt-Morette and Dillard-Bleick [44], Hirsh [91], Sato [153], Narasimham [136], Sharpe [162] and also Husemoller [97], which covers more, including characteristic classes.

Proposition 28.14 shows that principal bundles are induced by suitable right actions, but we still need sufficient conditions to guarantee conditions (a), (b) and (c). The special situation of homogeneous spaces is considered in the next section.
28.6 Proper and Free Actions, Homogeneous Spaces Revisited

Now that we have introduced the notion of principal bundle, we can revisit the notion of homogeneous space given in Definition 5.9, which only applies to groups and sets without any topology or differentiable structure. We state stronger versions of the results about manifolds arising from group actions given in Section 19.2.

Before stating the main results of this section, observe that in the definition of a fibre bundle (Definition 28.1), the local trivialization maps are of the form

$$\varphi_{\alpha} : \pi^{-1}(U_{\alpha}) \to U_{\alpha} \times F,$$

where the fibre $F$ appears on the right. In particular, for a principal fibre bundle $\xi$, the fibre $F$ is equal to the structure group $G$, and this is the reason why $G$ acts on the right on the total space $E$ of $\xi$ (see Proposition 28.13). To be more precise, we could call such bundles right bundles. We can also define left bundles as bundles whose local trivialization maps are of the form

$$\varphi_{\alpha} : \pi^{-1}(U_{\alpha}) \to F \times U_{\alpha}.$$

Then, if $\xi$ is a left principal bundle, the group $G$ acts on $E$ on the left. Duistermaat and Kolk [64] address this issue at the end of their Appendix B, and prove the theorem stated below (Chapter 1, Section 11). Beware that in Duistermaat and Kolk [64], this theorem is stated for left bundles, which is a departure from the prevalent custom of dealing with right bundles. However, the weaker version that does not mention principal bundles is usually stated for left actions; for instance, in Lee [117] (Chapter 9, Theorem 9.16). We formulate both versions at the same time.

**Theorem 28.19.** Let $M$ be a smooth manifold, $G$ be a Lie group, and let $\cdot : G \times M \to M$ be a left (resp. right) smooth action which is proper and free. Then, $M/G$ is a principal left $G$-bundle (resp. right $G$-bundle) of dimension $\dim M - \dim G$. Moreover, the canonical projection $\pi : G \to M/G$ is a submersion (which means that $d\pi_g$ is surjective for all $g \in G$), and there is a unique manifold structure on $M/G$ with this property.

Theorem 28.19 has some interesting corollaries. Because a closed subgroup $H$ of a Lie group $G$ is a Lie group, and because the action of a closed subgroup is free and proper, we get the following result (proofs can also be found in Bröcker and tom Dieck [31] (Chapter I, Section 4) and in Duistermaat and Kolk [64] (Chapter 1, Section 11)).

**Theorem 28.20.** If $G$ is a Lie group and $H$ is a closed subgroup of $G$, then the right action of $H$ on $G$ defines a principal (right) $H$-bundle $\xi = (G, \pi, G/H, H)$, where $\pi : G \to G/H$ is the canonical projection. Moreover, $\pi$ is a submersion (which means that $d\pi_g$ is surjective for all $g \in G$), and there is a unique manifold structure on $G/H$ with this property.
In the special case where $G$ acts transitively on $M$, for any $x \in M$, if $G_x$ is the stabilizer of $x$, then with $H = G_x$, we get Proposition 28.21. Recall the definition of a homogeneous space.

**Definition 28.14.** A **homogeneous space** is a smooth manifold $M$ together with a smooth transitive action $\cdot : G \times M \to M$, of a Lie group $G$ on $M$.

The following result can be shown as a corollary of Theorem 28.20 and Theorem 19.10 (see Lee [117], Chapter 9, Theorem 9.24). It is also mostly proved in Bröcker and tom Dieck [31], Chapter I, Section 4):

**Proposition 28.21.** Let $\cdot : G \times M \to M$ be smooth transitive action of a Lie group $G$ on a manifold $M$. Then, $G/G_x$ and $M$ are diffeomorphic, and $G$ is the total space of a principal bundle $\xi = (G, \pi, M, G_x)$, where $G_x$ is the stabilizer of any element $x \in M$. Furthermore, the projection $\pi: G \to G/G_x$ is a submersion.

Thus, we finally see that homogeneous spaces induce principal bundles. Going back to some of the examples of Section 5.2, we see that

1. $\text{SO}(n+1)$ is a principal $\text{SO}(n)$-bundle over the sphere $S^n$ (for $n \geq 0$).
2. $\text{SU}(n+1)$ is a principal $\text{SU}(n)$-bundle over the sphere $S^{2n+1}$ (for $n \geq 0$).
3. $\text{SL}(2, \mathbb{R})$ is a principal $\text{SO}(2)$-bundle over the upper-half space $H$.
4. $\text{GL}(n, \mathbb{R})$ is a principal $\text{O}(n)$-bundle over the space $\text{SPD}(n)$ of symmetric, positive definite matrices.
5. $\text{GL}^+(n, \mathbb{R})$, is a principal $\text{SO}(n)$-bundle over the space $\text{SPD}(n)$ of symmetric, positive definite matrices, with fibre $\text{SO}(n)$.
6. $\text{SO}(n+1)$ is a principal $\text{O}(n)$-bundle over the real projective space $\mathbb{RP}^n$ (for $n \geq 0$).
7. $\text{SU}(n+1)$ is a principal $\text{U}(n)$-bundle over the complex projective space $\mathbb{CP}^n$ (for $n \geq 0$).
8. $\text{O}(n)$ is a principal $\text{O}(k) \times \text{O}(n-k)$-bundle over the Grassmannian $G(k, n)$.
9. $\text{SO}(n)$ is a principal $S(\text{O}(k) \times \text{O}(n-k))$-bundle over the Grassmannian $G(k, n)$.
10. $\text{SO}(n)$ is a principal $\text{SO}(n-k)$-bundle over the Stiefel manifold $S(k, n)$, with $1 \leq k \leq n-1$.
11. From Section 5.4, we see that the Lorentz group, $\text{SO}_0(n, 1)$, is a principal $\text{SO}(n)$-bundle over the space $\mathcal{H}_n^+(1)$, consisting of one group of the hyperbolic paraboloid $\mathcal{H}_n^+(1)$.

Thus, we see that both $\text{SO}(n+1)$ and $\text{SO}_0(n, 1)$ are principal $\text{SO}(n)$-bundles, the difference being that the base space for $\text{SO}(n+1)$ is the sphere $S^n$, which is compact, whereas the base space for $\text{SO}_0(n, 1)$ is the (connected) surface $\mathcal{H}_n^+(1)$, which is not compact. Many more examples can be given, for instance, see Arvanitoyeorgos [11].
Chapter 29

Connections and Curvature in Vector Bundles

29.1 Connections and Connection Forms in Vector Bundles

The goal of this chapter is to generalize the notions of connection and curvature to vector bundles. Among other things, this material has applications to theoretical physics. This chapter makes heavy use of differential forms (and tensor products), so the reader may want to brush up on these notions before reading it.

Given a manifold $M$, as $\mathfrak{X}(M) = \Gamma(M, TM) = \Gamma(TM)$, the set of smooth sections of the tangent bundle $TM$, it is natural that for a vector bundle $\xi = (E, \pi, B, V)$, a connection on $\xi$ should be some kind of bilinear map,

$$\mathfrak{X}(B) \times \Gamma(\xi) \longrightarrow \Gamma(\xi),$$

that tells us how to take the covariant derivative of sections.

Technically, it turns out that it is cleaner to define a connection on a vector bundle $\xi$, as an $\mathbb{R}$-linear map

$$\nabla : \Gamma(\xi) \to \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$$

that satisfies the “Leibniz rule”

$$\nabla(fs) = df \otimes s + f\nabla s,$$

with $s \in \Gamma(\xi)$ and $f \in C^\infty(B)$, where $\Gamma(\xi)$ and $\mathcal{A}^1(B)$ are treated as $C^\infty(B)$-modules. Since
\( \mathcal{A}^1(B) = \Gamma(B, T^*B) = \Gamma(T^*B) \) is the space of 1-forms on \( B \), and by Proposition 28.12,

\[
\mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) = \Gamma(T^*B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \Gamma(T^*B \otimes \xi) \\
\cong \Gamma(\text{Hom}(TB, \xi)) \\
\cong \text{Hom}_{C^\infty(B)}(\Gamma(TB), \Gamma(\xi)) \\
= \text{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi)),
\]

the range of \( \nabla \) can be viewed as a space of \( \Gamma(\xi) \)-valued differential forms on \( B \). Milnor and Stasheff [129] (Appendix C) use the version where

\[
\nabla: \Gamma(\xi) \to \Gamma(T^*B \otimes \xi),
\]

and Madsen and Tornehave [119] (Chapter 17) use the equivalent version stated in \((*)\). A thorough presentation of connections on vector bundles and the various ways to define them can be found in Postnikov [144] which also constitutes one of the most extensive references on differential geometry. Set

\[
\mathcal{A}^1(\xi) = \mathcal{A}^1(B; \xi) = \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi),
\]

and more generally, for any \( i \geq 0 \), set

\[
\mathcal{A}^i(\xi) = \mathcal{A}^i(B; \xi) = \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \Gamma\left( \bigwedge^i T^*B \otimes \xi \right).
\]

Obviously, \( \mathcal{A}^0(\xi) = \Gamma(\xi) \) (and recall that \( \mathcal{A}^0(B) = C^\infty(B) \)). The space of differential forms \( \mathcal{A}^i(B; \xi) \) with values in \( \Gamma(\xi) \) is a generalization of the space \( \mathcal{A}^i(M, F) \) of differential forms with values in \( F \) encountered in Section 23.4.

If we use the isomorphism

\[
\mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \text{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi)),
\]

then a connection is an \( \mathbb{R} \)-linear map

\[
\nabla: \Gamma(\xi) \to \text{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi))
\]

satisfying a Leibniz-type rule, or equivalently, an \( \mathbb{R} \)-bilinear map

\[
\nabla: \mathfrak{X}(B) \times \Gamma(\xi) \to \Gamma(\xi)
\]

such that, for any \( X \in \mathfrak{X}(B) \) and \( s \in \Gamma(\xi) \), if we write \( \nabla_Xs \) instead of \( \nabla(X, s) \), then the following properties hold for all \( f \in C^\infty(B) \):

\[
\nabla_{fX}s = f\nabla_Xs \\
\nabla_X(fs) = X[f]s + f\nabla_Xs.
\]
This second version may be considered simpler than the first since it does not involve a tensor product. Since

$$\mathcal{A}^1(B) = \Gamma(T^*B) \cong \text{Hom}_{C^\infty(B)}(\mathfrak{X}(B), C^\infty(B)) = \mathfrak{X}(B)^*,$$

using Proposition 21.26, the isomorphism

$$\alpha: \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \cong \text{Hom}_{C^\infty(B)}(\mathfrak{X}(B), \Gamma(\xi))$$

can be described in terms of the evaluation map

$$\text{Ev}_X: \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \to \Gamma(\xi),$$

given by

$$\text{Ev}_X(\omega \otimes s) = \omega(X)s, \quad X \in \mathfrak{X}(B), \ \omega \in \mathcal{A}^1(B), \ s \in \Gamma(\xi).$$

Namely, for any $\theta \in \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$,

$$\alpha(\theta)(X) = \text{Ev}_X(\theta).$$

In particular, we have

$$\text{Ev}_X(df \otimes s) = df(X)s = X[f]s.$$

Then, it is easy to see that we pass from the first version of $\nabla$, where

$$\nabla: \Gamma(\xi) \to \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi) \quad (*)$$

with the Leibniz rule

$$\nabla(fs) = df \otimes s + f \nabla s,$$

to the second version of $\nabla$, denoted $\nabla'$, where

$$\nabla': \mathfrak{X}(B) \times \Gamma(\xi) \to \Gamma(\xi) \quad (**)$$

is $\mathbb{R}$-bilinear and where the two conditions

$$\nabla'_{fX}s = f\nabla'_Xs$$

$$\nabla'_X(fs) = X[f]s + f\nabla'_Xs$$

hold, via the equation

$$\nabla'_X = \text{Ev}_X \circ \nabla.$$

From now on, we will simply write $\nabla_X s$ instead of $\nabla'_X s$, unless confusion arise. As summary of the above discussion, we make the following definition:
**Definition 29.1.** Let $\xi = (E, \pi, B, V)$ be a smooth real vector bundle. A connection on $\xi$ is an $\mathbb{R}$-linear map

$$\nabla : \Gamma(\xi) \to \mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$$

such that the Leibniz rule

$$\nabla(fs) = df \otimes s + f \nabla s$$

holds, for all $s \in \Gamma(\xi)$ and all $f \in C^\infty(B)$. For every $X \in \mathfrak{X}(B)$, we let

$$\nabla_X = Ev_X \circ \nabla,$$

and for every $s \in \Gamma(\xi)$, we call $\nabla_X s$ the covariant derivative of $s$ relative to $X$. Then, the family $(\nabla_X)$ induces a $\mathbb{R}$-bilinear map also denoted $\nabla$,

$$\nabla : \mathfrak{X}(B) \times \Gamma(\xi) \to \Gamma(\xi),$$

such that the following two conditions hold:

$$\nabla_{fX}s = f \nabla_X s \quad \nabla_X(fs) = X[f]s + f \nabla_X s,$$

for all $s \in \Gamma(\xi)$, all $X \in \mathfrak{X}(B)$ and all $f \in C^\infty(B)$. We refer to (*) as the first version of a connection and to (**) as the second version of a connection.

Observe that in terms of the $\mathcal{A}^i(\xi)$'s, a connection is a linear map,

$$\nabla : \mathcal{A}^0(\xi) \to \mathcal{A}^1(\xi),$$

satisfying the Leibniz rule. When $\xi = TB$, a connection (second version) is what is known as an affine connection on the manifold $B$.

**Remark:** Given two connections, $\nabla^1$ and $\nabla^2$, we have

$$\nabla^1(fs) - \nabla^2(fs) = df \otimes s + f \nabla^1 s - df \otimes s - f \nabla^2 s = f(\nabla^1 s - \nabla^2 s),$$

which shows that $\nabla^1 - \nabla^2$ is a $C^\infty(B)$-linear map from $\Gamma(\xi)$ to $\mathcal{A}^1(B) \otimes_{C^\infty(B)} \Gamma(\xi)$. However

\[
\text{Hom}_{C^\infty(B)}(\mathcal{A}^0(\xi), \mathcal{A}^i(\xi)) = \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi)) \\
\cong \Gamma(\xi)^* \otimes_{C^\infty(B)} (\mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi)) \\
\cong \mathcal{A}^i(B) \otimes_{C^\infty(B)} \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi)) \\
\cong \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\text{Hom}(\xi, \xi)) \\
= \mathcal{A}^i(\text{Hom}(\xi, \xi)).
\]
Therefore, $\nabla^1 - \nabla^2 \in \mathcal{A}^1(\text{Hom}(\xi, \xi))$, that is, it is a one-form with values in $\Gamma(\text{Hom}(\xi, \xi))$. But then, the vector space $\Gamma(\text{Hom}(\xi, \xi))$ acts on the space of connections (by addition) and makes the space of connections into an affine space. Given any connection, $\nabla$ and any one-form $\omega \in \Gamma(\text{Hom}(\xi, \xi))$, the expression $\nabla + \omega$ is also a connection. Equivalently, any affine combination of connections is also a connection.

A basic property of $\nabla$ is that it is a local operator.

**Proposition 29.1.** Let $\xi = (E, \pi, B, V)$ be a smooth real vector bundle and let $\nabla$ be a connection on $\xi$. For every open subset $U \subseteq B$, for every section $s \in \Gamma(\xi)$, if $s \equiv 0$ on $U$, then $\nabla s \equiv 0$ on $U$; that is, $\nabla$ is a local operator.

**Proof.** By Proposition 9.2 applied to the constant function with value 1, for every $p \in U$, there is some open subset, $V \subseteq U$, containing $p$ and a smooth function, $f : B \to \mathbb{R}$, such that $\text{supp } f \subseteq U$ and $f \equiv 1$ on $V$. Consequently, $fs$ is a smooth section which is identically zero. By applying the Leibniz rule, we get

$$0 = \nabla(fs) = df \otimes s + f \nabla s,$$

which, evaluated at $p$ yields $(\nabla s)(p) = 0$, since $f(p) = 1$ and $df \equiv 0$ on $V$. 

As an immediate consequence of Proposition 29.1, if $s_1$ and $s_2$ are two sections in $\Gamma(\xi)$ that agree on $U$, then $s_1 - s_2$ is zero on $U$, so $\nabla(s_1 - s_2) = \nabla s_1 - \nabla s_2$ is zero on $U$, that is, $\nabla s_1$ and $\nabla s_2$ agree on $U$.

Proposition 29.1 also implies that a connection, $\nabla$, on $\xi$, restricts to a connection, $\nabla | U$ on the vector bundle, $\xi | U$, for every open subset, $U \subseteq B$. Indeed, let $s$ be a section of $\xi$ over $U$. Pick any $b \in U$ and define $(\nabla s)(b)$ as follows: Using Proposition 9.2, there is some open subset, $V_1 \subseteq U$, containing $b$ and a smooth function, $f_1 : B \to \mathbb{R}$, such that $\text{supp } f_1 \subseteq U$ and $f_1 \equiv 1$ on $V_1$ so, let $s_1 = f_1 s$, a global section of $\xi$. Clearly, $s_1 = s$ on $V_1$, and set

$$(\nabla s)(b) = (\nabla s_1)(b).$$

This definition does not depend on $(V_1, f_1)$, because if we had used another pair, $(V_2, f_2)$, as above, since $b \in V_1 \cap V_2$, we have

$$s_1 = f_1 s = s = f_2 s = s_2 \quad \text{on} \quad V_1 \cap V_2$$

so, by Proposition 29.1,

$$\nabla s_1)(b) = (\nabla s_2)(b).$$

It should also be noted that $(\nabla_X s)(b)$ only depends on $X(b)$.

**Proposition 29.2.** for any two vector fields $X, Y \in \mathfrak{X}(B)$, if $X(b) = Y(b)$ for some $b \in B$, then

$$(\nabla_X s)(b) = (\nabla_Y s)(b), \quad \text{for every } s \in \Gamma(\xi).$$
Proof. As above, by linearity, it it enough to prove that if \( X(b) = 0 \), then \( (\nabla_X s)(b) = 0 \) (this argument is due to O’Neill [138], Chapter 2, Lemma 3). To prove this, pick any local trivialization, \((U, \varphi)\), with \( b \in U \). Then, we can write
\[
X \mid U = \sum_{i=1}^{d} X_i \frac{\partial}{\partial x_i}.
\]
However, as before, we can find a pair, \((V, f)\), with \( b \in V \subseteq U \), \( \text{supp } f \subseteq U \) and \( f = 1 \) on \( V \), so that \( f \frac{\partial}{\partial x_i} \) is a smooth vector field on \( B \) and \( f \frac{\partial}{\partial x_i} \) agrees with \( \frac{\partial}{\partial x_i} \) on \( V \), for \( i = 1, \ldots, n \).

Clearly, \( fX_i \in C^\infty(B) \) and \( fX_i \) agrees with \( X_i \) on \( V \) so if we write \( \tilde{X} = f^2 X \), then
\[
\tilde{X} = f^2 X = \sum_{i=1}^{d} fX_i \frac{\partial}{\partial x_i}
\]
and we have
\[
f^2 \nabla_X s = \nabla_{\tilde{X}} s = \sum_{i=1}^{d} fX_i \nabla f \frac{\partial}{\partial x_i} s.
\]
Since \( X_i(b) = 0 \) and \( f(b) = 1 \), we get \( (\nabla_X s)(b) = 0 \), as claimed. \( \square \)

Using the above property, for any point, \( p \in B \), we can define the covariant derivative \( (\nabla_u s)(p) \) of a section \( s \in \Gamma(\xi) \), with respect to a tangent vector \( u \in T_p B \).

Indeed, pick any vector field \( X \in \mathfrak{X}(B) \) such that \( X(p) = u \) (such a vector field exists locally over the domain of a chart and then extend it using a bump function) and set \( (\nabla_u s)(p) = (\nabla_X s)(p) \). By the above property, if \( X(p) = Y(p) \), then \( (\nabla_X s)(p) = (\nabla_Y s)(p) \) so \( (\nabla_u s)(p) \) is well-defined. Since \( \nabla \) is a local operator, \( (\nabla_u s)(p) \) is also well defined for any tangent vector \( u \in T_p B \), and any local section \( s \in \Gamma(U, \xi) \) defined in some open subset \( U \), with \( p \in U \). From now on, we will use this property without any further justification.

Since \( \xi \) is locally trivial, it is interesting to see what \( \nabla \mid U \) looks like when \((U, \varphi)\) is a local trivialization of \( \xi \).

Fix once and for all some basis \( (v_1, \ldots, v_n) \) of the typical fibre \( V \) \( (n = \text{dim}(V)) \). To every local trivialization \( \varphi: \pi^{-1}(U) \to U \times V \) of \( \xi \) (for some open subset, \( U \subseteq B \)), we associate the frame \( (s_1, \ldots, s_n) \) over \( U \), given by
\[
s_i(b) = \varphi^{-1}(b, v_i), \quad b \in U.
\]
Then, every section \( s \) over \( U \) can be written uniquely as \( s = \sum_{i=1}^{n} f_i s_i \), for some functions \( f_i \in C^\infty(U) \), and we have
\[
\nabla s = \sum_{i=1}^{n} \nabla (f_i s_i) = \sum_{i=1}^{n} (df_i \otimes s_i + f_i \nabla s_i).
\]
On the other hand, each $\nabla s_i$ can be written as

$$\nabla s_i = \sum_{j=1}^{n} \omega_{ij} \otimes s_j,$$

for some $n \times n$ matrix $\omega = (\omega_{ij})$ of one-forms $\omega_{ij} \in \mathcal{A}^1(U)$, so we get

$$\nabla s = \sum_{i=1}^{n} df_i \otimes s_i + \sum_{i=1}^{n} f_i \nabla s_i = \sum_{i=1}^{n} df_i \otimes s_i + \sum_{i,j=1}^{n} f_i \omega_{ij} \otimes s_j = \sum_{j=1}^{n} (df_j + \sum_{i=1}^{n} f_i \omega_{ij}) \otimes s_j.$$

With respect to the frame $(s_1, \ldots, s_n)$, the connection $\nabla$ has the matrix form

$$\nabla(f_1, \ldots, f_n) = (df_1, \ldots, df_n) + (f_1, \ldots, f_n) \omega,$$

and the matrix $\omega = (\omega_{ij})$ of one-forms $\omega_{ij} \in \mathcal{A}^1(U)$ is called the connection form or connection matrix of $\nabla$ with respect to $\varphi: \pi^{-1}(U) \to U \times V$.

The above computation also shows that on $U$, any connection is uniquely determined by a matrix of one-forms, $\omega_{ij} \in \mathcal{A}^1(U)$. In particular, the connection on $U$ for which

$$\nabla s_1 = 0, \ldots, \nabla s_n = 0,$$

corresponding to the zero matrix is called the flat connection on $U$ (w.r.t. $(s_1, \ldots, s_n)$).

Some authors (such as Morita [133]) use a notation involving subscripts and superscripts, namely

$$\nabla s_i = \sum_{j=1}^{n} \omega^j_i \otimes s_j.$$

But, beware, the expression $\omega = (\omega^j_i)$ denotes the $n \times n$-matrix whose rows are indexed by $j$ and whose columns are indexed by $i$, and we have $\omega_{ij} = \omega^j_i$. Accordingly, if $\theta = \omega \eta$, then

$$\theta^j_i = \sum_{k} \omega^j_k \eta^k_i.$$

The matrix, $(\omega^j_i)$ is thus the transpose of our matrix $(\omega_{ij})$. This has the effects that some of the results differ either by a sign (as in $\omega \wedge \omega$) or by a permutation of matrices (as in the formula for a change of frame).

**Remark:** If $(\theta_1, \ldots, \theta_n)$ is the dual frame of $(s_1, \ldots, s_n)$, that is, $\theta_i \in \mathcal{A}^1(U)$ is the one-form defined so that

$$\theta_i(b)(s_j(b)) = \delta_{ij}, \quad \text{for all} \quad b \in U, \ 1 \leq i, j \leq n,$$

then we can write $\omega_{ik} = \sum_{j=1}^{n} \Gamma^k_{ji} \theta_j$ and so,

$$\nabla s_i = \sum_{j,k=1}^{n} \Gamma^k_{ji}(\theta_j \otimes s_k),$$

where the $\Gamma^k_{ji} \in C^\infty(U)$ are the Christoffel symbols.
Proposition 29.3. Every vector bundle $\xi$ possesses a connection.

Proof. Since $\xi$ is locally trivial, we can find a locally finite open cover $(U_\alpha)_\alpha$ of $B$ such that $\pi^{-1}(U_\alpha)$ is trivial. If $(f_\alpha)$ is a partition of unity subordinate to the cover $(U_\alpha)_\alpha$ and if $\nabla^\alpha$ is any flat connection on $\xi|_{U_\alpha}$, then it is immediately verified that

$$
\nabla = \sum_\alpha f_\alpha \nabla^\alpha
$$

is a connection on $\xi$. \qed

If $\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi_\beta: \pi^{-1}(U_\beta) \to U_\beta \times V$ are two overlapping trivializations, we know that for every $b \in U_\alpha \cap U_\beta$, we have

$$
\varphi_\alpha \circ \varphi_\beta^{-1}(b, u) = (b, g_{\alpha\beta}(b)u),
$$

where $g_{\alpha\beta}: U_\alpha \cap U_\beta \to \text{GL}(V)$ is the transition function. As

$$
\varphi_\beta^{-1}(b, u) = \varphi_\alpha^{-1}(b, g_{\alpha\beta}(b)u),
$$

if $(s_1, \ldots, s_n)$ is the frame over $U_\alpha$ associated with $\varphi_\alpha$ and $(t_1, \ldots, t_n)$ is the frame over $U_\beta$ associated with $\varphi_\beta$, we see that

$$
t_i = \sum_{j=1}^n g_{ij} s_j,
$$

where $g_{\alpha\beta} = (g_{ij})$.

Proposition 29.4. With the notations as above, the connection matrices, $\omega_\alpha$ and $\omega_\beta$ respectively over $U_\alpha$ and $U_\beta$ obey the transformation rule

$$
\omega_\beta = g_{\alpha\beta} \omega_\alpha g_{\alpha\beta}^{-1} + (dg_{\alpha\beta}) g_{\alpha\beta}^{-1},
$$

where $dg_{\alpha\beta} = (dg_{ij})$.

Proof. To prove the above proposition, apply $\nabla$ to both side of the equations

$$
t_i = \sum_{j=1}^n g_{ij} s_j
$$

and use $\omega_\alpha$ and $\omega_\beta$ to express $\nabla t_i$ and $\nabla s_j$. The details are left as an exercise. \qed

In Morita [133] (Proposition 5.22), the order of the matrices in the equation of Proposition 29.4 must be reversed.
29.1. CONNECTIONS IN VECTOR BUNDLES AND RIEMANNIAN MANIFOLDS

If $\xi = TM$, the tangent bundle of some smooth manifold $M$, then a connection on $TM$, also called a connection on $M$, is a linear map

$$\nabla: \mathfrak{X}(M) \rightarrow \mathcal{A}^1(M) \otimes_{C^\infty(M)} \mathfrak{X}(M) \cong \text{Hom}_{C^\infty(M)}(\mathfrak{X}(M), (\mathfrak{X}(M))),$$

since $\Gamma(TM) = \mathfrak{X}(M)$. Then, for fixed $Y \in \mathcal{X}(M)$, the map $\nabla Y$ is $C^\infty(M)$-linear, which implies that $\nabla Y$ is a $(1,1)$ tensor. In a local chart, $(U, \varphi)$, we have

$$\nabla_{\partial / \partial x_i} \left( \partial / \partial x_j \right) = \sum_{k=1}^{n} \Gamma^k_{ij} \frac{\partial}{\partial x_k},$$

where the $\Gamma^k_{ij}$ are Christoffel symbols.

The covariant derivative $\nabla_X$ given by a connection $\nabla$ on $TM$ can be extended to a covariant derivative $\nabla_{r,s}$, defined on tensor fields in $\Gamma(M, T^{r,s}(M))$, for all $r, s \geq 0$, where

$$T^{r,s}(M) = T^{r} \otimes M \otimes (T^* M)^{\otimes s}.$$

We already have $\nabla^1_0 = \nabla_X$ and it is natural to set $\nabla^0_0 f = X[f] = df(X)$. Recall that there is an isomorphism between the set of tensor fields $\Gamma(M, T^{r,s}(M))$, and the set of $C^\infty(M)$-multilinear maps

$$\Phi: \mathcal{A}^1(M) \times \cdots \times \mathcal{A}^1(M) \times \mathfrak{X}(M) \times \cdots \times \mathfrak{X}(M) \rightarrow C^\infty(M),$$

where $\mathcal{A}^1(M)$ and $\mathfrak{X}(M)$ are $C^\infty(M)$-modules.

The next proposition is left as an exercise. For help, see O’Neill [138], Chapter 2, Proposition 13 and Theorem 15.

**Proposition 29.5.** for every vector field $X \in \mathfrak{X}(M)$, there is a unique family of $\mathbb{R}$-linear map $\nabla_{r,s}: \Gamma(M, T^{r,s}(M)) \rightarrow \Gamma(M, T^{r,s}(M))$, with $r, s \geq 0$, such that

(a) $\nabla^0_0 f = df(X)$, for all $f \in C^\infty(M)$ and $\nabla^1_0 = \nabla_X$, for all $X \in \mathfrak{X}(M)$.

(b) $\nabla^{r_1+r_2,s_1+s_2}_X (S \otimes T) = \nabla^{r_1,s_1}_X (S) \otimes T + S \otimes \nabla^{r_2,s_2}_X (T)$, for all $S \in \Gamma(M, T^{r_1,s_1}(M))$ and all $T \in \Gamma(M, T^{r_2,s_2}(M))$.

(c) $\nabla^{r-1,s-1}_{X}(c_{ij}(S)) = c_{ij}(\nabla^s_X (S))$, for all $S \in \Gamma(M, T^{r,s}(M))$ and all contractions, $c_{ij}$, of $\Gamma(M, T^{r,s}(M))$.

Furthermore,

$$\nabla^{0,1}_X \theta(Y) = X[\theta(Y)] - \theta(\nabla_X Y),$$
for all \(X, Y \in \mathfrak{X}(M)\) and all one-forms, \(\theta \in \mathcal{A}^1(M)\), and for every \(S \in \Gamma(M, T^{r,s}(M))\), with \(r + s \geq 2\), the covariant derivative \(\nabla^r_s X(S)\) is given by

\[
(\nabla^r_s X(S)(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s) = X[S(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s)] - \sum_{i=1}^r S(\theta_1, \ldots, \nabla^0_{X} \theta_i, \ldots, \theta_r, X_1, \ldots, X_s) - \sum_{j=1}^s S(\theta_1, \ldots, \theta_r, X_1, \ldots, \nabla X_j, \ldots, X_s),
\]

for all \(X_1, \ldots, X_s \in \mathfrak{X}(M)\) and all one-forms, \(\theta_1, \ldots, \theta_r \in \mathcal{A}^1(M)\).

In particular, for \(S = g\), the Riemannian metric on \(M\) (a \((0, 2)\) tensor), we get

\[
\nabla_X(g(Y, Z)) = X(g(Y, Z)) - g(\nabla_X Y, Z) - g(Y, \nabla_X Z),
\]

for all \(X, Y, Z \in \mathfrak{X}(M)\). We will see later on that a connection on \(M\) is compatible with a metric \(g\) iff \(\nabla_X(g) = 0\).

We define the covariant differential \(\nabla^r_s X S\) of a tensor \(S \in \Gamma(M, T^{r,s}(M))\) as the \((r, s + 1)\)-tensor given by

\[
(\nabla^r_s X S)(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s) = (\nabla^r_s \theta)(\theta_1, \ldots, \theta_r, X_1, \ldots, X_s),
\]

for all \(X, X_j \in \mathfrak{X}(M)\) and all \(\theta_i \in \mathcal{A}^1(M)\). For simplicity of notation we usually omit the superscripts \(r\) and \(s\). In particular, if \(r = 1\) and \(s = 0\), in which case \(S\) is a vector field, the covariant derivative \(\nabla S\) is defined so that

\[
(\nabla S)(X) = \nabla_X S.
\]

Everything we did in this section applies to complex vector bundles by considering complex vector spaces instead of real vector spaces, \(\mathbb{C}\)-linear maps instead of \(\mathbb{R}\)-linear map, and the space of smooth complex-valued functions, \(C^\infty(B; \mathbb{C}) \cong C^\infty(B) \otimes_{\mathbb{R}} \mathbb{C}\). We also use spaces of complex-valued differentials forms,

\[
\mathcal{A}^i(B; \mathbb{C}) = \mathcal{A}^i(B) \otimes_{C^\infty(B)} C^\infty(B; \mathbb{C}) \cong \Gamma\left(\bigwedge^i T^* B \otimes \mathcal{E}^1_{\mathbb{C}}\right),
\]

where \(\mathcal{E}^1_{\mathbb{C}}\) is the trivial complex line bundle, \(B \times \mathbb{C}\), and we define \(\mathcal{A}^i(\xi)\) as

\[
\mathcal{A}^i(\xi) = \mathcal{A}^i(B; \mathbb{C}) \otimes_{C^\infty(B; \mathbb{C})} \Gamma(\xi).
\]

A connection is a \(\mathbb{C}\)-linear map, \(\nabla : \Gamma(\xi) \to \mathcal{A}^1(\xi)\), that satisfies the same Leibniz-type rule as before. Obviously, every differential form in \(\mathcal{A}^i(B; \mathbb{C})\) can be written uniquely as \(\omega + i\eta\), with \(\omega, \eta \in \mathcal{A}^i(B)\). The exterior differential,

\[
d : \mathcal{A}^i(B; \mathbb{C}) \to \mathcal{A}^{i+1}(B; \mathbb{C})
\]

is defined by \(d(\omega + i\eta) = d\omega + id\eta\). We obtain complex-valued de Rham cohomology groups,

\[
H^i_{DR}(M; \mathbb{C}) = H^i_{DR}(M) \otimes_{\mathbb{R}} \mathbb{C}.
\]
29.2 Curvature, Curvature Form and Curvature Matrix

If $\xi = B \times V$ is the trivial bundle and $\nabla$ is a flat connection on $\xi$, we obviously have

$$\nabla_X \nabla_Y - \nabla_Y \nabla_X = \nabla_{[X,Y]}$$

where $[X,Y]$ is the Lie bracket of the vector fields $X$ and $Y$. However, for general bundles and arbitrary connections, the above fails. The error term,

$$R(X,Y) = \nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]}$$

measures what’s called the curvature of the connection. The curvature of a connection also turns up as the failure of a certain sequence involving the spaces $A^i(\xi)$ to be a cochain complex. Recall that a connection on $\xi$ is a linear map $\nabla: A^0(\xi) \to A^1(\xi)$ satisfying a Leibniz-type rule. It is natural to ask whether $\nabla$ can be extended to a family of operators, $d^\nabla: A^i(\xi) \to A^{i+1}(\xi)$, with properties analogous to $d$ on $A^*(B)$.

This is indeed the case, and we get a sequence of maps

$$0 \to A^0(\xi) \xrightarrow{\nabla} A^1(\xi) \xrightarrow{d^\nabla} A^2(\xi) \xrightarrow{\cdots} A^i(\xi) \xrightarrow{d^\nabla} A^{i+1}(\xi) \xrightarrow{\cdots},$$

but in general, $d^\nabla \circ d^\nabla = 0$ fails. In particular, $d^\nabla \circ \nabla = 0$ generally fails. The term $R^\nabla = d^\nabla \circ \nabla$ is the curvature form (or curvature tensor) of the connection $\nabla$. As we will see, it yields our previous curvature $R$, back.

Our next goal is to define $d^\nabla$. For this, we first define a $C^\infty(B)$-bilinear map

$$\wedge: A^i(\xi) \times A^j(\eta) \to A^{i+j}(\xi \otimes \eta)$$

as follows:

$$(\omega \otimes s) \wedge (\tau \otimes t) = (\omega \wedge \tau) \otimes (s \otimes t),$$

where $\omega \in A^i(B)$, $\tau \in A^j(B)$, $s \in \Gamma(\xi)$, and $t \in \Gamma(\eta)$, and where we used the fact that

$$\Gamma(\xi \otimes \eta) = \Gamma(\xi) \otimes_{C^\infty(B)} \Gamma(\eta).$$

First, consider the case where $\xi = \epsilon^1 = B \times \mathbb{R}$, the trivial line bundle over $B$. In this case, $A^i(\xi) = A^i(B)$ and we have a bilinear map

$$\wedge: A^i(B) \times A^j(\eta) \to A^{i+j}(\eta)$$

given by

$$\omega \wedge (\tau \otimes t) = (\omega \wedge \tau) \otimes t.$$
For \( j = 0 \), we have the bilinear map
\[
\wedge : \mathcal{A}^i(B) \times \Gamma(\eta) \to \mathcal{A}^i(\eta)
\]
given by
\[
\omega \wedge t = \omega \otimes t.
\]
It is clear that the bilinear map
\[
\wedge : \mathcal{A}^r(B) \times \mathcal{A}^s(\eta) \to \mathcal{A}^{r+s}(\eta)
\]
has the following properties:
\[
(\omega \wedge \tau) \wedge \theta = \omega \wedge (\tau \wedge \theta)
\]
\[
1 \wedge \theta = \theta,
\]
for all \( \omega \in \mathcal{A}^i(B), \tau \in \mathcal{A}^j(B), \theta \in \mathcal{A}^k(\xi) \), and where 1 denotes the constant function in \( C^\infty(B) \) with value 1.

**Proposition 29.6.** For every vector bundle \( \xi \), for all \( j \geq 0 \), there is a unique \( \mathbb{R} \)-linear map (resp. \( \mathbb{C} \)-linear if \( \xi \) is a complex VB) \( d^\nabla : \mathcal{A}^j(\xi) \to \mathcal{A}^{j+1}(\xi) \), such that

(i) \( d^\nabla = \nabla \) for \( j = 0 \).

(ii) \( d^\nabla (\omega \wedge t) = d\omega \wedge t + (-1)^i \omega \wedge d^\nabla t \), for all \( \omega \in \mathcal{A}^i(B) \) and all \( t \in \mathcal{A}^j(\xi) \).

**Proof.** Recall that \( \mathcal{A}^i(B) = \mathcal{A}^i(\xi) \otimes_{C^\infty(B)} \Gamma(\xi) \), and define \( d^\nabla : \mathcal{A}^i(B) \times \Gamma(\xi) \to \mathcal{A}^{i+1}(\xi) \) by
\[
d^\nabla (\omega, s) = d\omega \otimes s + (-1)^i \omega \wedge \nabla s,
\]
for all \( \omega \in \mathcal{A}^i(B) \) and all \( s \in \Gamma(\xi) \). We claim that \( d^\nabla \) induces an \( \mathbb{R} \)-linear map on \( \mathcal{A}^i(\xi) \), but there is a complication as \( d^\nabla \) is not \( C^\infty(B) \)-bilinear. The way around this problem is to use Proposition 21.27. For this, we need to check that \( d^\nabla \) satisfies the condition of Proposition 21.27, where the right action of \( C^\infty(B) \) on \( \mathcal{A}^i(B) \) is equal to the left action, namely wedging:
\[
f \wedge \omega = \omega \wedge f \quad f \in C^\infty(B) = \mathcal{A}^0(B), \omega \in \mathcal{A}^i(B).
\]
As \( \wedge \) is \( C^\infty(B) \)-bilinear and \( \tau \otimes s = \tau \wedge s \) for all \( \tau \in \mathcal{A}^i(B) \) and all \( s \in \Gamma(\xi) \), we have
\[
d^\nabla (\omega f, s) = d(\omega f) \otimes s + (-1)^i (\omega f) \wedge \nabla s
\]
\[
= d(\omega f) \wedge s + (-1)^i f \omega \wedge \nabla s
\]
\[
= ((d\omega)f + (-1)^i \omega \wedge df) \wedge s + (-1)^j f \omega \wedge \nabla s
\]
\[
= f d\omega \wedge s + (-1)^j \omega \wedge df \wedge s + (-1)^j f \omega \wedge \nabla s
\]
and

\[ d^\nabla (\omega, fs) = d\omega \otimes (fs) + (-1)^i \omega \land \nabla (fs) \]
\[ = d\omega \land (fs) + (-1)^i \omega \land \nabla (fs) \]
\[ = f d\omega \land s + (-1)^i \omega \land (df \otimes s + f \nabla s) \]
\[ = f d\omega \land s + (-1)^i \omega \land (df \land s + f \nabla s) \]
\[ = f d\omega \land s + (-1)^i \omega \land df \land s + (-1)^i f \omega \land \nabla s. \]

Thus, \( d^\nabla (\omega, fs) = d^\nabla (\omega, fs) \), and Proposition 21.27 shows that \( d^\nabla : A^i(\xi) \to A^{i+1}(\xi) \) is a well-defined \( \mathbb{R} \)-linear map for all \( j \geq 0 \). Furthermore, it is clear that \( d^\nabla = \nabla \) for \( j = 0 \). Now, for \( \omega \in A^i(B) \) and \( t = \tau \otimes s \in A^j(\xi) \) we have

\[ d^\nabla (\omega \land (\tau \otimes s)) = d^\nabla ((\omega \land \tau) \otimes s) \]
\[ = d(\omega \land \tau) \otimes s + (-1)^{i+j} (\omega \land \tau) \land \nabla s \]
\[ = (d\omega \land \tau) \otimes s + (-1)^i (\omega \land d\tau) \otimes s + (-1)^{i+j} (\omega \land \tau) \land \nabla s \]
\[ = d\omega \land (\tau \otimes s) + (-1)^i \omega \land (d\tau \otimes s) + (-1)^{i+j} \omega \land (\tau \land \nabla s) \]
\[ = d\omega \land (\tau \otimes s) + (-1)^i \omega \land d^\nabla (\tau \land s), \]
\[ = d\omega \land (\tau \otimes s) + (-1)^i \omega \land d^\nabla (\tau \otimes s), \]

which proves (ii).

As a consequence, we have the following sequence of linear maps

\[ 0 \longrightarrow A^0(\xi) \overset{\nabla}{\longrightarrow} A^1(\xi) \overset{d^\nabla}{\longrightarrow} A^2(\xi) \longrightarrow \cdots \longrightarrow A^i(\xi) \overset{d^\nabla}{\longrightarrow} A^{i+1}(\xi) \longrightarrow \cdots. \]

but in general, \( d^\nabla \circ d^\nabla = 0 \) fails. Although generally \( d^\nabla \circ \nabla = 0 \) fails, the map \( d^\nabla \circ \nabla \) is \( C^\infty(B) \)-linear. Indeed,

\[ (d^\nabla \circ \nabla)(fs) = d^\nabla (df \otimes s + f \nabla s) \]
\[ = d^\nabla (df \land s + f \land \nabla s) \]
\[ = ddf \land s - df \land \nabla s + df \land \nabla s + f \land d^\nabla (\nabla s) \]
\[ = f((d^\nabla \circ \nabla)(s)). \]

Therefore, \( d^\nabla \circ \nabla : A^0(\xi) \to A^2(\xi) \) is a \( C^\infty(B) \)-linear map. However, recall that just before Proposition 29.1 we showed that

\[ \text{Hom}_{C^\infty(B)}(A^0(\xi), A^i(\xi)) \cong A^i(\text{Hom}(\xi, \xi)), \]

therefore, \( d^\nabla \circ \nabla \in A^2(\text{Hom}(\xi, \xi)) \); that is, \( d^\nabla \circ \nabla \) is a two-form with values in \( \Gamma(\text{Hom}(\xi, \xi)) \).

Although this is far from obvious from Definition 29.2, the curvature form \( R^\nabla \) is related to the curvature \( R(X, Y) \) defined at the beginning of Section 29.2. For this, we need to explain
how to define $R^\nabla_{X,Y}(s)$, for any two vector fields $X, Y \in \mathfrak{X}(B)$ and any section $s \in \Gamma(\xi)$. For any section $s \in \Gamma(\xi)$, the value $\nabla s$ can be written as a linear combination of elements of the form $\omega \otimes t$, with $\omega \in \mathcal{A}^1(B)$ and $t \in \Gamma(\xi)$. If $\nabla s = \omega \otimes t = \omega \wedge t$, as above, we have

$$d^\nabla (\nabla s) = d^\nabla (\omega \wedge t)$$

$$= d\omega \otimes t - \omega \wedge \nabla t.$$

But, $\nabla t$ itself is a linear combination of the form

$$\nabla t = \sum_j \eta_j \otimes t_j$$

for some 1-forms $\eta_j \in \mathcal{A}^1(B)$ and some sections $t_j \in \Gamma(\xi)$, so we have

$$d^\nabla (\nabla s) = d\omega \otimes t - \sum_j (\omega \wedge \eta_j) \otimes t_j.$$

Thus, it makes sense to define $R^\nabla_{X,Y}(s)$ by

$$R^\nabla_{X,Y}(s) = d\omega(X,Y)t - \sum_j (\omega \wedge \eta_j)(X,Y)t_j$$

$$= d\omega(X,Y)t - \sum_j (\omega(X)\eta_j(Y) - \omega(Y)\eta_j(X))t_j$$

$$= d\omega(X,Y)t - \left(\omega(X) \sum_j \eta_j(Y)t_j - \omega(Y) \sum_j \eta_j(X)t_j\right)$$

$$= d\omega(X,Y)t - (\omega(X)\nabla_Y t - \omega(Y)\nabla_X t),$$

since $\nabla_X t = \sum_j \eta_j(X)t_j$ because $\nabla t = \sum_j \eta_j \otimes t_j$, and similarly for $\nabla_Y t$. We extend this formula by linearity when $\nabla s$ is a linear combinations of elements of the form $\omega \otimes t$. A clean way to define $R^\nabla_{X,Y}$ is to define the evaluation map

$$\text{Ev}_{X,Y} : \mathcal{A}^2(\mathcal{H}om(\xi, \xi)) \to \mathcal{A}^0(\mathcal{H}om(\xi, \xi)) = \Gamma(\mathcal{H}om(\xi, \xi)) \cong \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi))$$

as follows: For all $X, Y \in \mathfrak{X}(B)$, all $\theta \otimes h \in \mathcal{A}^2(\mathcal{H}om(\xi, \xi)) = \mathcal{A}^2(B) \otimes_{C^\infty(B)} \Gamma(\mathcal{H}om(\xi, \xi))$, set

$$\text{Ev}_{X,Y}(\theta \otimes h) = \theta(X,Y)h.$$}

It is clear that this map is $C^\infty(B)$-linear and thus well-defined on $\mathcal{A}^2(\mathcal{H}om(\xi, \xi))$. (Recall that $\mathcal{A}^0(\mathcal{H}om(\xi, \xi)) = \Gamma(\mathcal{H}om(\xi, \xi)) = \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi))$.) We write

$$R^\nabla_{X,Y} = \text{Ev}_{X,Y}(R^\nabla) \in \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi)).$$

Since $R^\nabla$ is a linear combination of the form

$$R^\nabla = \sum_j \theta_j \otimes h_j$$
for some 2-forms $\theta_j \in \mathcal{A}^2(B)$ and some sections $h_j \in \Gamma(\mathcal{H}om(\xi, \xi))$, for any section $s \in \Gamma(\xi)$, we have

$$R_{X,Y}^\nabla(s) = \sum_j \theta_j(X,Y)h_j(s),$$

where $h_j(s)$ is some section in $\Gamma(\xi)$, and then we use the formula obtained above when $\nabla s$ is a linear combination of terms of the form $\omega \otimes s$ for some 1-forms $\mathcal{A}^1(B)$ and some sections $s \in \Gamma(\xi)$.

**Proposition 29.7.** For any vector bundle $\xi$, and any connection $\nabla$ on $\xi$, for all $X,Y \in \mathfrak{X}(B)$, if we let

$$R(X,Y) = \nabla_X \circ \nabla_Y - \nabla_Y \circ \nabla_X - \nabla_{[X,Y]},$$

then

$$R(X,Y) = R_{X,Y}^\nabla.$$

**Proof.** Since for any section $s \in \Gamma(\xi)$, the value $\nabla s$ can be written as a linear combination of elements of the form $\omega \otimes t = \omega \wedge t$, with $\omega \in \mathcal{A}^1(B)$ and $t \in \Gamma(\xi)$, it is sufficient to compute $R_{X,Y}^\nabla(s)$ when $\nabla s = \omega \otimes t$, and we get

$$R_{X,Y}^\nabla(s) = d\omega(X,Y)t - (\omega(X)\nabla_Y t - \omega(Y)\nabla_X t)$$

$$= (X(\omega(Y)) - Y(\omega(X)) - \omega([X,Y]))s - (\omega(X)\nabla_Y t - \omega(Y)\nabla_X t)$$

$$= \nabla_X(\omega(Y)t) - \nabla_Y(\omega(X)t) - \omega([X,Y])t$$

$$= \nabla_X(\nabla_Y s) - \nabla_Y(\nabla_X s) - \nabla_{[X,Y]} s,$$

since $\nabla_X s = \omega(X)t$ because $\nabla s = \omega \otimes t$ (and similarly for the other terms involving $\omega$). \qed

**Remark:** Proposition 29.7 implies that $R(Y,X) = -R(X,Y)$ and that $R(X,Y)(s)$ is $C^\infty(B)$-linear in $X, Y$ and $s$.

**Definition 29.2.** For any vector bundle $\xi$ and any connection $\nabla$ on $\xi$, the vector-valued two-form $R^\nabla = d\nabla \circ \nabla \in \mathcal{A}^2(\mathcal{H}om(\xi, \xi))$ is the curvature form (or curvature tensor) of the connection $\nabla$. We say that $\nabla$ is a flat connection iff $R^\nabla = 0$.

The expression $R^\nabla$ is also denoted $F^\nabla$ or $K^\nabla$.

As in the case of a connection, we can express $R^\nabla$ locally in any local trivialization $\varphi: \pi^{-1}(U) \to U \times V$ of $\xi$. Since $R^\nabla \in \mathcal{A}^2(\xi) = \mathcal{A}^2(B) \otimes_{C^\infty(B)} \Gamma(\mathcal{H}om(\xi, \xi))$, if $(s_1, \ldots, s_n)$ is the frame associated with $(\varphi, U)$, then

$$R^\nabla(s_i) = \sum_{j=1}^n \Omega_{ij} \otimes s_j,$$

for some matrix $\Omega = (\Omega_{ij})$ of two forms $\Omega_{ij} \in \mathcal{A}^2(B)$. We call $\Omega = (\Omega_{ij})$ the curvature matrix (or curvature form) associated with the local trivialization. The relationship between the connection form $\omega$ and the curvature form $\Omega$ is simple.
Proposition 29.8. (Structure Equations) Let $\xi$ be any vector bundle and let $\nabla$ be any connection on $\xi$. For every local trivialization $\varphi: \pi^{-1}(U) \to U \times V$, the connection matrix $\omega = (\omega_{ij})$ and the curvature matrix $\Omega = (\Omega_{ij})$ associated with the local trivialization $(\varphi, U)$, are related by the structure equation:

$$\Omega = d\omega - \omega \wedge \omega.$$  

Proof. By definition,

$$\nabla(s_i) = \sum_{j=1}^{n} \omega_{ij} \otimes s_j,$$

so if we apply $d\nabla$ and use property (ii) of Proposition 29.6 we get

$$d\nabla(\nabla(s_i)) = \sum_{k=1}^{n} \Omega_{ik} \otimes s_k$$

$$= \sum_{j=1}^{n} d\nabla(\omega_{ij} \otimes s_j)$$

$$= \sum_{j=1}^{n} d\omega_{ij} \otimes s_j - \sum_{j=1}^{n} \omega_{ij} \wedge \nabla s_j$$

$$= \sum_{j=1}^{n} d\omega_{ij} \otimes s_j - \sum_{j=1}^{n} \omega_{ij} \wedge \sum_{k=1}^{n} \omega_{jk} \otimes s_k$$

$$= \sum_{k=1}^{n} d\omega_{ik} \otimes s_k - \sum_{k=1}^{n} \left( \sum_{j=1}^{n} \omega_{ij} \wedge \omega_{jk} \right) \otimes s_k,$$

and so,

$$\Omega_{ik} = d\omega_{ik} - \sum_{j=1}^{n} \omega_{ij} \wedge \omega_{jk},$$

which, means that

$$\Omega = d\omega - \omega \wedge \omega,$$

as claimed.\hfill \Box

Some other texts, including Morita [133] (Theorem 5.21) state the structure equations as

$$\Omega = d\omega + \omega \wedge \omega.$$

If $\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi_\beta: \pi^{-1}(U_\beta) \to U_\beta \times V$ are two overlapping trivializations, the relationship between the curvature matrices $\Omega_\alpha$ and $\Omega_\beta$, is given by the following proposition which is the counterpart of Proposition 29.4 for the curvature matrix:
Proposition 29.9. If $\varphi_\alpha: \pi^{-1}(U_\alpha) \to U_\alpha \times V$ and $\varphi_\beta: \pi^{-1}(U_\beta) \to U_\beta \times V$ are two overlapping trivializations of a vector bundle $\xi$, then we have the following transformation rule for the

curvature matrices $\Omega_\alpha$ and $\Omega_\beta$:

$$\Omega_\beta = g_{\alpha\beta} \Omega_\alpha g_{\alpha\beta}^{-1},$$

where $g_{\alpha\beta}: U_\alpha \cap U_\beta \to \text{GL}(V)$ is the transition function.

Proof sketch. Use the structure equations (Proposition 29.8) and apply $d$ to the equations of Proposition 29.4. $\square$

Proposition 29.8 also yields a formula for $d\Omega$, known as Bianchi’s identity (in local form).

Proposition 29.10. (Bianchi’s Identity) For any vector bundle $\xi$ and any connection $\nabla$ on $\xi$, if $\omega$ and $\Omega$ are respectively the connection matrix and the curvature matrix, in some local trivialization, then

$$d\Omega = \omega \wedge \Omega - \Omega \wedge \omega.$$  

Proof. If we apply $d$ to the structure equation, $\Omega = d\omega - \omega \wedge \omega$, we get

$$d\Omega = \begin{align*} dd\omega - d\omega \wedge \omega &+ \omega \wedge d\omega \\ = -(\Omega + \omega \wedge \omega) &+ \omega \wedge (\Omega + \omega \wedge \omega) \\ = -\Omega &+ \omega - \omega \wedge \omega \wedge \omega + \omega \wedge \Omega + \omega \wedge \omega \wedge \omega \\ = \omega &+ \Omega - \Omega \wedge \omega, \end{align*}$$

as claimed. $\square$

We conclude this section by giving a formula for $d^\nabla \circ d^\nabla(t)$, for any $t \in \mathcal{A}^i(\xi)$. Consider the special case of the bilinear map

$$\wedge: \mathcal{A}^i(\xi) \times \mathcal{A}^j(\eta) \to \mathcal{A}^{i+j}(\xi \otimes \eta)$$

defined just before Proposition 29.6 with $j = 2$ and $\eta = \text{Hom}(\xi, \xi)$. This is the $C^\infty$-bilinear map

$$\wedge: \mathcal{A}^i(\xi) \times \mathcal{A}^2(\text{Hom}(\xi, \xi)) \to \mathcal{A}^{i+2}(\xi \otimes \text{Hom}(\xi, \xi)).$$

We also have the evaluation map,

$$\text{ev}: \mathcal{A}^i(\xi \otimes \text{Hom}(\xi, \xi)) \cong \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi) \otimes_{C^\infty(B)} \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi))$$

$$\to \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi) = \mathcal{A}^i(\xi),$$

given by

$$\text{ev}(\omega \otimes s \otimes h) = \omega \otimes h(s),$$

with $\omega \in \mathcal{A}^i(B)$, $s \in \Gamma(\xi)$ and $h \in \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \Gamma(\xi))$. Let

$$\wedge: \mathcal{A}^i(\xi) \times \mathcal{A}^2(\text{Hom}(\xi, \xi)) \to \mathcal{A}^{i+2}(\xi)$$
be the composition
\[
\mathcal{A}^i(\xi) \times \mathcal{A}^2(\text{Hom}(\xi, \xi)) \xrightarrow{\wedge} \mathcal{A}^{i+2}(\xi \otimes \text{Hom}(\xi, \xi)) \xrightarrow{\text{ev}} \mathcal{A}^{i+2}(\xi).
\]

More explicitly, the above map is given (on generators) by
\[(\omega \otimes s) \wedge H = \omega \wedge H(s),\]
for any \(\omega \in \mathcal{A}^i(B), s \in \Gamma(\xi)\) and \(H \in \text{Hom}_{C^\infty(B)}(\Gamma(\xi), \mathcal{A}^2(\xi)) \cong \mathcal{A}^2(\text{Hom}(\xi, \xi)).\)

**Proposition 29.11.** For any vector bundle \(\xi\) and any connection \(\nabla\) on \(\xi\), the composition \(d\nabla \circ d\nabla: \mathcal{A}^i(\xi) \to \mathcal{A}^{i+2}(\xi)\) maps \(t\) to \(t \wedge R^\nabla\), for any \(t \in \mathcal{A}^i(\xi)\).

**Proof.** Any \(t \in \mathcal{A}^i(\xi)\) is some linear combination of elements \(\omega \otimes s \in \mathcal{A}^i(B) \otimes_{C^\infty(B)} \Gamma(\xi)\) and by Proposition 29.6, we have
\[
d\nabla \circ d\nabla (\omega \otimes s) = d\nabla (d\omega \otimes s + (-1)^i \omega \wedge \nabla s)
= dd\omega \otimes s + (-1)^{i+1} d\omega \wedge \nabla s + (-1)^i (-1)^i (d\omega \otimes s + (-1)^i \omega \wedge d\nabla \circ \nabla s)
= \omega \wedge d\nabla \circ \nabla s
= (\omega \otimes s) \wedge R^\nabla,
\]
as claimed. \(\square\)

Proposition 29.11 shows that \(d\nabla \circ d\nabla = 0\) iff \(R^\nabla = d\nabla \circ \nabla = 0\), that is, iff the connection \(\nabla\) is flat. Thus, the sequence
\[
0 \to \mathcal{A}^0(\xi) \xrightarrow{\nabla} \mathcal{A}^1(\xi) \xrightarrow{d\nabla} \mathcal{A}^2(\xi) \to \cdots \to \mathcal{A}^i(\xi) \xrightarrow{d\nabla} \mathcal{A}^{i+1}(\xi) \to \cdots,
\]
is a cochain complex iff \(\nabla\) is flat.

Again, everything we did in this section applies to complex vector bundles.

### 29.3 Parallel Transport

The notion of connection yields the notion of parallel transport in a vector bundle. First, we need to define the covariant derivative of a section along a curve.

**Definition 29.3.** Let \(\xi = (E, \pi, B, V)\) be a vector bundle and let \(\gamma: [a, b] \to B\) be a smooth curve in \(B\). A smooth section along the curve \(\gamma\) is a smooth map \(X: [a, b] \to E\), such that \(\pi(X(t)) = \gamma(t)\), for all \(t \in [a, b]\). When \(\xi = TB\), the tangent bundle of the manifold \(B\), we use the terminology smooth vector field along \(\gamma\).

Recall that the curve \(\gamma: [a, b] \to B\) is smooth iff \(\gamma\) is the restriction to \([a, b]\) of a smooth curve on some open interval containing \([a, b]\). Since a section \(X\) along a curve \(\gamma\) does not necessarily extend to an open subset of \(B\) (for example, if the image of \(\gamma\) is dense in \(B\)), the covariant derivative \((\nabla_{\gamma(t_0)} X)_{\gamma(t_0)}\) may not be defined, so we need a proposition showing that the covariant derivative of a section along a curve makes sense.
Proposition 29.12. Let $\xi$ be a vector bundle, $\nabla$ be a connection on $\xi$, and $\gamma: [a, b] \to B$ be a smooth curve in $B$. There is a $\mathbb{R}$-linear map $D/dt$, defined on the vector space of smooth sections $X$ along $\gamma$, which satisfies the following conditions:

1. For any smooth function $f: [a, b] \to \mathbb{R}$,
   
   $\frac{D(fX)}{dt} = \frac{df}{dt} X + f \frac{DX}{dt}$

2. If $X$ is induced by a global section $s \in \Gamma(\xi)$, that is, if $X(t_0) = s(\gamma(t_0))$ for all $t_0 \in [a, b]$, then
   
   $\frac{DX}{dt}(t_0) = (\nabla_{\gamma'(t_0)} s)_{\gamma(t_0)}$.

Proof. Since $\gamma([a, b])$ is compact, it can be covered by a finite number of open subsets $U_\alpha$ such that $(U_\alpha, \varphi_\alpha)$ is a local trivialization. Thus, we may assume that $\gamma: [a, b] \to U$ for some local trivialization, $(U, \varphi)$. As $\varphi \circ \gamma: [a, b] \to \mathbb{R}^n$, we can write

   $\varphi \circ \gamma(t) = (u_1(t), \ldots, u_n(t))$,

where each $u_i = pr_i \circ \varphi \circ \gamma$ is smooth. Now (see Definition 7.14), for every $g \in C^\infty(B)$, as

   $d\gamma_{t_0} \left( \frac{d}{dt} \right)_{t_0} (g) = \frac{d}{dt} (g \circ \gamma)_{t_0} = \frac{d}{dt} ((g \circ \varphi^{-1} \circ (\varphi \circ \gamma))_{t_0} = \sum_{i=1}^n \frac{du_i}{dt} \left( \frac{\partial}{\partial x_i} \right)_{\gamma(t_0)} g$,

since by definition of $\gamma'(t_0)$,

   $\gamma'(t_0) = d\gamma_{t_0} \left( \frac{d}{dt} \right)_{t_0}$

(see the end of Section 7.2), we have

   $\gamma'(t_0) = \sum_{i=1}^n \frac{du_i}{dt} \left( \frac{\partial}{\partial x_i} \right)_{\gamma(t_0)}$.

If $(s_1, \ldots, s_n)$ is a frame over $U$, we can write

   $X(t) = \sum_{i=1}^n X_i(t)s_i(\gamma(t))$,

for some smooth functions, $X_i$. Then, conditions (1) and (2) imply that

   $\frac{DX}{dt} = \sum_{j=1}^n \left( \frac{dX_j}{dt} s_j(\gamma(t)) + X_j(t)\nabla_{\gamma'(t)} (s_j(\gamma(t))) \right)$.
and since
\[\gamma'(t) = \sum_{i=1}^{n} \frac{du_i}{dt} \left( \frac{\partial}{\partial x_i} \right)_{\gamma(t)},\]
there exist some smooth functions, \(\Gamma^k_{ij}\), so that
\[\nabla_{\gamma'(t)}(s_j(\gamma(t))) = \sum_{i=1}^{n} \frac{du_i}{dt} \frac{\partial}{\partial x_i} (s_j(\gamma(t))) = \sum_{i,k} \frac{du_i}{dt} \Gamma^k_{ij} s_k(\gamma(t)).\]

It follows that
\[\frac{DX}{dt} = \sum_{k=1}^{n} \left( \frac{dX_k}{dt} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} X_j \right) s_k(\gamma(t)).\]

Conversely, the above expression defines a linear operator, \(D/dt\), and it is easy to check that it satisfies (1) and (2).

The operator \(D/dt\) is often called covariant derivative along \(\gamma\) and it is also denoted by \(\nabla_{\gamma'(t)}\) or simply \(\nabla_{\gamma'}\).

**Definition 29.4.** Let \(\xi\) be a vector bundle and let \(\nabla\) be a connection on \(\xi\). For every curve \(\gamma: [a, b] \to B\) in \(B\), a section \(X\) along \(\gamma\) is parallel (along \(\gamma\)) iff
\[\frac{DX}{dt}(t_0) = 0 \quad \text{for all } t_0 \in [a, b].\]

If \(\xi\) was the tangent bundle of a smooth manifold \(M\) embedded in \(\mathbb{R}^d\) (for some \(d\)), then to say that \(X\) is parallel along \(\gamma\) would mean that the directional derivative, \((D_\gamma X)(\gamma(t))\), is normal to \(T_{\gamma(t)}M\).

The following proposition can be shown using the existence and uniqueness of solutions of ODE’s (in our case, linear ODE’s) and its proof is omitted:

**Proposition 29.13.** Let \(\xi\) be a vector bundle and let \(\nabla\) be a connection on \(\xi\). For every \(C^1\) curve \(\gamma: [a, b] \to B\) in \(B\), for every \(t \in [a, b]\) and every \(v \in \pi^{-1}(\gamma(t))\), there is a unique parallel section \(X\) along \(\gamma\) such that \(X(t) = v\).

**Proof.** For the proof of Proposition 29.13 it is sufficient to consider the portions of the curve \(\gamma\) contained in some local trivialization. In such a trivialization, \((U, \varphi)\), as in the proof of Proposition 29.12, using a local frame, \((s_1, \ldots, s_n)\), over \(U\), we have
\[\frac{DX}{dt} = \sum_{k=1}^{n} \left( \frac{dX_k}{dt} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} X_j \right) s_k(\gamma(t)),\]
29.4. CONNECTIONS COMPATIBLE WITH A METRIC

with \( u_i = pr_i \circ \varphi \circ \gamma \). Consequently, \( X \) is parallel along our portion of \( \gamma \) iff the system of linear ODE’s in the unknowns, \( X_k \),

\[
\frac{dX_k}{dt} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} X_j = 0, \quad k = 1, \ldots, n,
\]
is satisfied.

Remark: Proposition 29.13 can be extended to piecewise \( C^1 \) curves.

Definition 29.5. Let \( \xi \) be a vector bundle and let \( \nabla \) be a connection on \( \xi \). For every curve \( \gamma: [a, b] \to B \) in \( B \), for every \( t \in [a, b] \), the parallel transport from \( \gamma(a) \) to \( \gamma(t) \) along \( \gamma \) is the linear map from the fibre \( \pi^{-1}(\gamma(a)) \) to the fibre \( \pi^{-1}(\gamma(t)) \), which associates to any \( v \in \pi^{-1}(\gamma(a)) \) the vector \( X_v(t) \in \pi^{-1}(\gamma(t)) \), where \( X_v \) is the unique parallel section along \( \gamma \) with \( X_v(a) = v \).

The following proposition is an immediate consequence of properties of linear ODE’s:

Proposition 29.14. Let \( \xi = (E, \pi, B, V) \) be a vector bundle and let \( \nabla \) be a connection on \( \xi \). For every \( C^1 \) curve \( \gamma: [a, b] \to B \) in \( B \), the parallel transport along \( \gamma \) defines for every \( t \in [a, b] \) a linear isomorphism \( P_\gamma: \pi^{-1}(\gamma(a)) \to \pi^{-1}(\gamma(t)) \) between the fibres \( \pi^{-1}(\gamma(a)) \) and \( \pi^{-1}(\gamma(t)) \).

In particular, if \( \gamma \) is a closed curve, that is, if \( \gamma(a) = \gamma(b) = p \), we obtain a linear isomorphism \( P_\gamma \) of the fibre \( E_p = \pi^{-1}(p) \), called the holonomy of \( \gamma \). The holonomy group of \( \nabla \) based at \( p \), denoted \( \text{Hol}_p(\nabla) \), is the subgroup of \( \text{GL}(V, \mathbb{R}) \) (where \( V \) is the fibre of the vector bundle \( \xi \)) given by

\[
\text{Hol}_p(\nabla) = \{ P_\gamma \in \text{GL}(V, \mathbb{R}) \mid \gamma \text{ is a closed curve based at } p \}.
\]

If \( B \) is connected, then \( \text{Hol}_p(\nabla) \) depends on the basepoint \( p \in B \) up to conjugation and so \( \text{Hol}_p(\nabla) \) and \( \text{Hol}_q(\nabla) \) are isomorphic for all \( p, q \in B \). In this case, it makes sense to talk about the holonomy group of \( \nabla \). If \( \xi = TB \), the tangent bundle of a manifold, \( B \), by abuse of language, we call \( \text{Hol}_p(\nabla) \) the holonomy group of \( B \).

29.4 Connections Compatible with a Metric; Levi-Civita Connections

If a vector bundle (or a Riemannian manifold) \( \xi \) has a metric, then it is natural to define when a connection \( \nabla \) on \( \xi \) is compatible with the metric. So, assume the vector bundle \( \xi \) has a metric \( \langle -, - \rangle \). We can use this metric to define pairings

\[
\mathcal{A}^1(\xi) \times \mathcal{A}^0(\xi) \to \mathcal{A}^1(B) \quad \text{and} \quad \mathcal{A}^0(\xi) \times \mathcal{A}^1(\xi) \to \mathcal{A}^1(B)
\]
as follows: Set (on generators)
\[ \langle \omega \otimes s_1, s_2 \rangle = \langle s_1, \omega \otimes s_2 \rangle = \omega \langle s_1, s_2 \rangle, \]
for all \( \omega \in A^1(B), s_1, s_2 \in \Gamma(\xi) \) and where \( \langle s_1, s_2 \rangle \) is the function in \( C^\infty(B) \) given by \( b \mapsto \langle s_1(b), s_2(b) \rangle \), for all \( b \in B \). More generally, we define a pairing
\[ \mathcal{A}^i(\xi) \times \mathcal{A}^j(\xi) \to \mathcal{A}^{i+j}(B), \]
by
\[ \langle \omega \otimes s_1, \eta \otimes s_2 \rangle = \langle s_1, s_2 \rangle \omega \wedge \eta, \]
for all \( \omega \in \mathcal{A}^i(B), \eta \in \mathcal{A}^j(B), s_1, s_2 \in \Gamma(\xi) \).

**Definition 29.6.** Given any metric \( \langle -, - \rangle \) on a vector bundle \( \xi \), a connection \( \nabla \) on \( \xi \) is *compatible with the metric*, for short, a *metric connection* iff
\[ d\langle s_1, s_2 \rangle = \langle \nabla s_1, s_2 \rangle + \langle s_1, \nabla s_2 \rangle, \]
for all \( s_1, s_2 \in \Gamma(\xi) \).

In terms of version-two of a connection, \( \nabla_X \) is a metric connection iff
\[ X(\langle s_1, s_2 \rangle) = \langle \nabla_X s_1, s_2 \rangle + \langle s_1, \nabla_X s_2 \rangle, \]
for every vector field, \( X \in \mathfrak{X}(B) \).

Definition 29.6 remains unchanged if \( \xi \) is a complex vector bundle. The condition of compatibility with a metric is nicely expressed in a local trivialization. Indeed, let \( (U, \varphi) \) be a local trivialization of the vector bundle \( \xi \) (of rank \( n \)). Then, using the Gram-Schmidt procedure, we obtain an orthonormal frame \( (s_1, \ldots, s_n) \), over \( U \).

**Proposition 29.15.** Using the above notations, if \( \omega = (\omega_{ij}) \) is the connection matrix of \( \nabla \) w.r.t. \( (s_1, \ldots, s_n) \), then \( \omega \) is skew-symmetric.

**Proof.** Since
\[ \nabla e_i = \sum_{j=1}^n \omega_{ij} \otimes s_j \]
and since \( \langle s_i, s_j \rangle = \delta_{ij} \) (as \( (s_1, \ldots, s_n) \) is orthonormal), we have \( d\langle s_i, s_j \rangle = 0 \) on \( U \). Consequently,
\[ 0 = d\langle s_i, s_j \rangle = \langle \nabla s_i, s_j \rangle + \langle s_i, \nabla s_j \rangle = \sum_{k=1}^n \omega_{ik} \otimes s_k, s_j \rangle + \langle s_i, \sum_{l=1}^n \omega_{jl} \otimes s_l \rangle = \sum_{k=1}^n \omega_{ik} \langle s_k, s_j \rangle + \sum_{l=1}^n \omega_{jl} \langle s_i, s_l \rangle = \omega_{ij} + \omega_{ji}, \]
as claimed. \( \square \)
In Proposition 29.15, if \( \xi \) is a complex vector bundle, then \( \omega \) is skew-Hermitian. This means that
\[
\overline{\omega}^\top = -\omega,
\]
where \( \overline{\omega} \) is the conjugate matrix of \( \omega \); that is, \( (\overline{\omega})_{ij} = \overline{\omega_{ij}} \). It is also easy to prove that metric connections exist.

**Proposition 29.16.** Let \( \xi \) be a rank \( n \) vector with a metric \( \langle - , - \rangle \). Then, \( \xi \) possesses metric connections.

**Proof.** We can pick a locally finite cover \( (U_\alpha)_\alpha \) of \( B \) such that \( (U_\alpha, \varphi_\alpha) \) is a local trivialization of \( \xi \). Then, for each \( (U_\alpha, \varphi_\alpha) \), we use the Gram-Schmidt procedure to obtain an orthonormal frame \( (s_\alpha^1, \ldots, s_\alpha^n) \) over \( U_\alpha \), and we let \( \nabla^\alpha \) be the trivial connection on \( \pi^{-1}(U_\alpha) \). By construction, \( \nabla^\alpha \) is compatible with the metric. We finish the argument by using a partition of unity, leaving the details to the reader. \( \square \)

If \( \xi \) is a complex vector bundle, then we use a Hermitian metric and we call a connection compatible with this metric a *Hermitian connection*. In any local trivialization, the connection matrix \( \omega \) is skew-Hermitian. The existence of Hermitian connections is clear.

If \( \nabla \) is a metric connection, then the curvature matrices are also skew-symmetric.

**Proposition 29.17.** Let \( \xi \) be a rank \( n \) vector bundle with a metric \( \langle - , - \rangle \). In any local trivialization of \( \xi \), the curvature matrix \( \Omega = (\Omega_{ij}) \) is skew-symmetric. If \( \xi \) is a complex vector bundle, then \( \Omega = (\Omega_{ij}) \) is skew-Hermitian.

**Proof.** By the structure equation (Proposition 29.8),
\[
\Omega = d\omega - \omega \wedge \omega,
\]
that is, \( \Omega_{ij} = d\omega_{ij} - \sum_{k=1}^n \omega_{ik} \wedge \omega_{kj} \). Using the skew symmetry of \( \omega_{ij} \) and wedge,
\[
\Omega_{ji} = d\omega_{ji} - \sum_{k=1}^n \omega_{jk} \wedge \omega_{ki} \\
= -d\omega_{ij} - \sum_{k=1}^n \omega_{kj} \wedge \omega_{ik} \\
= -d\omega_{ij} + \sum_{k=1}^n \omega_{ik} \wedge \omega_{kj} \\
= -\Omega_{ij},
\]
as claimed. \( \square \)
We now restrict our attention to a Riemannian manifold; that is, to the case where our bundle \( \xi \) is the tangent bundle \( \xi = TM \) of some Riemannian manifold \( M \). We know from Proposition 29.16 that metric connections on \( TM \) exist. However, there are many metric connections on \( TM \), and none of them seems more relevant than the others. If \( M \) is a Riemannian manifold, the metric \( \langle -, - \rangle \) on \( M \) is often denoted \( g \). In this case, for every chart \( (U, \varphi) \), we let \( g_{ij} \in C^\infty(M) \) be the function defined by

\[
g_{ij}(p) = \left( \frac{\partial}{\partial x_i} \right)_p \left( \frac{\partial}{\partial x_j} \right)_p.
\]

(Note the unfortunate clash of notation with the transitions functions!)

The notations \( g = \sum_{ij} g_{ij} dx_i \otimes dx_j \) or simply \( g = \sum_{ij} g_{ij} dx_i dx_j \) are often used to denote the metric in local coordinates.

We observed immediately after stating Proposition 29.5 that the covariant differential \( \nabla g \) of the Riemannian metric \( g \) on \( M \) is given by

\[
\nabla_X(g)(Y, Z) = d(g(Y, Z))(X) - g(\nabla_X Y, Z) - g(Y, \nabla_X Z),
\]

for all \( X, Y, Z \in \mathfrak{X}(M) \). Therefore, a connection \( \nabla \) on a Riemannian manifold \( (M, g) \) is compatible with the metric iff

\[
\nabla g = 0.
\]

It is remarkable that if we require a certain kind of symmetry on a metric connection, then it is uniquely determined. Such a connection is known as the Levi–Civita connection. The Levi–Civita connection can be characterized in several equivalent ways, a rather simple way involving the notion of torsion of a connection.

Recall that one way to introduce the curvature is to view it as the “error term”

\[
R(X, Y) = \nabla_X \nabla_Y - \nabla_Y \nabla_X - [X, Y].
\]

Another natural error term is the torsion \( T(X, Y) \), of the connection \( \nabla \), given by

\[
T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y],
\]

which measures the failure of the connection to behave like the Lie bracket. Then, the Levi–Civita connection is the unique metric and torsion-free connection \( (T(X, Y) = 0) \) on the Riemannian manifold.

Another way to characterize the Levi-Civita connection uses the cotangent bundle \( T^*M \). It turns out that a connection \( \nabla \) on a vector bundle (metric or not) \( \xi \) naturally induces a connection \( \nabla^* \) on the dual bundle \( \xi^* \). Now, if \( \nabla \) is a connection on \( TM \), then \( \nabla^* \) is a connection on \( T^*M \), namely, a linear map, \( \nabla^*: \Gamma(T^*M) \to \mathcal{A}^1(M) \otimes_{C^\infty(B)} \Gamma(T^*M) \); that is

\[
\nabla^*: \mathcal{A}^1(M) \to \mathcal{A}^1(M) \otimes_{C^\infty(B)} \mathcal{A}^1(M) \cong \Gamma(T^*M \otimes T^*M),
\]
since $\Gamma(T^*M) = \mathcal{A}^1(M)$. If we compose this map with $\wedge$, we get the map

$$
\mathcal{A}^1(M) \xrightarrow{\nabla} \mathcal{A}^1(M) \otimes_{C^\infty(B)} \mathcal{A}^1(M) \xrightarrow{\wedge} \mathcal{A}^2(M).
$$

Then miracle, a metric connection is the Levi-Civita connection iff

$$
d = \wedge \circ \nabla^*,
$$

where $d: \mathcal{A}^1(M) \to \mathcal{A}^2(M)$ is exterior differentiation. There is also a nice local expression of the above equation.

The first characterization of the Levi–Civita connection was given in Proposition 11.8.

Let us now consider the second approach to torsion-freeness. For this, we have to explain how a connection $\nabla$ on a vector bundle $\xi = (E, \pi, B, V)$ induces a connection $\nabla^*$ on the dual bundle $\xi^*$. First, there is an evaluation map $\Gamma(\xi \otimes \xi^*) \to \Gamma(\epsilon_1)$ (where $\epsilon_1 = B \times \mathbb{R}$, the trivial line bundle over $B$), or equivalently

$$
\langle -, - \rangle: \Gamma(\xi) \otimes_{C^\infty(B)} \text{Hom}_{C^\infty(B)}(\Gamma(\xi), C^\infty(B)) \to C^\infty(B),
$$

given by

$$
\langle s_1, s_2^* \rangle = s_2^*(s_1), \quad s_1 \in \Gamma(\xi), \ s_2^* \in \text{Hom}_{C^\infty(B)}(\Gamma(\xi), C^\infty(B)),
$$

and thus a map

$$
\mathcal{A}^k(\xi \otimes \xi^*) = \mathcal{A}^k(B) \otimes_{C^\infty(B)} \Gamma(\xi \otimes \xi^*) \xrightarrow{\text{id} \otimes \langle -, - \rangle} \mathcal{A}^k(B) \otimes_{C^\infty(B)} C^\infty(B) \cong \mathcal{A}^k(B).
$$

Using this map, we obtain a pairing

$$
\langle -, - \rangle: \mathcal{A}^i(\xi) \otimes \mathcal{A}^j(\xi^*) \xrightarrow{\wedge} \mathcal{A}^{i+j}(\xi \otimes \xi^*) \to \mathcal{A}^{i+j}(B)
$$

given by

$$
(\omega \otimes s_1, \eta \otimes s_2^*) = (\omega \wedge \eta) \otimes \langle s_1, s_2^* \rangle,
$$

where $\omega \in \mathcal{A}^i(B), \eta \in \mathcal{A}^j(B), s_1 \in \Gamma(\xi), s_2^* \in \Gamma(\xi^*)$. It is easy to check that this pairing is non-degenerate. Then, given a connection $\nabla$ on a rank $n$ vector bundle $\xi$, we define $\nabla^*$ on $\xi^*$ by

$$
d\langle s_1, s_2^* \rangle = (\nabla(s_1), s_2^*) + (s_1, \nabla^*(s_2^*))
$$

where $s_1 \in \Gamma(\xi)$ and $s_2^* \in \Gamma(\xi^*)$. Because the pairing $\langle -, - \rangle$ is non-degenerate, $\nabla^*$ is well-defined, and it is immediately that it is a connection on $\xi^*$. Let us see how it is expressed locally.

If $(U, \varphi)$ is a local trivialization and $(s_1, \ldots, s_n)$ is the frame over $U$ associated with $(U, \varphi)$, then let $(\theta_1, \ldots, \theta_n)$ be the dual frame (called a coframe). We have

$$
\langle s_j, \theta_i \rangle = \theta_i(s_j) = \delta_{ij}, \quad 1 \leq i, j \leq n.
$$
Recall that
\[ \nabla s_j = \sum_{k=1}^{n} \omega_{jk} \otimes s_k, \]
and write
\[ \nabla^* \theta_i = \sum_{k=1}^{n} \omega_{ik}^* \otimes \theta_k. \]

Applying \( d \) to the equation \( \langle s_j, \theta_i \rangle = \delta_{ij} \) and using the equation defining \( \nabla^* \), we get
\[
0 = \quad d \langle s_j, \theta_i \rangle \\
= \quad (\nabla (s_j), \theta_i) + (s_j, \nabla^* (\theta_i)) \\
= \quad \left( \sum_{k=1}^{n} \omega_{jk} \otimes s_k, \theta_i \right) + \left( s_j, \sum_{l=1}^{n} \omega_{il}^* \otimes \theta_l \right) \\
= \quad \sum_{k=1}^{n} \omega_{jk} \langle s_k, \theta_i \rangle + \sum_{l=1}^{n} \omega_{il}^* \langle s_j, \theta_l \rangle \\
= \quad \omega_{ji} + \omega_{ij}^*. 
\]

Therefore, if we write \( \omega^* = (\omega_{ij}^*) \), we have
\[ \omega^* = -\omega^T. \]

If \( \nabla \) is a metric connection, then \( \omega \) is skew-symmetric; that is, \( \omega^T = -\omega \). In this case, \( \omega^* = -\omega^T = \omega \).

If \( \xi \) is a complex vector bundle, then there is a problem because if \((s_1, \ldots, s_n)\) is a frame over \( U \), then the \( \theta_j(b) \)'s defined by
\[ \langle s_i(b), \theta_j(b) \rangle = \delta_{ij} \]
are not linear, but instead conjugate-linear. (Recall that a linear form \( \theta \) is conjugate linear (or semi-linear) iff \( \theta(\lambda u) = \overline{\lambda} \theta(u) \), for all \( \lambda \in \mathbb{C} \).)

Instead of \( \xi^* \), we need to consider the bundle \( \overline{\xi}^* \), which is the bundle whose fibre over \( b \in B \) consist of all conjugate-linear forms over \( \pi^{-1}(b) \). In this case, the evaluation pairing \( \langle s, \theta \rangle \) is conjugate-linear in \( s \), and we find that \( \omega^* = -\overline{\omega}^T \), where \( \omega^* \) is the connection matrix of \( \overline{\xi} \) over \( U \).

If \( \xi \) is a Hermitian bundle, as \( \omega \) is skew-Hermitian, we find that \( \omega^* = \omega \), which makes sense since \( \xi \) and \( \overline{\xi}^* \) are canonically isomorphic. However, this does not give any information on \( \xi^* \). For this, we consider the conjugate bundle \( \overline{\xi} \). This is the bundle obtained from \( \xi \) by redefining the vector space structure on each fibre \( \pi^{-1}(b) \), with \( b \in B \), so that
\[ (x + iy)v = (x - iy)v, \]
for every \( v \in \pi^{-1}(b) \). If \( \omega \) is the connection matrix of \( \xi \) over \( U \), then \( \overline{\omega} \) is the connection matrix of \( \overline{\xi} \) over \( U \). If \( \xi \) has a Hermitian metric, it is easy to prove that \( \xi^* \) and \( \overline{\xi} \) are canonically isomorphic (see Proposition 29.27). In fact, the Hermitian product \( \langle - , - \rangle \) establishes a pairing between \( \overline{\xi} \) and \( \xi^* \), and basically as above, we can show that if \( \omega \) is the connection matrix of \( \xi \) over \( U \), then \( \omega^* = -\omega^\top \) is the connection matrix of \( \xi^* \) over \( U \). As \( \omega \) is skew-Hermitian, \( \omega^* = \omega \).

Going back to a connection \( \nabla \) on a manifold \( M \), the connection \( \nabla^* \) is a linear map
\[
\nabla^*_\omega(X,Y) = X(\theta(Y)) - \theta(\nabla_X Y),
\]
for every one-form \( \theta \in \mathcal{A}^1(M) \) and all vector fields \( X, Y \in \mathfrak{X}(M) \). Applying \( \wedge \), we get
\[
\nabla^*_\theta(X,Y) - \nabla^*_\theta(Y,X) = X(\theta(Y)) - \theta(\nabla_X Y) - Y(\theta(X)) + \theta(\nabla_Y X) = X(\theta(Y)) - Y(\theta(X)) - \theta(\nabla_X Y - \nabla_Y X).
\]

However, recall that
\[
d\theta(X,Y) = X(\theta(Y)) - Y(\theta(X)) - \theta([X,Y]),
\]
so we get
\[
(\wedge \circ \nabla^*)(\theta)(X,Y) = \nabla^*_\theta(X,Y) - \nabla^*_\theta(Y,X) = d\theta(X,Y) - \theta(\nabla_X Y - \nabla_Y X - [X,Y]).
\]

It follows that for every \( \theta \in \mathcal{A}^1(M) \), we have \( (\wedge \circ \nabla^*)\theta = d\theta \) iff \( \theta(T(X,Y)) = 0 \) for all \( X, Y \in \mathfrak{X}(M) \), that is iff \( T(X,Y) = 0 \), for all \( X, Y \in \mathfrak{X}(M) \). We record this as

**Proposition 29.18.** Let \( M \) be a manifold with connection \( \nabla \). Then, \( \nabla \) is torsion-free (i.e., \( T(X,Y) = \nabla_X Y - \nabla_Y X - [X,Y] = 0 \), for all \( X, Y \in \mathfrak{X}(M) \)) iff
\[
\wedge \circ \nabla^* = d,
\]
where \( d: \mathcal{A}^1(M) \to \mathcal{A}^2(M) \) is exterior differentiation.

Proposition 29.18 together with Proposition 11.8 yield a second version of the Levi-Civita Theorem:

**Proposition 29.19.** (Levi-Civita, Version 2) Let \( M \) be any Riemannian manifold. There is a unique, metric connection \( \nabla \) on \( M \), such that
\[
\wedge \circ \nabla^* = d,
\]
where \( d: \mathcal{A}^1(M) \to \mathcal{A}^2(M) \) is exterior differentiation. This connection is equal to the Levi-Civita connection in Proposition 11.8.
Remark: If $\nabla$ is the Levi-Civita connection of some Riemannian manifold $M$, for every chart $(U, \varphi)$, we have $\omega^* = \omega$, where $\omega$ is the connection matrix of $\nabla$ over $U$ and $\omega^*$ is the connection matrix of the dual connection $\nabla^*$. This implies that the Christoffel symbols of $\nabla$ and $\nabla^*$ over $U$ are identical. Furthermore, $\nabla^*$ is a linear map

$$\nabla^*: \mathcal{A}^1(M) \longrightarrow \Gamma(T^*M \otimes T^*M).$$

Thus, locally in a chart $(U, \varphi)$, if (as usual) we let $x_i = pr_i \circ \varphi$, then we can write

$$\nabla^*(dx_k) = \sum_{ij} \Gamma^k_{ij} dx_i \otimes dx_j.$$

Now, if we want $\wedge \circ \nabla^* = d$, we must have $\wedge \nabla^*(dx_k) = ddx_k = 0$, that is

$$\Gamma^k_{ij} = \Gamma^k_{ji},$$

for all $i, j$. Therefore, torsion-freeness can indeed be viewed as a symmetry condition on the Christoffel symbols of the connection $\nabla$.

Our third version is a local version due to Élie Cartan. Recall that locally in a chart $(U, \varphi)$, the connection $\nabla^*$ is given by the matrix, $\omega^*$, such that $\omega^* = -\omega^\top$, where $\omega$ is the connection matrix of $TM$ over $U$. That is, we have

$$\nabla^* \theta_i = \sum_{j=1}^n -\omega_{ji} \otimes \theta_j,$$

for some one-forms $\omega_{ij} \in \mathcal{A}^1(M)$. Then,

$$\wedge \circ \nabla^* \theta_i = -\sum_{j=1}^n \omega_{ji} \wedge \theta_j$$

so the requirement that $d = \wedge \circ \nabla^*$ is expressed locally by

$$d\theta_i = -\sum_{j=1}^n \omega_{ji} \wedge \theta_j.$$

In addition, since our connection is metric, $\omega$ is skew-symmetric, and so $\omega^* = \omega$. Then, it is not too surprising that the following proposition holds:

**Proposition 29.20.** Let $M$ be a Riemannian manifold with metric $\langle -, - \rangle$. For every chart $(U, \varphi)$, if $(s_1, \ldots, s_n)$ is the frame over $U$ associated with $(U, \varphi)$ and $(\theta_1, \ldots, \theta_n)$ is the corresponding coframe (dual frame), then there is a unique matrix $\omega = (\omega_{ij})$ of one-forms $\omega_{ij} \in \mathcal{A}^1(M)$, so that the following conditions hold:

(i) $\omega_{ji} = -\omega_{ij}$. 

(ii) \( d \theta_i = \sum_{j=1}^{n} \omega_{ij} \wedge \theta_j \), or in matrix form, \( d \theta = \omega \wedge \theta \).

**Proof.** There is a direct proof using a combinatorial trick. For instance, see Morita [133], Chapter 5, Proposition 5.32, or Milnor and Stasheff [129], Appendix C, Lemma 8. On the other hand, if we view \( \omega = (\omega_{ij}) \) as a connection matrix, then we observed that (i) asserts that the connection is metric and (ii) that it is torsion-free. We conclude by applying Proposition 29.19.

As an example, consider an orientable (compact) surface \( M \), with a Riemannian metric. Pick any chart \((U, \varphi)\), and choose an orthonormal coframe of one-forms \((\theta_1, \theta_2)\), such that \( \text{Vol} = \theta_1 \wedge \theta_2 \) on \( U \). Then, we have

\[
\begin{align*}
  d \theta_1 &= a_1 \theta_1 \wedge \theta_2 \\
  d \theta_2 &= a_2 \theta_1 \wedge \theta_2
\end{align*}
\]

for some functions, \( a_1, a_2 \), and we let

\[
\omega_{12} = a_1 \theta_1 + a_2 \theta_2.
\]

Clearly,

\[
\begin{pmatrix}
  0 & \omega_{12} \\
  -\omega_{12} & 0
\end{pmatrix}
\begin{pmatrix}
  \theta_1 \\
  \theta_2
\end{pmatrix}
= \begin{pmatrix}
  0 & a_1 \theta_1 + a_2 \theta_2 \\
  -(a_1 \theta_1 + a_2 \theta_2) & 0
\end{pmatrix}
\begin{pmatrix}
  \theta_1 \\
  \theta_2
\end{pmatrix}
= \begin{pmatrix}
  d \theta_1 \\
  d \theta_2
\end{pmatrix}
\]

which shows that

\[
\omega = \omega^* = \begin{pmatrix}
  0 & \omega_{12} \\
  -\omega_{12} & 0
\end{pmatrix}
\]

corresponds to the Levi-Civita connection on \( M \). Since \( \Omega = d \omega - \omega \wedge \omega \), we see that

\[
\Omega = \begin{pmatrix}
  0 & d \omega_{12} \\
  -d \omega_{12} & 0
\end{pmatrix}.
\]

As \( M \) is oriented and as \( M \) has a metric, the transition functions are in \( \text{SO}(2) \). We easily check that

\[
\begin{pmatrix}
  \cos t & \sin t \\
  -\sin t & \cos t
\end{pmatrix}
\begin{pmatrix}
  0 & d \omega_{12} \\
  -d \omega_{12} & 0
\end{pmatrix}
\begin{pmatrix}
  \cos t & -\sin t \\
  \sin t & \cos t
\end{pmatrix}
= \begin{pmatrix}
  0 & d \omega_{12} \\
  -d \omega_{12} & 0
\end{pmatrix},
\]

which shows that \( \Omega \) is a global two-form called the Gauss-Bonnet 2-form of \( M \). Then, there is a function \( \kappa \), the Gaussian curvature of \( M \), such that

\[
d \omega_{12} = -\kappa \text{Vol},
\]
where Vol is the oriented volume form on $M$. The Gauss-Bonnet Theorem for orientable surfaces asserts that

$$\int_M d\omega_{12} = 2\pi \chi(M),$$

where $\chi(M)$ is the Euler characteristic of $M$.

**Remark:** The Levi-Civita connection induced by a Riemannian metric $g$ can also be defined in terms of the Lie derivative of the metric $g$. This is the approach followed in Petersen [140] (Chapter 2). If $\theta_X$ is the one-form given by

$$\theta_X = i_X g;$$

that is, $(i_X g)(Y) = g(X,Y)$ for all $X, Y \in \mathfrak{X}(M)$, and if $L_X g$ is the Lie derivative of the symmetric $(0, 2)$ tensor $g$, defined so that

$$(L_X g)(Y, Z) = X(g(Y, Z)) - g(L_X Y, Z) - g(Y, L_X Z)$$

(see Proposition 23.18), then it is proved in Petersen [140] (Chapter 2, Theorem 1) that the Levi-Civita connection is defined implicitly by the formula

$$2g(\nabla_X Y, Z) = (L_Y g)(X, Z) + (d\theta_Y)(X, Z).$$

### 29.5 Duality between Vector Fields and Differential Forms and their Covariant Derivatives

If $(M, \langle - , - \rangle)$ is a Riemannian manifold, then the inner product $\langle - , - \rangle_p$ on $T_pM$, establishes a canonical duality between $T_pM$ and $T^*_pM$, as explained in Section 21.2. Namely, we have the isomorphism $\flat : T_pM \rightarrow T^*_pM$, defined such that for every $u \in T_pM$, the linear form $u^\flat \in T^*_pM$ is given by

$$u^\flat(v) = \langle u, v \rangle_p, \quad v \in T_pM.$$ 

The inverse isomorphism $\sharp : T^*_pM \rightarrow T_pM$ is defined such that for every $\omega \in T^*_pM$, the vector $\omega^\sharp$ is the unique vector in $T_pM$ so that

$$\langle \omega^\sharp, v \rangle_p = \omega(v), \quad v \in T_pM.$$ 

The isomorphisms $\flat$ and $\sharp$ induce isomorphisms between vector fields $X \in \mathfrak{X}(M)$ and one-forms $\omega \in A^1(M)$: A vector field $X \in \mathfrak{X}(M)$ yields the one-form $X^\flat \in A^1(M)$ given by

$$(X^\flat)_p = (X_p)^\flat,$$

and a one-form $\omega \in A^1(M)$ yields the vector field $\omega^\sharp \in \mathfrak{X}(M)$ given by

$$(\omega^\sharp)_p = (\omega_p)^\sharp,$$
so that 

$$\omega_p(v) = \langle (\omega_p)^\sharp, v \rangle_p, \quad v \in T_p M, \ p \in M.$$ 

In particular, for every smooth function $f \in C^\infty(M)$, the vector field corresponding to the one-form $df$ is the gradient $\text{grad} f$, of $f$. The gradient of $f$ is uniquely determined by the condition 

$$\langle (\text{grad} f)_p, v \rangle_p = df_p(v), \quad v \in T_p M, \ p \in M.$$ 

Recall from Proposition 29.5 that the covariant derivative $\nabla_X \omega$ of any one-form $\omega \in \mathcal{A}^1(M)$ is the one-form given by 

$$(\nabla_X \omega)(Y) = X(\omega(Y)) - \omega(\nabla_X Y).$$ 

If $\nabla$ is a metric connection, then the vector field $(\nabla_X \omega)^\sharp$ corresponding to $\nabla_X \omega$ is nicely expressed in terms of $\omega^\sharp$: Indeed, we have 

$$(\nabla_X \omega)^\sharp = \nabla_X \omega^\sharp.$$ 

The proof goes as follows: 

$$(\nabla_X \omega)(Y) = X(\omega(Y)) - \omega(\nabla_X Y)$$

$$= X(\langle \omega^\sharp, Y \rangle) - \langle \omega^\sharp, \nabla_X Y \rangle$$

$$= \langle \nabla_X \omega^\sharp, Y \rangle + \langle \omega^\sharp, \nabla_X Y \rangle - \langle \omega^\sharp, \nabla_X Y \rangle$$

$$= \langle \nabla_X \omega^\sharp, Y \rangle,$$

where we used the fact that the connection is compatible with the metric in the third line and so, 

$$(\nabla_X \omega)^\sharp = \nabla_X \omega^\sharp,$$

as claimed.

### 29.6 Pontrjagin Classes and Chern Classes, a Glimpse

This section can be omitted at first reading. Its purpose is to introduce the reader to Pontrjagin Classes and Chern Classes, which are fundamental invariants of real (resp. complex) vector bundles. We focus on motivations and intuitions and omit most proofs, but we give precise pointers to the literature for proofs.

Given a real (resp. complex) rank $n$ vector bundle $\xi = (E, \pi, B, V)$, we know that locally, $\xi$ “looks like” a trivial bundle $U \times V$, for some open subset $U$ of the base space $B$. Globally, $\xi$ can be very twisted, and one of the main issues is to understand and quantify “how twisted” $\xi$ really is. Now, we know that every vector bundle admit a connection, say $\nabla$, and the curvature $R^\nabla$ of this connection is some measure of the twisting of $\xi$. However, $R^\nabla$ depends on $\nabla$, so curvature is not intrinsic to $\xi$, which is unsatisfactory as we seek invariants that depend only on $\xi$. 

Pontrjagin, Stiefel and Chern (starting from the late 1930’s) discovered that invariants with “good” properties could be defined if we took these invariants to belong to various cohomology groups associated with $B$. Such invariants are usually called characteristic classes. Roughly, there are two main methods for defining characteristic classes: one using topology, and the other due to Chern and Weil, using differential forms.

A masterly exposition of these methods is given in the classic book by Milnor and Stasheff [129]. Amazingly, the method of Chern and Weil using differential forms is quite accessible for someone who has reasonably good knowledge of differential forms and de Rham cohomology, as long as one is willing to gloss over various technical details.

As we said earlier, one of the problems with curvature is that it depends on a connection. The way to circumvent this difficulty rests on the simple, yet subtle observation, that locally, given any two overlapping local trivializations $(U_\alpha, \varphi_\alpha)$ and $(U_\beta, \varphi_\beta)$, the transformation rule for the curvature matrices $\Omega_\alpha$ and $\Omega_\beta$ is

$$\Omega_\beta = g_{\alpha\beta} \Omega_\alpha g_{\alpha\beta}^{-1},$$

where $g_{\alpha\beta}: U_\alpha \cap U_\beta \to \text{GL}(V)$ is the transition function. The matrices of two-forms $\Omega_\alpha$ are local, but the stroke of genius is to glue them together to form a global form using invariant polynomials.

Indeed, the $\Omega_\alpha$ are $n \times n$ matrices, so consider the algebra of polynomials $\mathbb{R}[X_1, \ldots, X_{n^2}]$ (or $\mathbb{C}[X_1, \ldots, X_{n^2}]$ in the complex case) in $n^2$ variables $X_1, \ldots, X_{n^2}$, considered as the entries of an $n \times n$ matrix. It is more convenient to use the set of variables $\{X_{ij} \mid 1 \leq i, j \leq n\}$, and to let $X$ be the $n \times n$ matrix $X = (X_{ij})$.

**Definition 29.7.** A polynomial $P \in \mathbb{R}[\{X_{ij} \mid 1 \leq i, j \leq n\}]$ (or $P \in \mathbb{C}[\{X_{ij} \mid 1 \leq i, j \leq n\}]$) is invariant iff

$$P(A X A^{-1}) = P(X),$$

for all $A \in \text{GL}(n, \mathbb{R})$ (resp. $A \in \text{GL}(n, \mathbb{C})$). The algebra of invariant polynomials over $n \times n$ matrices is denoted by $I_n$.

Examples of invariant polynomials are, the trace $\text{tr}(X)$, and the determinant $\text{det}(X)$, of the matrix $X$. We will characterize shortly the algebra $I_n$.

Now comes the punch line: For any homogeneous invariant polynomial $P \in I_n$ of degree $k$, we can substitute $\Omega_\alpha$ for $X$; that is, substitute $\omega_{ij}$ for $X_{ij}$, and evaluate $P(\Omega_\alpha)$. This is because $\Omega$ is a matrix of two-forms, and the wedge product is commutative for forms of even degree. Therefore, $P(\Omega_\alpha) \in \mathcal{A}^{2k}(U_\alpha)$. But, the formula for a change of trivialization yields

$$P(\Omega_\alpha) = P(g_{\alpha\beta} \Omega_\alpha g_{\alpha\beta}^{-1}) = P(\Omega_\beta),$$

so the forms $P(\Omega_\alpha)$ and $P(\Omega_\beta)$ agree on overlaps, and thus they define a global form denoted $P(R^\nabla) \in \mathcal{A}^{2k}(B)$. 

Now, we know how to obtain global $2k$-forms $P(R^\nabla) \in A^{2k}(B)$, but they still seem to depend on the connection, and how do they define a cohomology class? Both problems are settled thanks to the following Theorems:

**Theorem 29.21.** For every real rank $n$ vector bundle $\xi$, for every connection $\nabla$ on $\xi$, for every invariant homogeneous polynomial $P$ of degree $k$, the $2k$-form $P(R^\nabla) \in A^{2k}(B)$ is closed. If $\xi$ is a complex vector bundle, then the $2k$-form $P(R^\nabla) \in A^{2k}(B;\mathbb{C})$ is closed.

Theorem 29.21 implies that the $2k$-form $P(R^\nabla) \in A^{2k}(B)$ defines a cohomology class $[P(R^\nabla)] \in H^{2k}_{DR}(B)$. We will come back to the proof of Theorem 29.21 later.

**Theorem 29.22.** For every real (resp. complex) rank $n$ vector bundle $\xi$, for every invariant homogeneous polynomial $P$ of degree $k$, the cohomology class $[P(R^\nabla)] \in H^{2k}_{DR}(B)$ (resp. $[P(R^\nabla)] \in H^{2k}_{DR}(B;\mathbb{C})$) is independent of the choice of the connection $\nabla$.

The cohomology class $[P(R^\nabla)]$, which does not depend on $\nabla$, is denoted $P(\xi)$, and is called the characteristic class of $\xi$ corresponding to $P$.

The proof of Theorem 29.22 involves a kind of homotopy argument; see Madsen and Tornehave [119] (Lemma 18.2), Morita [133] (Proposition 5.28), or Milnor and Stasheff [129] (Appendix C).

The upshot is that Theorems 29.21 and 29.22 give us a method for producing invariants of a vector bundle that somehow reflect how curved (or twisted) the bundle is. However, it appears that we need to consider infinitely many invariants. Fortunately, we can do better because the algebra $I_n$ of invariant polynomials is finitely generated, and in fact, has very nice sets of generators. For this, we recall the elementary symmetric functions in $n$ variables.

Given $n$ variables $\lambda_1, \ldots, \lambda_n$, we can write

$$\prod_{i=1}^{n}(1 + t\lambda_i) = 1 + \sigma_1 t + \sigma_2 t^2 + \cdots + \sigma_n t^n,$$

where the $\sigma_i$ are symmetric, homogeneous polynomials of degree $i$ in $\lambda_1, \ldots, \lambda_n$, called elementary symmetric functions in $n$ variables. For example,

$$\sigma_1 = \sum_{i=1}^{n} \lambda_i, \quad \sigma_1 = \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j, \quad \sigma_n = \lambda_1 \cdots \lambda_n.$$

To be more precise, we write $\sigma_i(\lambda_1, \ldots, \lambda_n)$ instead of $\sigma_i$.

Given any $n \times n$ matrix $X = (X_{ij})$, we define $\sigma_i(X)$ by the formula

$$\det(I + tX) = 1 + \sigma_1(X) t + \sigma_2(X) t^2 + \cdots + \sigma_n(X) t^n.$$

We claim that

$$\sigma_i(X) = \sigma_i(\lambda_1, \ldots, \lambda_n),$$
where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $X$. Indeed, $\lambda_1, \ldots, \lambda_n$ are the roots the the polynomial $\det(\lambda I - X) = 0$, and as

$$det(\lambda I - X) = \prod_{i=1}^{n} (\lambda - \lambda_i) = \lambda^n + \sum_{i=1}^{n} (-1)^i \sigma_i(\lambda_1, \ldots, \lambda_n) \lambda^{n-i},$$

by factoring $\lambda^n$ and replacing $\lambda^{-1}$ by $-\lambda^{-1}$, we get

$$det(I + (-\lambda^{-1})X) = 1 + \sum_{i=1}^{n} \sigma_i(\lambda_1, \ldots, \lambda_n)(-\lambda^{-1})^n,$$

which proves our claim.

Observe that

$$\sigma_1(X) = \text{tr}(X), \quad \sigma_n(X) = \det(X).$$

Also, $\sigma_k(X^\top) = \sigma_k(X)$, since $\det(I + tX) = \det((I + tX)^\top) = \det(I + tX^\top)$. It is not very difficult to prove the following theorem:

**Theorem 29.23.** The algebra $I_n$ of invariant polynomials in $n^2$ variables is generated by $\sigma_1(X), \ldots, \sigma_n(X)$; that is,

$$I_n \cong \mathbb{R}[\sigma_1(X), \ldots, \sigma_n(X)] \quad (\text{resp. } I_n \cong \mathbb{C}[\sigma_1(X), \ldots, \sigma_n(X)]).$$

For a proof of Theorem 29.23, see Milnor and Stasheff [129] (Appendix C, Lemma 6), Madsen and Tornehave [119] (Appendix B), or Morita [133] (Theorem 5.26). The proof uses the fact that for every matrix $X$, there is an upper-triangular matrix $T$, and an invertible matrix $B$, so that

$$X = BTB^{-1}.$$ 

Then, we can replace $B$ by the matrix $\text{diag}(\epsilon, \epsilon^2, \ldots, \epsilon^n) B$, where $\epsilon$ is very small, and make the off diagonal entries arbitrarily small. By continuity, it follows that $P(X)$ depends only on the diagonal entries of $BTB^{-1}$, that is, on the eigenvalues of $X$. So, $P(X)$ must be a symmetric function of these eigenvalues, and the classical theory of symmetric functions completes the proof.

It turns out that there are situations where it is more convenient to use another set of generators instead of $\sigma_1, \ldots, \sigma_n$. Define $s_i(X)$ by

$$s_i(X) = \text{tr}(X^i).$$

Of course,

$$s_i(X) = \lambda_1^i + \cdots + \lambda_n^i,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $X$. Now, the $\sigma_i(X)$ and $s_i(X)$ are related to each other by *Newton’s formula*, namely:

$$s_i(X) - \sigma_1(X)s_{i-1}(X) + \sigma_2(X)s_{i-2}(X) + \cdots + (-1)^{i-1}\sigma_{i-1}(X)s_1(X) + (-1)^i \sigma_i(X) = 0,$$
with $1 \leq i \leq n$. A “cute” proof of the Newton formulae is obtained by computing the derivative of $\log(h(t))$, where

$$h(t) = \prod_{i=1}^{n} (1 + t\lambda_i) = 1 + \sigma_1 t + \sigma_2 t^2 + \cdots + \sigma_n t^n,$$

see Madsen and Tornehave [119] (Appendix B) or Morita [133] (Exercise 5.7).

Consequently, we can inductively compute $s_i$ in terms of $\sigma_1, \ldots, \sigma_i$, and conversely $\sigma_i$ in terms of $s_1, \ldots, s_i$. For example,

$$s_1 = \sigma_1, \quad s_2 = \sigma_1^2 - 2\sigma_2, \quad s_3 = \sigma_1^3 - 3\sigma_1 \sigma_2 + 3\sigma_3.$$

It follows that

$$I_n \cong \mathbb{R}[s_1(X), \ldots, s_n(X)] \quad \text{(resp. } I_n \cong \mathbb{C}[s_1(X), \ldots, s_n(X)]\text{)}.$$

Using the above, we can give a simple proof of Theorem 29.21, using Theorem 29.23.

**Proof.** (Proof of Theorem 29.21). Since $s_1, \ldots, s_n$ generate $I_n$, it is enough to prove that $s_i(R^N)$ is closed. We need to prove that $ds_i(R^N) = 0$, and for this, it is enough to prove it in every local trivialization $(U, \varphi)$. To simplify notation, we write $\Omega$ for $\Omega^i$.

We have

$$ds_i(\Omega) = d\Omega = \omega \wedge \Omega - \Omega \wedge \omega.$$

However, the entries in $\omega$ are one-forms, the entries in $\Omega$ are two-forms, and since

$$\eta \wedge \theta = \theta \wedge \eta$$

for all $\eta \in \mathcal{A}^1(B)$ and all $\theta \in \mathcal{A}^2(B)$ and $\text{tr}(XY) = \text{tr}(YX)$ for all matrices $X$ and $Y$ with commuting entries, we get

$$\text{tr}(d\Omega^i) = \text{tr}(\omega \wedge \Omega^i - \Omega^i \wedge \omega) = \text{tr}(\omega \wedge \Omega^i) - \text{tr}(\Omega^i \wedge \omega) = 0,$$

as required. □
A more elegant proof (also using Bianchi’s identity) can be found in Milnor and Stasheff [129] (Appendix C, page 296-298).

For real vector bundles, only invariant polynomials of even degrees matter.

**Proposition 29.24.** If \( \xi \) is a real vector bundle, then for every homogeneous invariant polynomial \( P \) of odd degree \( k \), we have \( P(\xi) = 0 \in H^{2k}_{\text{DR}}(B) \).

**Proof.** As \( I_n \cong \mathbb{R}[s_1(X), \ldots, s_n(X)] \) and \( s_i(X) \) is homogeneous of degree \( i \), it is enough to prove Proposition 29.24 for \( s_i(X) \) with \( i \) odd. By Theorem 29.22, we may assume that we pick a metric connection on \( \xi \), so that \( \Omega_\alpha \) is skew-symmetric in every local trivialization. Then, \( \Omega^i_\alpha \) is also skew symmetric and

\[
\text{tr}(\Omega^i_\alpha) = 0,
\]

since the diagonal entries of a real skew-symmetric matrix are all zero. It follows that \( s_i(\Omega_\alpha) = \text{tr}(\Omega^i_\alpha) = 0 \). \( \square \)

Proposition 29.24 implies that for a real vector bundle \( \xi \), non-zero characteristic classes can only live in the cohomology groups \( H^{4k}_{\text{DR}}(B) \) of dimension \( 4k \). This property is specific to real vector bundles and generally fails for complex vector bundles.

Before defining Pontrjagin and Chern classes, we state another important properties of the homology classes \( P(\xi) \):

**Proposition 29.25.** If \( \xi = (E, \pi, B, V) \) and \( \xi' = (E', \pi', B', V) \) are real (resp. complex) vector bundles, for every bundle map

\[
\begin{array}{ccc}
E & \xrightarrow{f_E} & E' \\
\downarrow \pi & & \downarrow \pi' \\
B & \xrightarrow{f} & B',
\end{array}
\]

for every homogeneous invariant polynomial \( P \) of degree \( k \), we have

\[
P(\xi) = f^*(P(\xi')) \in H^{2k}_{\text{DR}}(B) \quad \text{(resp. } P(\xi) = f^*(P(\xi')) \in H^{2k}_{\text{DR}}(B; \mathbb{C}) \text{)}.
\]

In particular, for every smooth map \( f : N \to B \), we have

\[
P(f^*\xi) = f^*(P(\xi)) \in H^{2k}_{\text{DR}}(N) \quad \text{(resp. } P(f^*\xi) = f^*(P(\xi)) \in H^{2k}_{\text{DR}}(N; \mathbb{C}) \text{)}.
\]

The above proposition implies that isomorphic vector bundles have identical characteristic classes. We finally define Pontrjagin classes and Chern classes.
29.6. Pontrjagin Classes and Chern Classes, a Glimpse

Definition 29.8. If $\xi$ be a real rank $n$ vector bundle, then the $k^{th}$ Pontrjagin class of $\xi$, denoted $p_k(\xi)$, where $1 \leq 2k \leq n$, is the cohomology class

$$p_k(\xi) = \left[ \frac{1}{(2\pi i)^{2k}} \sigma_{2k}(R^\nabla) \right] \in H^{2k}_{\text{DR}}(B),$$

for any connection $\nabla$ on $\xi$.

If $\xi$ be a complex rank $n$ vector bundle, then the $k^{th}$ Chern class of $\xi$, denoted $c_k(\xi)$, where $1 \leq k \leq n$, is the cohomology class

$$c_k(\xi) = \left[ \left( \frac{-1}{2\pi i} \right)^k \sigma_k(R^\nabla) \right] \in H^{2k}_{\text{DR}}(B),$$

for any connection $\nabla$ on $\xi$. We also set $p_0(\xi) = 1$, and $c_0(\xi) = 1$ in the complex case.

The strange coefficient in $p_k(\xi)$ is present so that our expression matches the topological definition of Pontrjagin classes. The equally strange coefficient in $c_k(\xi)$ is there to insure that $c_k(\xi)$ actually belongs to the real cohomology group $H^{2k}_{\text{DR}}(B)$, as stated (from the definition, we can only claim that $c_k(\xi) \in H^{2k}_{\text{DR}}(B; \mathbb{C})$).

This requires a proof which can be found in Morita [133] (Proposition 5.30), or in Madsen and Tornehave [119] (Chapter 18). One can use the fact that every complex vector bundle admits a Hermitian connection. Locally, the curvature matrices are skew-Hermitian and this easily implies that the Chern classes are real, since if $\Omega$ is skew-Hermitian, then $i\Omega$ is Hermitian. (Actually, the topological version of Chern classes shows that $c_k(\xi) \in H^{2k}(B; \mathbb{Z})$.)

If $\xi$ is a real rank $n$ vector bundle and $n$ is odd, say $n = 2m + 1$, then the “top” Pontrjagin class $p_m(\xi)$ corresponds to $\sigma_{2m}(R^\nabla)$, which is not $\text{det}(R^\nabla)$. However, if $n$ is even, say $n = 2m$, then the “top” Pontrjagin class $p_m(\xi)$ corresponds to $\text{det}(R^\nabla)$.

It is also useful to introduce the Pontrjagin polynomial $p(\xi)(t) \in H^*_{\text{DR}}(B)[t]$, given by

$$p(\xi)(t) = \det \left( I + \frac{t}{2\pi i} R^\nabla \right) = 1 + p_1(\xi)t + p_2(\xi)t^2 + \cdots + p_{\lfloor \frac{n}{2} \rfloor}(\xi)t^{\lfloor \frac{n}{2} \rfloor}$$

and the Chern polynomial $c(\xi)(t) \in H^*_{\text{DR}}(B)[t]$, given by

$$c(\xi)(t) = \det \left( I - \frac{t}{2\pi i} R^\nabla \right) = 1 + c_1(\xi)t + c_2(\xi)t^2 + \cdots + c_n(\xi)t^n.$$
case, the tangent bundle $TB$ is a rank $n$ complex vector bundle over the real manifold of dimension $2n$, and thus, it has $n$ Chern classes, denoted $c_1(B), \ldots, c_n(B)$.

The determination of the Pontrjagin classes (or Chern classes) of a manifold is an important step for the study of the geometric/topological structure of the manifold. For example, it is possible to compute the Chern classes of complex projective space $\mathbb{CP}^n$ (as a complex manifold).

The Pontrjagin classes of a real vector bundle $\xi$ are related to the Chern classes of its complexification $\xi_C = \xi \otimes \epsilon_1^C$ (where $\epsilon_1^C$ is the trivial complex line bundle $B \times \mathbb{C}$).

**Proposition 29.26.** For every real rank $n$ vector bundle $\xi = (E, \pi, B, V)$, if $\xi_C = \xi \otimes \epsilon_1^C$ is the complexification of $\xi$, then

$$p_k(\xi) = (-1)^k c_{2k}(\xi_C) \in H_{\text{DR}}^{4k}(B) \quad k \geq 0.$$ 

Basically, the reason why Proposition 29.26 holds is that

$$\frac{1}{(2\pi)^{2k}} = (-1)^k \left( \frac{-1}{2\pi i} \right)^{2k}$$

We conclude this section by stating a few more properties of Chern classes.

**Proposition 29.27.** For every complex rank $n$ vector bundle $\xi$, the following properties hold:

1. If $\xi$ has a Hermitian metric, then we have a canonical isomorphism $\xi^* \cong \overline{\xi}$.

2. The Chern classes of $\xi$, $\xi^*$ and $\overline{\xi}$ satisfy:

$$c_k(\xi^*) = c_k(\overline{\xi}) = (-1)^k c_k(\xi).$$

3. For any complex vector bundles $\xi$ and $\eta$,

$$c_k(\xi \oplus \eta) = \sum_{i=0}^{k} c_i(\xi)c_{k-i}(\eta),$$

or equivalently

$$c(\xi \oplus \eta)(t) = c(\xi)(t)c(\eta)(t),$$

and similarly for Pontrjagin classes when $\xi$ and $\eta$ are real vector bundles.

To prove (2), we can use the fact that $\xi$ can be given a Hermitian metric. Then, we saw earlier that if $\omega$ is the connection matrix of $\xi$ over $U$ then $\overline{\omega} = -\omega^\top$ is the connection matrix of $\overline{\xi}$ over $U$. However, it is clear that $\sigma_k(-\overline{\Omega}_\alpha^\top) = (-1)^k \sigma_k(\Omega_\alpha)$, and so $c_k(\overline{\xi}) = (-1)^k c_k(\xi)$.

**Remark:** For a real vector bundle $\xi$, it is easy to see that $(\xi_C)^* = (\xi^*)_C$, which implies that $c_k((\xi_C)^*) = c_k(\xi_C)$ (as $\xi \cong \xi^*$) and (2) implies that $c_k(\xi_C) = 0$ for $k$ odd. This proves again that the Pontrjagin classes exit only in dimension $4k$.

A complex rank $n$ vector bundle $\xi$ can also be viewed as a rank $2n$ vector bundle $\xi_R$ and we have:
Proposition 29.28. For every rank \( n \) complex vector bundle \( \xi \), if \( p_k = p_k(\xi_\mathbb{R}) \) and \( c_k = c_k(\xi) \), then we have

\[
1 - p_1 + p_2 + \cdots + (-1)^n p_n = (1 + c_1 + c_2 + \cdots + c_n)(1 - c_1 + c_2 + \cdots + (-1)^n c_n).
\]

29.7 Euler Classes and The Generalized Gauss-Bonnet Theorem

Let \( \xi = (E, \pi, B, V) \) be a real vector bundle of rank \( n = 2m \) and let \( \nabla \) be any metric connection on \( \xi \). Then, if \( \xi \) is orientable (as defined in Section 28.4, see Definition 28.11 and the paragraph following it), it is possible to define a global form \( \text{eu}(R^\nabla) \in \mathcal{A}^{2m}(B) \), which turns out to be closed. Furthermore, the cohomology class \( [\text{eu}(R^\nabla)] \in H^{2m}_{\text{DR}}(B) \) is independent of the choice of \( \nabla \). This cohomology class, denoted \( e(\xi) \), is called the Euler class of \( \xi \), and has some very interesting properties. For example, \( p_m(\xi) = e(\xi)^2 \).

As \( \nabla \) is a metric connection, in a trivialization \( (U_\alpha, \varphi_\alpha) \), the curvature matrix \( \Omega_\alpha \) is a skew symmetric \( 2m \times 2m \) matrix of 2-forms. Therefore, we can substitute the 2-forms in \( \Omega_\alpha \) for the variables of the Pfaffian of degree \( m \) (see Section 22.10), and we obtain the 2-\( m \)-form, \( \text{Pf}(\Omega_\alpha) \in \mathcal{A}^{2m}(B) \). Now, as \( \xi \) is orientable, the transition functions take values in \( \text{SO}(2m) \), so by Proposition 29.9, since

\[
\Omega_\beta = g_{\alpha\beta} \Omega_\alpha g_{\alpha\beta}^{-1},
\]

we conclude from Proposition 22.30 (ii) that

\[
\text{Pf}(\Omega_\alpha) = \text{Pf}(\Omega_\beta).
\]

Therefore, the local 2\( m \)-forms \( \text{Pf}(\Omega_\alpha) \) patch and define a global form \( \text{Pf}(R^\nabla) \in \mathcal{A}^{2m}(B) \).

The following propositions can be shown:

Proposition 29.29. For every real, orientable, rank 2\( m \) vector bundle \( \xi \), for every metric connection \( \nabla \) on \( \xi \), the 2\( m \)-form \( \text{Pf}(R^\nabla) \in \mathcal{A}^{2m}(B) \) is closed.

Proposition 29.30. For every real, orientable, rank 2\( m \) vector bundle \( \xi \), the cohomology class \( [\text{Pf}(R^\nabla)] \in H^{2m}_{\text{DR}}(B) \) is independent of the metric connection \( \nabla \) on \( \xi \).

Proofs of Propositions 29.29 and 29.30 can be found in Madsen and Tornehave [119] (Chapter 19) or Milnor and Stasheff [129] (Appendix C) (also see Morita [133], Chapters 5 and 6).

Definition 29.9. Let \( \xi = (E, \pi, B, V) \) be any real, orientable, rank 2\( m \) vector bundle. For any metric connection \( \nabla \) on \( \xi \), the Euler form associated with \( \nabla \) is the closed form

\[
\text{eu}(R^\nabla) = \frac{1}{(2\pi)^n} \text{Pf}(R^\nabla) \in \mathcal{A}^{2m}(B),
\]
and the *Euler class* of $\xi$ is the cohomology class

$$e(\xi) = [\text{eu}(R^\nabla)] \in H^{2m}_{DR}(B),$$

which does not depend on $\nabla$.

Some authors, including Madsen and Tornehave [119], have a negative sign in front of $R^\nabla$ in their definition of the Euler form; that is, they define $\text{eu}(R^\nabla)$ by

$$\text{eu}(R^\nabla) = \frac{1}{(2\pi)^n} \text{Pf}(-R^\nabla).$$

However these authors use a Pfaffian with the opposite sign convention from ours and this Pfaffian differs from ours by the factor $(-1)^n$ (see the warning in Section 22.10). Madsen and Tornehave [119] seem to have overlooked this point and with their definition of the Pfaffian (which is the one we have adopted) Proposition 29.32 is incorrect.

Here is the relationship between the Euler class $e(\xi)$, and the top Pontrjagin class $p_m(\xi)$:

**Proposition 29.31.** For every real, orientable, rank $2m$ vector bundle $\xi = (E, \pi, B, V)$, we have

$$p_m(\xi) = e(\xi)^2 \in H^{4m}_{DR}(B).$$

**Proof.** The top Pontrjagin class $p_m(\xi)$ is given by

$$p_m(\xi) = \left[ \frac{1}{(2\pi)^{2m}} \det(R^\nabla) \right],$$

for any (metric) connection $\nabla$, and

$$e(\xi) = [\text{eu}(R^\nabla)],$$

with

$$\text{eu}(R^\nabla) = \frac{1}{(2\pi)^n} \text{Pf}(R^\nabla).$$

From Proposition 22.30 (i), we have

$$\det(R^\nabla) = \text{Pf}(R^\nabla)^2,$$

which yields the desired result. \qed

A rank $m$ complex vector bundle $\xi = (E, \pi, B, V)$ can be viewed as a real rank $2m$ vector bundle $\xi_\mathbb{R}$, by viewing $V$ as a $2m$ dimensional real vector space. Then, it turns out that $\xi_\mathbb{R}$ is naturally orientable. Here is the reason.

For any basis, $(e_1, \ldots, e_m)$, of $V$ over $\mathbb{C}$, observe that $(e_1, ie_1, \ldots, e_m, ie_m)$ is a basis of $V$ over $\mathbb{R}$ (since $v = \sum_{i=1}^{m} \lambda_i + i\mu_i)e_i = \sum_{i=1}^{m} \lambda_i e_i + \sum_{i=1}^{m} \mu_i ie_i$). But, any $m \times m$ invertible
29.7. EULER CLASSES AND THE GENERALIZED GAUSS-BONNET THEOREM 957

matrix $A$, over $\mathbb{C}$ becomes a real $2m \times 2m$ invertible matrix $A_R$, obtained by replacing the entry $a_{jk} + ib_{jk}$ in $A$ by the real $2 \times 2$ matrix

$$
\begin{pmatrix}
  a_{jk} & -b_{jk} \\
  b_{jk} & a_{jk}
\end{pmatrix}
$$

Indeed, if $v_k = \sum_{j=1}^{m} a_{jk} e_j + \sum_{j=1}^{m} b_{jk} i e_j$, then $iv_k = \sum_{j=1}^{m} -b_{jk} e_j + \sum_{j=1}^{m} a_{jk} i e_j$ and when we express $v_k$ and $iv_k$ over the basis $(e_1, ie_1, \ldots, e_m, ie_m)$, we get a matrix $A_R$ consisting of $2 \times 2$ blocks as above. Clearly, the map $r: A \mapsto A_R$ is a continuous injective homomorphism from $\text{GL}(m, \mathbb{C})$ to $\text{GL}(2m, \mathbb{R})$. Now, it is known that $\text{GL}(m, \mathbb{C})$ is connected, thus $\text{Im}(r) = r(\text{GL}(m, \mathbb{C}))$ is connected, and as $\det(I_{2m}) = 1$, we conclude that all matrices in $\text{Im}(r)$ have positive determinant.\(^1\) Therefore, the transition functions of $\xi_R$, which take values in $\text{Im}(r)$ have positive determinant, and $\xi_R$ is orientable. We can give $\xi_R$ an orientation by fixing some basis of $V$ over $\mathbb{R}$. Then, we have the following relationship between $e(\xi_R)$ and the top Chern class, $c_m(\xi)$:

**Proposition 29.32.** For every complex, rank $m$ vector bundle $\xi = (E, \pi, B, V)$, we have

$$c_m(\xi) = e(\xi) \in H^{2m}_{DR}(B).$$

**Proof.** Pick some metric connection $\nabla$. Recall that

$$c_m(\xi) = \left[\left(\frac{-1}{2\pi i}\right)^m \det(R^\nabla)\right] = i^m \left[\left(\frac{1}{2\pi}\right)^m \det(R^\nabla)\right].$$

On the other hand,

$$e(\xi) = \left[\frac{1}{(2\pi)^m} \text{Pf}(R^\nabla)\right].$$

Here, $R^\nabla$ denotes the global $2m$-form, which locally, is equal to $\Omega_R$, where $\Omega$ is the $m \times m$ curvature matrix of $\xi$ over some trivialization. By Proposition 22.31,

$$\text{Pf}(\Omega_R) = i^n \det(\Omega),$$

so $c_m(\xi) = e(\xi)$, as claimed. \(\square\)

The Euler class enjoys many other nice properties. For example, if $f: \xi_1 \to \xi_2$ is an orientation preserving bundle map, then

$$e(f^*\xi_2) = f^*(e(\xi_2)),$$

where $f^*\xi_2$ is given the orientation induced by $\xi_2$. Also, the Euler class can be defined by topological means and it belongs to the integral cohomology group $H^{2m}(B; \mathbb{Z})$.

\(^1\)One can also prove directly that every matrix in $\text{Im}(r)$ has positive determinant by expressing $r(A)$ as a product of simple matrices whose determinants are easily computed.
Although this result lies beyond the scope of these notes, we cannot resist stating one of the most important and most beautiful theorems of differential geometry usually called the Generalized Gauss-Bonnet Theorem or Gauss-Bonnet-Chern Theorem.

For this, we need the notion of Euler characteristic. Since we haven’t discussed triangulations of manifolds, we will use a definition in terms of cohomology. Although concise, this definition is hard to motivate, and we apologize for this. Given a smooth $n$-dimensional manifold $M$, we define its Euler characteristic $\chi(M)$, as

$$\chi(M) = \sum_{i=0}^{n} (-1)^i \dim(H^i_{\text{DR}}).$$

The integers $b_i = \dim(H^i_{\text{DR}})$ are known as the Betti numbers of $M$. For example, $\chi(S^2) = 2$.

It turns out that if $M$ is an odd dimensional manifold, then $\chi(M) = 0$. This explains partially why the Euler class is only defined for even dimensional bundles.

The Generalized Gauss-Bonnet Theorem (or Gauss-Bonnet-Chern Theorem) is a generalization of the Gauss-Bonnet Theorem for surfaces. In the general form stated below it was first proved by Allendoerfer and Weil (1943), and Chern (1944).

**Theorem 29.33.** (Generalized Gauss-Bonnet Formula) Let $M$ be an orientable, smooth, compact manifold of dimension $2m$. For every metric connection $\nabla$ on $TM$, (in particular, the Levi-Civita connection for a Riemannian manifold), we have

$$\int_M e\text{u}(R^\nabla) = \chi(M).$$

A proof of Theorem 29.33 can be found in Madsen and Tornehave [119] (Chapter 21), but beware of some sign problems. The proof uses another famous theorem of differential topology, the Poincaré-Hopf Theorem. A sketch of the proof is also given in Morita [133], Chapter 5.

Theorem 29.33 is remarkable because it establishes a relationship between the geometry of the manifold (its curvature) and the topology of the manifold (the number of “holes”), somehow encoded in its Euler characteristic.

Characteristic classes are a rich and important topic and we’ve only scratched the surface. We refer the reader to the texts mentioned earlier in this section as well as to Bott and Tu [24] for comprehensive expositions.
Chapter 30
Clifford Algebras, Clifford Groups, and the Groups Pin($n$) and Spin($n$)

30.1 Introduction: Rotations As Group Actions

The main goal of this chapter is to explain how rotations in $\mathbb{R}^n$ are induced by the action of a certain group $\text{Spin}(n)$ on $\mathbb{R}^n$, in a way that generalizes the action of the unit complex numbers $U(1)$ on $\mathbb{R}^2$, and the action of the unit quaternions $SU(2)$ on $\mathbb{R}^3$ (i.e., the action is defined in terms of multiplication in a larger algebra containing both the group $\text{Spin}(n)$ and $\mathbb{R}^n$). The group $\text{Spin}(n)$, called a spinor group, is defined as a certain subgroup of units of an algebra $\text{Cl}_n$, the Clifford algebra associated with $\mathbb{R}^n$. Furthermore, for $n \geq 3$, we are lucky, because the group $\text{Spin}(n)$ is topologically simpler than the group $\text{SO}(n)$. Indeed, for $n \geq 3$, the group $\text{Spin}(n)$ is simply connected (a fact that it not so easy to prove without some machinery), whereas $\text{SO}(n)$ is not simply connected. Intuitively speaking, $\text{SO}(n)$ is more twisted than $\text{Spin}(n)$. In fact, we will see that $\text{Spin}(n)$ is a double cover of $\text{SO}(n)$.

Since the spinor groups are certain well chosen subgroups of units of Clifford algebras, it is necessary to investigate Clifford algebras to get a firm understanding of spinor groups. This chapter provides a tutorial on Clifford algebra and the groups $\text{Spin}$ and $\text{Pin}$, including a study of the structure of the Clifford algebra $\text{Cl}_{p,q}$ associated with a nondegenerate symmetric bilinear form of signature $(p,q)$ and culminating in the beautiful “8-periodicity theorem” of Elie Cartan and Raoul Bott (with proofs). We also explain when $\text{Spin}(p,q)$ is a double-cover of $\text{SO}(p,q)$. The reader should be warned that a certain amount of algebraic (and topological) background is expected. This being said, perseverant readers will be rewarded by being exposed to some beautiful and nontrivial concepts and results, including Elie Cartan and Raoul Bott “8-periodicity theorem.”

Going back to rotations as transformations induced by group actions, recall that if $V$ is a vector space, a linear action (on the left) of a group $G$ on $V$ is a map $\alpha: G \times V \rightarrow V$ satisfying the following conditions, where, for simplicity of notation, we denote $\alpha(g,v)$ by $g \cdot v$.
CHAPTER 30. CLIFFORD ALGEBRAS, CLIFFORD GROUPS, PIN AND SPIN

(1) \( g \cdot (h \cdot v) = (gh) \cdot v \), for all \( g, h \in G \) and \( v \in V \);

(2) \( 1 \cdot v = v \), for all \( v \in V \), where 1 is the identity of the group \( G \);

(3) The map \( v \mapsto g \cdot v \) is a linear isomorphism of \( V \) for every \( g \in G \).

For example, the (multiplicative) group \( \mathbb{U}(1) \) of unit complex numbers acts on \( \mathbb{R}^2 \) (by identifying \( \mathbb{R}^2 \) and \( \mathbb{C} \)) via complex multiplication: For every \( z = a + ib \) (with \( a^2 + b^2 = 1 \)), for every \( (x, y) \in \mathbb{R}^2 \) (viewing \( (x, y) \) as the complex number \( x + iy \)),

\[ z \cdot (x, y) = (ax - by, ay + bx). \]

Now, every unit complex number is of the form \( \cos \theta + i \sin \theta \), and thus the above action of \( z = \cos \theta + i \sin \theta \) on \( \mathbb{R}^2 \) corresponds to the rotation of angle \( \theta \) around the origin. In the case \( n = 2 \), the groups \( \mathbb{U}(1) \) and \( \text{SO}(2) \) are isomorphic, but this is an exception.

To represent rotations in \( \mathbb{R}^3 \) and \( \mathbb{R}^4 \), we need the quaternions. For our purposes, it is convenient to define the quaternions as certain \( 2 \times 2 \) complex matrices. Let \( 1, i, j, k \) be the matrices

\[ 1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad i = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad j = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad k = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \]

and let \( \mathbb{H} \) be the set of all matrices of the form

\[ X = a1 + bi + c j + dk, \quad a, b, c, d \in \mathbb{R}. \]

Thus, every matrix in \( \mathbb{H} \) is of the form

\[ X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R}. \]

The quaternions \( 1, i, j, k \) satisfy the famous identities discovered by Hamilton:

\[ i^2 = j^2 = k^2 = ijk = -1, \]

\[ ij = -ji = k, \]

\[ jk = -kj = i, \]

\[ ki = -ik = j. \]

As a consequence, it can be verified that \( \mathbb{H} \) is a skew field (a noncommutative field) called the quaternions. It is also a real vector space of dimension 4 with basis \( (1, i, j, k) \); thus as a vector space, \( \mathbb{H} \) is isomorphic to \( \mathbb{R}^4 \). The unit quaternions are the quaternions such that

\[ \det(X) = a^2 + b^2 + c^2 + d^2 = 1. \]

Given any quaternion \( X = a1 + bi + cj + dk \), the conjugate \( \overline{X} \) of \( X \) is given by

\[ \overline{X} = a1 - bi - cj - dk. \]
It is easy to check that the matrices associated with the unit quaternions are exactly the matrices in $\mathbf{SU}(2)$. Thus, we call $\mathbf{SU}(2)$ the group of unit quaternions.

Now we can define an action of the group of unit quaternions $\mathbf{SU}(2)$ on $\mathbb{R}^3$. For this, we use the fact that $\mathbb{R}^3$ can be identified with the pure quaternions in $\mathbb{H}$, namely, the quaternions of the form $x_1i + x_2j + x_3k$, where $(x_1, x_2, x_3) \in \mathbb{R}^3$. Then, we define the action of $\mathbf{SU}(2)$ over $\mathbb{R}^3$ by

$$Z \cdot X = ZXZ^{-1} = ZX\overline{Z},$$

where $Z \in \mathbf{SU}(2)$ and $X$ is any pure quaternion. Now, it turns out that the map $\rho_Z$ (where $\rho_Z(X) = ZX\overline{Z}$) is indeed a rotation, and that the map $\rho: Z \mapsto \rho_Z$ is a surjective homomorphism $\rho: \mathbf{SU}(2) \to \mathbf{SO}(3)$ whose kernel is $\{ -1, 1 \}$, where 1 denotes the multiplicative unit quaternion. (For details, see Gallier [72], Chapter 8).

We can also define an action of the group $\mathbf{SU}(2) \times \mathbf{SU}(2)$ over $\mathbb{R}^4$, by identifying $\mathbb{R}^4$ with the quaternions. In this case,

$$(Y, Z) \cdot X = YX\overline{Z},$$

where $(Y, Z) \in \mathbf{SU}(2) \times \mathbf{SU}(2)$ and $X \in \mathbb{H}$ is any quaternion. Then, the map $\rho_{Y,Z}$ is a rotation (where $\rho_{Y,Z}(X) = YX\overline{Z}$), and the map $\rho: (Y, Z) \mapsto \rho_{Y,Z}$ is a surjective homomorphism $\rho: \mathbf{SU}(2) \times \mathbf{SU}(2) \to \mathbf{SO}(4)$ whose kernel is $\{(1, 1), (-1, -1)\}$. (For details, see Gallier [72], Chapter 8).

Thus, we observe that for $n = 2, 3, 4$, the rotations in $\mathbf{SO}(n)$ can be realized via the linear action of some group (the case $n = 1$ is trivial, since $\mathbf{SO}(1) = \{1, -1\}$). It is also the case that the action of each group can be somehow be described in terms of multiplication in some larger algebra “containing” the original vector space $\mathbb{R}^n$ ($\mathbb{C}$ for $n = 2$, $\mathbb{H}$ for $n = 3, 4$). However, these groups appear to have been discovered in an ad hoc fashion, and there does not appear to be any universal way to define the action of these groups on $\mathbb{R}^n$. It would certainly be nice if the action was always of the form

$$Z \cdot X = ZXZ^{-1}(= ZX\overline{Z}).$$

A systematic way of constructing groups realizing rotations in terms of linear action, using a uniform notion of action, does exist. Such groups are the spinor groups, to be described in the following sections.

### 30.2 Clifford Algebras

We explained in Section 30.1 how the rotations in $\mathbf{SO}(3)$ can be realized by the linear action of the group of unit quaternions $\mathbf{SU}(2)$ on $\mathbb{R}^3$, and how the rotations in $\mathbf{SO}(4)$ can be realized by the linear action of the group $\mathbf{SU}(2) \times \mathbf{SU}(2)$ on $\mathbb{R}^4$.

The main reasons why the rotations in $\mathbf{SO}(3)$ can be represented by unit quaternions are the following:
(1) For every nonzero vector $u \in \mathbb{R}^3$, the reflection $s_u$ about the hyperplane perpendicular to $u$ is represented by the map

$$v \mapsto -uvu^{-1},$$

where $u$ and $v$ are viewed as pure quaternions in $\mathbb{H}$ (i.e., if $u = (u_1, u_2, u_3)$, then view $u$ as $u_1i + u_2j + u_3k$, and similarly for $v$).

(2) The group $\text{SO}(3)$ is generated by the reflections.

As one can imagine, a successful generalization of the quaternions, i.e., the discovery of a group $G$ inducing the rotations in $\text{SO}(n)$ via a linear action, depends on the ability to generalize properties (1) and (2) above. Fortunately, it is true that the group $\text{SO}(n)$ is generated by the hyperplane reflections. In fact, this is also true for the orthogonal group $\text{O}(n)$, and more generally for the group of direct isometries $\text{O}(\Phi)$ of any nondegenerate quadratic form $\Phi$, by the Cartan-Dieudonné theorem (for instance, see Bourbaki [25], or Gallier [72], Chapter 7, Theorem 7.2.1). In order to generalize (2), we need to understand how the group $G$ acts on $\mathbb{R}^n$. Now, the case $n = 3$ is special, because the underlying space $\mathbb{R}^3$ on which the rotations act can be embedded as the pure quaternions in $\mathbb{H}$. The case $n = 4$ is also special, because $\mathbb{R}^4$ is the underlying space of $\mathbb{H}$. The generalization to $n \geq 5$ requires more machinery, namely, the notions of Clifford groups and Clifford algebras.

As we will see, for every $n \geq 2$, there is a compact, connected (and simply connected when $n \geq 3$) group $\text{Spin}(n)$, the “spinor group,” and a surjective homomorphism $\rho: \text{Spin}(n) \to \text{SO}(n)$ whose kernel is $\{-1, 1\}$. This time, $\text{Spin}(n)$ acts directly on $\mathbb{R}^n$, because $\text{Spin}(n)$ is a certain subgroup of the group of units of the Clifford algebra $\text{Cl}_n$, and $\mathbb{R}^n$ is naturally a subspace of $\text{Cl}_n$.

The group of unit quaternions $\text{SU}(2)$ turns out to be isomorphic to the spinor group $\text{Spin}(3)$. Because $\text{Spin}(3)$ acts directly on $\mathbb{R}^3$, the representation of rotations in $\text{SO}(3)$ by elements of $\text{Spin}(3)$ may be viewed as more natural than the representation by unit quaternions. The group $\text{SU}(2) \times \text{SU}(2)$ turns out to be isomorphic to the spinor group $\text{Spin}(4)$, but this isomorphism is less obvious.

In summary, we are going to define a group $\text{Spin}(n)$ representing the rotations in $\text{SO}(n)$, for any $n \geq 1$, in the sense that there is a linear action of $\text{Spin}(n)$ on $\mathbb{R}^n$ which induces a surjective homomorphism $\rho: \text{Spin}(n) \to \text{SO}(n)$ whose kernel is $\{-1, 1\}$. Furthermore, the action of $\text{Spin}(n)$ on $\mathbb{R}^n$ is given in terms of multiplication in an algebra $\text{Cl}_n$ containing $\text{Spin}(n)$, and in which $\mathbb{R}^n$ is also embedded.

It turns out that as a bonus, for $n \geq 3$, the group $\text{Spin}(n)$ is topologically simpler than $\text{SO}(n)$, since $\text{Spin}(n)$ is simply connected, but $\text{SO}(n)$ is not. By being astute, we can also construct a group $\text{Pin}(n)$ and a linear action of $\text{Pin}(n)$ on $\mathbb{R}^n$ that induces a surjective homomorphism $\rho: \text{Pin}(n) \to \text{O}(n)$ whose kernel is $\{-1, 1\}$. The difficulty here is the presence of the negative sign in (2). We will see how Atiyah, Bott and Shapiro circumvent this problem by using a “twisted adjoint action,” as opposed to the usual adjoint action (where $v \mapsto uvu^{-1}$).
Our presentation is heavily influenced by Bröcker and tom Dieck [31] (Chapter 1, Section 6), where most details can be found. This Chapter is almost entirely taken from the first 11 pages of the beautiful and seminal paper by Atiyah, Bott and Shapiro [14], Clifford Modules, and we highly recommend it. Another excellent (but concise) exposition can be found in Kirillov [101]. A very thorough exposition can be found in two places:

1. Lawson and Michelsohn [115], where the material on $\text{Pin}(p, q)$ and $\text{Spin}(p, q)$ can be found in Chapter I.

2. Lounesto’s excellent book [118].

One may also want to consult Baker [16], Curtis [46], Porteous [143], Fulton and Harris (Lecture 20) [70], Choquet-Bruhat [43], Bourbaki [25], and Chevalley [42], a classic. The original source is Elie Cartan’s book (1937) whose translation in English appears in [34].

We begin by recalling what is an algebra over a field. Let $K$ denote any (commutative) field, although for our purposes we may assume that $K = \mathbb{R}$ (and occasionally, $K = \mathbb{C}$). Since we will only be dealing with associative algebras with a multiplicative unit, we only define algebras of this kind.

**Definition 30.1.** Given a field $K$, a $K$-algebra is a $K$-vector space $A$ together with a bilinear operation $\cdot : A \times A \to A$, called multiplication, which makes $A$ into a ring with unity 1 (or $1_A$, when we want to be very precise). This means that $\cdot$ is associative and that there is a multiplicative identity element 1 so that $1 \cdot a = a \cdot 1 = a$, for all $a \in A$. Given two $K$-algebras $A$ and $B$, a $K$-algebra homomorphism $h : A \to B$ is a linear map that is also a ring homomorphism, with $h(1_A) = 1_B$.

For example, the ring $M_n(K)$ of all $n \times n$ matrices over a field $K$ is a $K$-algebra.

There is an obvious notion of ideal of a $K$-algebra: An ideal $\mathfrak{a} \subseteq A$ is a linear subspace of $A$ that is also a two-sided ideal with respect to multiplication in $A$. If the field $K$ is understood, we usually simply say an algebra instead of a $K$-algebra.

We will also need tensor products. A rather detailed exposition of tensor products is given in Chapter 21 and the reader may want to review Section 21.2. For the reader’s convenience, we recall the definition of the tensor product of vector spaces. The basic idea is that tensor products allow us to view multilinear maps as linear maps. The maps become simpler, but the spaces (product spaces) become more complicated (tensor products). For more details, see Section 21.2 or Atiyah and Macdonald [12].

**Definition 30.2.** Given two $K$-vector spaces $E$ and $F$, a tensor product of $E$ and $F$ is a pair $(E \otimes F, \otimes)$, where $E \otimes F$ is a $K$-vector space and $\otimes : E \times F \to E \otimes F$ is a bilinear map, so that for every $K$-vector space $G$ and every bilinear map $f : E \times F \to G$, there is a unique linear map $f_\otimes : E \otimes F \to G$ with

$$f(u, v) = f_\otimes(u \otimes v) \quad \text{for all } u \in E \text{ and all } v \in V,$$
as in the diagram below:

\[ \begin{array}{ccc}
E \times F & \rightarrow & E \otimes F \\
\downarrow f & & \downarrow f \otimes \\
G & & 
\end{array} \]

The vector space \( E \otimes F \) is defined up to isomorphism. The vectors \( u \otimes v \), where \( u \in E \) and \( v \in F \), generate \( E \otimes F \).

**Remark:** We should really denote the tensor product of \( E \) and \( F \) by \( E \otimes_K F \), since it depends on the field \( K \). Since we usually deal with a fixed field \( K \), we use the simpler notation \( E \otimes F \).

As shown in Section 21.4, we have natural isomorphisms

\[(E \otimes F) \otimes G \approx E \otimes (F \otimes G) \quad \text{and} \quad E \otimes F \approx F \otimes E.\]

Given two linear maps \( f: E \rightarrow F \) and \( g: E' \rightarrow F' \), we have a unique bilinear map \( f \times g: E \times E' \rightarrow F \times F' \) so that

\[(f \times g)(a,a') = (f(a),g(a')) \quad \text{for all } a \in E \text{ and all } a' \in E'.\]

Thus, we have the bilinear map \( \otimes \circ (f \times g): E \times E' \rightarrow F \otimes F' \), and so, there is a unique linear map \( f \otimes g: E \otimes E' \rightarrow F \otimes F' \) so that

\[(f \otimes g)(a \otimes a') = f(a) \otimes g(a') \quad \text{for all } a \in E \text{ and all } a' \in E'.\]

Let us now assume that \( E \) and \( F \) are \( K \)-algebras. We want to make \( E \otimes F \) into a \( K \)-algebra. Since the multiplication operations \( m_E: E \times E \rightarrow E \) and \( m_F: F \times F \rightarrow F \) are bilinear, we get linear maps \( m'_E: E \otimes E \rightarrow E \) and \( m'_F: F \otimes F \rightarrow F \), and thus the linear map

\[ m'_E \otimes m'_F: (E \otimes E) \otimes (F \otimes F) \rightarrow E \otimes F. \]

Using the isomorphism \( \tau: (E \otimes E) \otimes (F \otimes F) \rightarrow (E \otimes F) \otimes (E \otimes F) \), we get a linear map

\[ m_{E \otimes F}: (E \otimes F) \otimes (E \otimes F) \rightarrow E \otimes F, \]

which defines a multiplication \( m \) on \( E \otimes F \) (namely, \( m(u,v) = m_{E \otimes F}(u \otimes v) \)). It is easily checked that \( E \otimes F \) is indeed a \( K \)-algebra under the multiplication \( m \). Using the simpler notation \( \cdot \) for \( m \), we have

\[(a \otimes a') \cdot (b \otimes b') = (ab) \otimes (a'b')\]

for all \( a,b \in E \) and all \( a',b' \in F \).

Given any vector space \( V \) over a field \( K \), there is a special \( K \)-algebra \( T(V) \) together with a linear map \( i: V \rightarrow T(V) \), with the following universal mapping property: Given any
$K$-algebra $A$, for any linear map $f: V \to A$, there is a unique $K$-algebra homomorphism $\overline{f}: T(V) \to A$ so that

$$f = \overline{f} \circ i,$$

as in the diagram below:

$$\begin{array}{ccc}
V & \xrightarrow{i} & T(V) \\
\downarrow{f} & & \downarrow{\overline{f}} \\
& A & 
\end{array}$$

The algebra $T(V)$ is the tensor algebra of $V$; see Section 21.6. The algebra $T(V)$ may be constructed as the direct sum

$$T(V) = \bigoplus_{i \geq 0} V^\otimes i,$$

where $V^0 = K$, and $V^\otimes i$ is the $i$-fold tensor product of $V$ with itself. For every $i \geq 0$, there is a natural injection $\iota_n: V^\otimes n \to T(V)$, and in particular, an injection $\iota_0: K \to T(V)$. The multiplicative unit $1$ of $T(V)$ is the image $\iota_0(1)$ in $T(V)$ of the unit $1$ of the field $K$. Since every $v \in T(V)$ can be expressed as a finite sum

$$v = v_1 + \cdots + v_k,$$

where $v_i \in V^\otimes n_i$ and the $n_i$ are natural numbers with $n_i \neq n_j$ if $i \neq j$, to define multiplication in $T(V)$, using bilinearity, it is enough to define the multiplication $V^{\otimes m} \times V^{\otimes n} \to V^{\otimes (m+n)}$. Of course, this is defined by

$$(v_1 \otimes \cdots \otimes v_m) \cdot (w_1 \otimes \cdots \otimes w_n) = v_1 \otimes \cdots \otimes v_m \otimes w_1 \otimes \cdots \otimes w_n.$$

(This has to be made rigorous by using isomorphisms involving the associativity of tensor products; for details, see see Atiyah and Macdonald [12].) The algebra $T(V)$ is an example of a graded algebra, where the homogeneous elements of rank $n$ are the elements in $V^{\otimes n}$.

**Remark:** It is important to note that multiplication in $T(V)$ is not commutative. Also, in all rigor, the unit $1$ of $T(V)$ is not equal to $1$, the unit of the field $K$. However, in view of the injection $\iota_0: K \to T(V)$, for the sake of notational simplicity, we will denote $1$ by $1$. More generally, in view of the injections $\iota_n: V^{\otimes n} \to T(V)$, we identify elements of $V^{\otimes n}$ with their images in $T(V)$.

Most algebras of interest arise as well-chosen quotients of the tensor algebra $T(V)$. This is true for the exterior algebra $\bigwedge^* V$ (also called Grassmann algebra), where we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v$, where $v \in V$, see Section 22.5.

A Clifford algebra may be viewed as a refinement of the exterior algebra, in which we take the quotient of $T(V)$ modulo the ideal generated by all elements of the form $v \otimes v - \Phi(v) \cdot 1$, where $\Phi$ is the quadratic form associated with a symmetric bilinear form $\varphi: V \times V \to K$, and $\cdot: K \times T(V) \to T(V)$ denotes the scalar product of the algebra $T(V)$. For simplicity, let us assume that we are now dealing with real algebras.
Definition 30.3. Let $V$ be a real finite-dimensional vector space together with a symmetric bilinear form $\varphi: V \times V \to \mathbb{R}$ and associated quadratic form $\Phi(v) = \varphi(v,v)$. A Clifford algebra associated with $V$ and $\Phi$ is a real algebra $\text{Cl}(V, \Phi)$ together with a linear map $i_\Phi: V \to \text{Cl}(V, \Phi)$ satisfying the condition $(i_\Phi(v))^2 = \Phi(v) \cdot 1$ for all $v \in V$, and so that for every real algebra $A$ and every linear map $f: V \to A$ with $(f(v))^2 = \Phi(v) \cdot 1$ for all $v \in V$, there is a unique algebra homomorphism $\overline{f}: \text{Cl}(V, \Phi) \to A$ so that

$$f = \overline{f} \circ i_\Phi,$$

as in the diagram below:

$$
\begin{array}{ccc}
V & \xrightarrow{i_\Phi} & \text{Cl}(V, \Phi) \\
\downarrow f & & \downarrow \overline{f} \\
A & & 
\end{array}
$$

We use the notation $\lambda \cdot u$ for the product of a scalar $\lambda \in \mathbb{R}$ and of an element $u$ in the algebra $\text{Cl}(V, \Phi)$, and juxtaposition $uv$ for the multiplication of two elements $u$ and $v$ in the algebra $\text{Cl}(V, \Phi)$.

By a familiar argument, any two Clifford algebras associated with $V$ and $\Phi$ are isomorphic. We often denote $i_\Phi$ by $i$.

To show the existence of $\text{Cl}(V, \Phi)$, observe that $T(V)/\mathfrak{a}$ does the job, where $\mathfrak{a}$ is the ideal of $T(V)$ generated by all elements of the form $v \otimes v - \Phi(v) \cdot 1$, where $v \in V$. The map $i_\Phi: V \to \text{Cl}(V, \Phi)$ is the composition

$$
V \xrightarrow{i_1} T(V) \xrightarrow{\pi} T(V)/\mathfrak{a},
$$

where $\pi$ is the natural quotient map. We often denote the Clifford algebra $\text{Cl}(V, \Phi)$ simply by $\text{Cl}(\Phi)$.

Remark: Observe that Definition 30.3 does not assert that $i_\Phi$ is injective or that there is an injection of $\mathbb{R}$ into $\text{Cl}(V, \Phi)$, but we will prove later that both facts are true when $V$ is finite-dimensional. Also, as in the case of the tensor algebra, the unit of the algebra $\text{Cl}(V, \Phi)$ and the unit of the field $\mathbb{R}$ are not equal.

Since

$$\Phi(u + v) - \Phi(u) - \Phi(v) = 2\varphi(u,v)$$

and

$$(i(u + v))^2 = (i(u))^2 + (i(v))^2 + i(u)i(v) + i(v)i(u),$$

using the fact that

$$i(u)^2 = \Phi(u) \cdot 1,$$
we get
\[ i(u)i(v) + i(v)i(u) = 2\varphi(u, v) \cdot 1. \]

As a consequence, if \((u_1, \ldots, u_n)\) is an orthogonal basis w.r.t. \(\varphi\) (which means that \(\varphi(u_j, u_k) = 0\) for all \(j \neq k\)), we have
\[ i(u_j)i(u_k) + i(u_k)i(u_j) = 0 \quad \text{for all } j \neq k. \]

**Remark:** Certain authors drop the unit 1 of the Clifford algebra \(\text{Cl}(V, \Phi)\) when writing the identities
\[ i(u)^2 = \Phi(u) \cdot 1 \]
and
\[ 2\varphi(u, v) \cdot 1 = i(u)i(v) + i(v)i(u), \]
where the second identity is often written as
\[ \varphi(u, v) = \frac{1}{2}(i(u)i(v) + i(v)i(u)). \]

This is very confusing and technically wrong, because we only have an injection of \(\mathbb{R}\) into \(\text{Cl}(V, \Phi)\), but \(\mathbb{R}\) is not a subset of \(\text{Cl}(V, \Phi)\).

We warn the readers that Lawson and Michelsohn [115] adopt the opposite of our sign convention in defining Clifford algebras, i.e., they use the condition
\[ (f(v))^2 = -\Phi(v) \cdot 1 \quad \text{for all } v \in V. \]

The most confusing consequence of this is that their \(\text{Cl}(p, q)\) is our \(\text{Cl}(q, p)\).

Observe that when \(\Phi \equiv 0\) is the quadratic form identically zero everywhere, then the Clifford algebra \(\text{Cl}(V, 0)\) is just the exterior algebra \(\wedge^* V\).

**Example 30.1.** Let \(V = \mathbb{R}, e_1 = 1\), and assume that \(\Phi(x_1e_1) = -x_1^2\). Then, \(\text{Cl}(\Phi)\) is spanned by the basis \((1, e_1)\). We have
\[ e_1^2 = -1. \]

Under the bijection
\[ e_1 \mapsto i, \]
the Clifford algebra \(\text{Cl}(\Phi)\), also denoted by \(\text{Cl}_1\), is isomorphic to the algebra of complex numbers \(\mathbb{C}\).

Now, let \(V = \mathbb{R}^2, (e_1, e_2)\) be the canonical basis, and assume that \(\Phi(x_1e_1 + x_2e_2) = -(x_1^2 + x_2^2)\). Then, \(\text{Cl}(\Phi)\) is spanned by the basis \((1, e_1, e_2, e_1e_2)\). Furthermore, we have
\[ e_2e_1 = -e_1e_2, \quad e_1^2 = -1, \quad e_2^2 = -1, \quad (e_1e_2)^2 = -1. \]
Under the bijection
\[ e_1 \mapsto i, \quad e_2 \mapsto j, \quad e_1 e_2 \mapsto k, \]
it is easily checked that the quaternion identities
\[
\begin{align*}
i^2 &= j^2 = k^2 = -1, \\
ij &= -ji = k, \\
jk &= -kj = i, \\
ki &= -ik = j,
\end{align*}
\]
hold, and thus the Clifford algebra \( \text{Cl}(\Phi) \), also denoted by \( \text{Cl}_2 \), is isomorphic to the algebra of quaternions \( \mathbb{H} \).

Our prime goal is to define an action of \( \text{Cl}(\Phi) \) on \( V \) in such a way that by restricting this action to some suitably chosen multiplicative subgroups of \( \text{Cl}(\Phi) \), we get surjective homomorphisms onto \( \text{O}(\Phi) \) and \( \text{SO}(\Phi) \), respectively. The key point is that a reflection in \( V \) about a hyperplane \( H \) orthogonal to a vector \( w \) can be defined by such an action, but some negative sign shows up. A correct handling of signs is a bit subtle and requires the introduction of a canonical anti-automorphism \( t \), and of a canonical automorphism \( \alpha \), defined as follows:

**Proposition 30.1.** Every Clifford algebra \( \text{Cl}(\Phi) \) possesses a canonical anti-automorphism \( t: \text{Cl}(\Phi) \rightarrow \text{Cl}(\Phi) \) satisfying the properties

\[
t(xy) = t(y)t(x), \quad t \circ t = \text{id}, \quad \text{and} \quad t(i(v)) = i(v),
\]
for all \( x, y \in \text{Cl}(\Phi) \) and all \( v \in V \). Furthermore, such an anti-automorphism is unique.

**Proof.** Consider the opposite algebra \( \text{Cl}(\Phi)^\circ \), in which the product of \( x \) and \( y \) is given by \( yx \). It has the universal mapping property. Thus, we get a unique isomorphism \( t \), as in the diagram below:

\[
\begin{array}{ccc}
V & \xrightarrow{i} & \text{Cl}(V, \Phi) \\
\downarrow{i} & & \downarrow{t} \\
& \text{Cl}(\Phi)^\circ &
\end{array}
\]

We also denote \( t(x) \) by \( x^t \). When \( V \) is finite-dimensional, for a more palatable description of \( t \) in terms of a basis of \( V \), see the paragraph following Theorem 30.4.

The canonical automorphism \( \alpha \) is defined using the proposition

**Proposition 30.2.** Every Clifford algebra \( \text{Cl}(\Phi) \) has a unique canonical automorphism \( \alpha: \text{Cl}(\Phi) \rightarrow \text{Cl}(\Phi) \) satisfying the properties

\[
\alpha \circ \alpha = \text{id}, \quad \text{and} \quad \alpha(i(v)) = -i(v),
\]
for all \( v \in V \).
Proof. Consider the linear map \( \alpha_0 : V \rightarrow \text{Cl}(\Phi) \) defined by \( \alpha_0(v) = -i(v) \), for all \( v \in V \). We get a unique homomorphism \( \alpha \) as in the diagram below:

\[
\begin{array}{ccc}
V & \xrightarrow{i} & \text{Cl}(V, \Phi) \\
\downarrow{\alpha_0} & & \downarrow{\alpha} \\
\text{Cl}(\Phi) & & \text{Cl}(\Phi)
\end{array}
\]

Furthermore, every \( x \in \text{Cl}(\Phi) \) can be written as \( x = x_1 \cdots x_m \), with \( x_j \in i(V) \), and since \( \alpha(x_j) = -x_j \), we get \( \alpha \circ \alpha = \text{id} \). It is clear that \( \alpha \) is bijective. \( \square \)

Again, when \( V \) is finite-dimensional, a more palatable description of \( \alpha \) in terms of a basis of \( V \) can be given. If \( (e_1, \ldots, e_n) \) is a basis of \( V \), then the Clifford algebra \( \text{Cl}(\Phi) \) consists of certain kinds of “polynomials,” linear combinations of monomials of the form \( \sum_J \lambda_J e_J \), where \( J = \{i_1, i_2, \ldots, i_k\} \) is any subset (possibly empty) of \( \{1, \ldots, n\} \), with \( 1 \leq i_1 < i_2 \cdots < i_k \leq n \), and the monomial \( e_J \) is the “product” \( e_{i_1} e_{i_2} \cdots e_{i_k} \). The map \( \alpha \) is the linear map defined on monomials by

\[
\alpha(e_{i_1} e_{i_2} \cdots e_{i_k}) = (-1)^k e_{i_1} e_{i_2} \cdots e_{i_k}.
\]

For a more rigorous explanation, see the paragraph following Theorem 30.4.

We now show that if \( V \) has dimension \( n \), then \( i \) is injective and \( \text{Cl}(\Phi) \) has dimension \( 2^n \). A clever way of doing this is to introduce a graded tensor product.

First, observe that

\[
\text{Cl}(\Phi) = \text{Cl}^0(\Phi) \oplus \text{Cl}^1(\Phi),
\]

where

\[
\text{Cl}^i(\Phi) = \{ x \in \text{Cl}(\Phi) \mid \alpha(x) = (-1)^i x \}, \quad \text{where } i = 0, 1.
\]

We say that we have a \( \mathbb{Z}/2 \)-grading, which means that if \( x \in \text{Cl}^i(\Phi) \) and \( y \in \text{Cl}^j(\Phi) \), then \( xy \in \text{Cl}^{i+j \mod 2}(\Phi) \).

When \( V \) is finite-dimensional, since every element of \( \text{Cl}(\Phi) \) is a linear combination of the form \( \sum_J \lambda_J e_J \) as explained earlier, in view of the description of \( \alpha \) given above, we see that the elements of \( \text{Cl}^0(\Phi) \) are those for which the monomials \( e_J \) are products of an even number of factors, and the elements of \( \text{Cl}^1(\Phi) \) are those for which the monomials \( e_J \) are products of an odd number of factors.

Remark: Observe that \( \text{Cl}^0(\Phi) \) is a subalgebra of \( \text{Cl}(\Phi) \), whereas \( \text{Cl}^1(\Phi) \) is not.

Given two \( \mathbb{Z}/2 \)-graded algebras \( A = A^0 \oplus A^1 \) and \( B = B^0 \oplus B^1 \), their graded tensor product \( A \hat{\otimes} B \) is defined by

\[
(A \hat{\otimes} B)^0 = (A^0 \otimes B^0) \oplus (A^1 \otimes B^1),
\]

\[
(A \hat{\otimes} B)^1 = (A^0 \otimes B^1) \oplus (A^1 \otimes B^0),
\]
with multiplication
\[(a' \otimes b)(a \otimes b') = (-1)^{ij}(a'a) \otimes (bb'),\]
for \(a \in A^i\) and \(b \in B^j\). The reader should check that \(A \hat{\otimes} B\) is indeed \(\mathbb{Z}/2\)-graded.

**Proposition 30.3.** Let \(V\) and \(W\) be finite dimensional vector spaces with quadratic forms \(\Phi\) and \(\Psi\). Then, there is a quadratic form \(\Phi \oplus \Psi\) on \(V \oplus W\) defined by
\[
(\Phi + \Psi)(v, w) = \Phi(v) + \Psi(w).
\]
If we write \(i: V \to \text{Cl}(\Phi)\) and \(j: W \to \text{Cl}(\Psi)\), we can define a linear map
\[
f: V \oplus W \to \text{Cl}(\Phi) \hat{\otimes} \text{Cl}(\Psi)
\]
by
\[
f(v, w) = i(v) \otimes 1 + 1 \otimes j(w).
\]
Furthermore, the map \(f\) induces an isomorphism (also denoted by \(f\))
\[
f: \text{Cl}(V \oplus W) \to \text{Cl}(\Phi) \hat{\otimes} \text{Cl}(\Psi).
\]

**Proof.** See Bröcker and tom Dieck [31], Chapter 1, Section 6, page 57.

As a corollary, we obtain the following result:

**Theorem 30.4.** For every vector space \(V\) of finite dimension \(n\), the map \(i: V \to \text{Cl}(\Phi)\) is injective. Given a basis \((e_1, \ldots, e_n)\) of \(V\), the \(2^n - 1\) products
\[
i(e_{i_1})i(e_{i_2}) \cdots i(e_{i_k}), \quad 1 \leq i_1 < i_2 \cdots < i_k \leq n,
\]
and 1 form a basis of \(\text{Cl}(\Phi)\). Thus, \(\text{Cl}(\Phi)\) has dimension \(2^n\).

**Proof.** The proof is by induction on \(n = \dim(V)\). For \(n = 1\), the tensor algebra \(T(V)\) is just the polynomial ring \(\mathbb{R}[X]\), where \(i(e_1) = X\). Thus, \(\text{Cl}(\Phi) = \mathbb{R}[X]/(X^2 - \Phi(e_1))\), and the result is obvious. Since
\[
i(e_j)i(e_k) - i(e_k)i(e_j) = 2 \varphi(e_i, e_j) \cdot 1,
\]
it is clear that the products
\[
i(e_{i_1})i(e_{i_2}) \cdots i(e_{i_k}), \quad 1 \leq i_1 < i_2 \cdots < i_k \leq n,
\]
and 1 generate \(\text{Cl}(\Phi)\). Now, there is always a basis that is orthogonal with respect to \(\varphi\) (for example, see Artin [10], Chapter 7, or Gallier [72], Chapter 6, Problem 6.14), and thus, we have a splitting
\[
(V, \Phi) = \bigoplus_{k=1}^{n} (V_k, \Phi_k),
\]
where \(V_k\) has dimension 1. Choosing a basis so that \(e_k \in V_k\), the theorem follows by induction from Proposition 30.3.
Since $i$ is injective, for simplicity of notation, from now on we write $u$ for $i(u)$. Theorem 30.4 implies that if $(e_1, \ldots, e_n)$ is an orthogonal basis of $V$, then $\text{Cl}(\Phi)$ is the algebra presented by the generators $(e_1, \ldots, e_n)$ and the relations

$$e_j^2 = \Phi(e_j) \cdot 1, \quad 1 \leq j \leq n,$$

$$e_j e_k = -e_k e_j, \quad 1 \leq j, k \leq n, \quad j \neq k.$$ 

If $V$ has finite dimension $n$ and $(e_1, \ldots, e_n)$ is a basis of $V$, by Theorem 30.4, the maps $t$ and $\alpha$ are completely determined by their action on the basis elements. Namely, $t$ is defined by

$$t(e_i) = e_i$$

$$t(e_{i_1} e_{i_2} \cdots e_{i_k}) = e_{i_k} e_{i_{k-1}} \cdots e_{i_1},$$

where $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and of course, $t(1) = 1$. The map $\alpha$ is defined by

$$\alpha(e_i) = -e_i$$

$$\alpha(e_{i_1} e_{i_2} \cdots e_{i_k}) = (-1)^k e_{i_1} e_{i_2} \cdots e_{i_k}$$

where $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and of course, $\alpha(1) = 1$. Furthermore, the even-graded elements (the elements of $\text{Cl}^0(\Phi)$) are those generated by $1$ and the basis elements consisting of an even number of factors $e_{i_1} e_{i_2} \cdots e_{i_k}$, and the odd-graded elements (the elements of $\text{Cl}^1(\Phi)$) are those generated by the basis elements consisting of an odd number of factors $e_{i_1} e_{i_2} \cdots e_{i_{2k+1}}$.

We are now ready to define the Clifford group and investigate some of its properties.

### 30.3 Clifford Groups

First, we define conjugation on a Clifford algebra $\text{Cl}(\Phi)$ as the map

$$x \mapsto \overline{x} = t(\alpha(x)) \quad \text{for all } x \in \text{Cl}(\Phi).$$

Observe that

$$t \circ \alpha = \alpha \circ t.$$ 

If $V$ has finite dimension $n$ and $(e_1, \ldots, e_n)$ is a basis of $V$, in view of previous remarks, conjugation is defined by

$$\overline{e_i} = -e_i$$

$$\overline{e_{i_1} e_{i_2} \cdots e_{i_k}} = (-1)^k e_{i_1} e_{i_{k-1}} \cdots e_{i_k}$$

where $1 \leq i_1 < i_2 \cdots < i_k \leq n$, and of course, $\overline{1} = 1$. Conjugation is an anti-automorphism.
CHAPTER 30. CLIFFORD ALGEBRAS, CLIFFORD GROUPS, PIN AND SPIN

The multiplicative group of invertible elements of \( Cl(\Phi) \) is denoted by \( Cl(\Phi)^* \). Observe that for any \( x \in V \), if \( \Phi(x) \neq 0 \), then \( x \) is invertible because \( x^2 = \Phi(x) \); that is, \( x \in Cl(\Phi)^* \).

We would like \( Cl(\Phi)^* \) to act on \( V \) via

\[
x \cdot v = \alpha(x)vx^{-1},
\]

where \( x \in Cl(\Phi)^* \) and \( v \in V \). In general, there is no reason why \( \alpha(x)vx^{-1} \) should be in \( V \) or why this action defines an automorphism of \( V \), so we restrict this map to the subset \( \Gamma(\Phi) \) of \( Cl(\Phi)^* \) as follows.

**Definition 30.4.** Given a finite dimensional vector space \( V \) and a quadratic form \( \Phi \) on \( V \), the \textit{Clifford group of} \( \Phi \) is the group

\[
\Gamma(\Phi) = \{ x \in Cl(\Phi)^* \mid \alpha(x)vx^{-1} \in V \text{ for all } v \in V \}.
\]

The map \( N: Cl(Q) \to Cl(Q) \) given by

\[
N(x) = x\overline{x}
\]

is called the \textit{norm} of \( Cl(\Phi) \).

For any \( x \in \Gamma(\Phi) \), let \( \rho_x: V \to V \) be the map defined by

\[
v \mapsto \alpha(x)vx^{-1}, \quad v \in V.
\]

It is not entirely obvious why the map \( \rho: \Gamma(\Phi) \to GL(V) \) given by \( x \mapsto \rho_x \) is a linear action, and for that matter, why \( \Gamma(\Phi) \) is a group. This is because \( V \) is finite-dimensional and \( \alpha \) is an automorphism.

**Proof.** For any \( x \in \Gamma(\Phi) \), the map \( \rho_x \) from \( V \) to \( V \) defined by

\[
v \mapsto \alpha(x)vx^{-1}
\]

is clearly linear. If \( \alpha(x)vx^{-1} = 0 \), since by hypothesis \( x \) is invertible and since \( \alpha \) is an automorphism \( \alpha(x) \) is also invertible, so \( v = 0 \). Thus our linear map is injective, and since \( V \) has finite dimension, it is bijective. To prove that \( x^{-1} \in \Gamma(\Phi) \), pick any \( v \in V \). Since the linear map \( \rho_x \) is bijective, there is some \( w \in V \) such that \( \rho_x(w) = v \), which means that \( \alpha(x)wx^{-1} = v \). Since \( x \) is invertible and \( \alpha \) is an automorphism, we get

\[
\alpha(x^{-1})vx = w,
\]

so \( \alpha(x^{-1})vx \in V \); since this holds for any \( v \in V \), we have \( x^{-1} \in \Gamma(\Phi) \). Since \( \alpha \) is an automorphism, if \( x, y \in \Gamma(\Phi) \), for any \( v \in V \) we have

\[
\rho_y(\rho_x(v)) = \alpha(y)\alpha(x)vx^{-1}y^{-1} = \alpha(\alpha(x)v)vx^{-1}y^{-1} = \rho_{yx}(v),
\]

which shows that \( \rho_{yx} \) is a linear automorphism of \( V \), so \( yx \in \Gamma(\Phi) \) and \( \rho \) is a homomorphism. Therefore, \( \Gamma(\Phi) \) is a group and \( \rho \) is a linear representation. \( \square \)
We also define the group \( \Gamma^+(\Phi) \), called the *special Clifford group*, by

\[
\Gamma^+(\Phi) = \Gamma(\Phi) \cap \text{Cl}^0(\Phi).
\]

Observe that \( N(v) = -\Phi(v) \cdot 1 \) for all \( v \in V \). Also, if \( (e_1, \ldots, e_n) \) is a basis of \( V \), we leave it as an exercise to check that

\[
N(e_{i_1} e_{i_2} \cdots e_{i_k}) = (-1)^k \Phi(e_{i_1}) \Phi(e_{i_2}) \cdots \Phi(e_{i_k}) \cdot 1.
\]

**Remark:** The map \( \rho: \Gamma(\Phi) \to \text{GL}(V) \) given by \( x \mapsto \rho_x \) is called the *twisted adjoint representation*. It was introduced by Atiyah, Bott and Shapiro [14]. It has the advantage of not introducing a spurious negative sign, i.e., when \( v \in V \) and \( \Phi(v) \neq 0 \), the map \( \rho_v \) is the reflection \( s_v \) about the hyperplane orthogonal to \( v \) (see Proposition 30.6). Furthermore, when \( \Phi \) is nondegenerate, the kernel \( \text{Ker}(\rho) \) of the representation \( \rho \) is given by \( \text{Ker}(\rho) = \mathbb{R}^* \cdot 1 \), where \( \mathbb{R}^* = \mathbb{R} - \{0\} \). The earlier *adjoint representation* (used by Chevalley [42] and others) is given by

\[
v \mapsto xv x^{-1}.
\]

Unfortunately, in this case, \( \rho_x \) represents \(-s_v\), where \( s_v \) is the reflection about the hyperplane orthogonal to \( v \). Furthermore, the kernel of the representation \( \rho \) is generally bigger than \( \mathbb{R}^* \cdot 1 \). This is the reason why the twisted adjoint representation is preferred (and must be used for a proper treatment of the \( \text{Pin} \) group).

**Proposition 30.5.** The maps \( \alpha \) and \( t \) induce an automorphism and an anti-automorphism of the Clifford group, \( \Gamma(\Phi) \).

**Proof.** It is not very instructive; see Bröcker and tom Dieck [31], Chapter 1, Section 6, page 58. \( \square \)

The following proposition shows why Clifford groups generalize the quaternions.

**Proposition 30.6.** Let \( V \) be a finite dimensional vector space and \( \Phi \) a quadratic form on \( V \). For every element \( x \) of the Clifford group \( \Gamma(\Phi) \), if \( x \in V \) and \( \Phi(x) \neq 0 \), then the map \( \rho_x: V \to V \) given by

\[
v \mapsto \alpha(x)vx^{-1} \quad \text{for all } v \in V
\]

is the reflection about the hyperplane \( H \) orthogonal to the vector \( x \).

**Proof.** Recall that the reflection \( s \) about the hyperplane \( H \) orthogonal to the vector \( x \) is given by

\[
s(u) = u - 2 \frac{\varphi(u, x)}{\Phi(x)} \cdot x.
\]

However, we have

\[
x^2 = \Phi(x) \cdot 1 \quad \text{and} \quad ux + xu = 2\varphi(u, x) \cdot 1.
\]
Thus, we have

\[ s(u) = u - 2 \frac{\varphi(u, x)}{\Phi(x)} \cdot x \]

\[ = u - 2\varphi(u, x) \cdot \left( \frac{1}{\Phi(x)} \cdot x \right) \]

\[ = u - 2\varphi(u, x) \cdot x^{-1} \]

\[ = u - 2\varphi(u, x) \cdot (1x^{-1}) \]

\[ = u - (2\varphi(u, x) \cdot 1)x^{-1} \]

\[ = u - (ux + xu)x^{-1} \]

\[ = -xux^{-1} \]

\[ = \alpha(x)ux^{-1}, \]

since \( \alpha(x) = -x \), for \( x \in V \).  

Recall that the linear representation

\[ \rho: \Gamma(\Phi) \to GL(V) \]

is given by

\[ \rho(x)(v) = \alpha(x)vx^{-1}, \]

for all \( x \in \Gamma(\Phi) \) and all \( v \in V \). We would like to show that \( \rho \) is a surjective homomorphism from \( \Gamma(\Phi) \) onto \( O(\varphi) \), and a surjective homomorphism from \( \Gamma^+(\Phi) \) onto \( SO(\varphi) \). For this, we will need to assume that \( \varphi \) is nondegenerate, which means that for every \( v \in V \), if \( \varphi(v, w) = 0 \) for all \( w \in V \), then \( v = 0 \). For simplicity of exposition, we first assume that \( \Phi \) is the quadratic form on \( \mathbb{R}^n \) defined by

\[ \Phi(x_1, \ldots, x_n) = -(x_1^2 + \cdots + x_n^2). \]

Let \( Cl_n \) denote the Clifford algebra \( Cl(\Phi) \) and \( \Gamma_n \) denote the Clifford group \( \Gamma(\Phi) \). The following lemma plays a crucial role:

**Lemma 30.7.** The kernel of the map \( \rho: \Gamma_n \to GL(n) \) is \( \mathbb{R}^* \cdot 1 \), the multiplicative group of nonzero scalar multiples of \( 1 \in Cl_n \).

**Proof.** If \( \rho(x) = id \), then

\[ \alpha(x)v = vx \quad \text{for all } v \in \mathbb{R}^n. \tag{1} \]

Since \( Cl_n = Cl_n^0 \oplus Cl_n^1 \), we can write \( x = x^0 + x^1 \), with \( x^i \in Cl_n^i \) for \( i = 0, 1 \). Then, equation (1) becomes

\[ x^0v = vx^0 \quad \text{and} \quad -x^1v = vx^1 \quad \text{for all } v \in \mathbb{R}^n. \tag{2} \]

Using Theorem 30.4, we can express \( x^0 \) as a linear combination of monomials in the canonical basis \( (e_1, \ldots, e_n) \), so that

\[ x^0 = a^0 + e_1b^1, \quad \text{with } a^0 \in Cl_n^0, \ b^1 \in Cl_n^1, \]
where neither $a^0$ nor $b^1$ contains a summand with a factor $e_1$. Applying the first relation in (2) to $v = e_1$, we get
\[ e_1^2 a^0 + e_1^2 b^1 = a^0 e_1 + e_1 b^1 e_1. \] (3)
Now, the basis $(e_1, \ldots, e_n)$ is orthogonal w.r.t. $\Phi$, which implies that
\[ e_j e_k = -e_k e_j \quad \text{for all } j \neq k. \]
Since each monomial in $a^0$ is of even degree and contains no factor $e_1$, we get
\[ a^0 e_1 = e_1 a^0. \]
Similarly, since $b^1$ is of odd degree and contains no factor $e_1$, we get
\[ e_1 b^1 e_1 = -e_1^2 b^1. \]
But then, from (3), we get
\[ e_1^2 a^0 + e_1^2 b^1 = a^0 e_1 + e_1 b^1 e_1 = e_1 a^0 - e_1^2 b^1, \]
and so, $e_1^2 b^1 = 0$. However, $e_1^2 = -1$, and so, $b_1 = 0$. Therefore, $x_0$ contains no monomial with a factor $e_1$. We can apply the same argument to the other basis elements $e_2, \ldots, e_n$, and thus, we just proved that $x^0 \in \mathbb{R} \cdot 1$.

A similar argument applying to the second equation in (2), with $x^1 = a^1 + e_1 b^0$ and $v = e_1$ shows that $b^0 = 0$. We also conclude that $x^1 \in \mathbb{R} \cdot 1$. However, $\mathbb{R} \cdot 1 \subseteq \text{Cl}_m^0$, and so $x^1 = 0$. Finally, $x = x^0 \in (\mathbb{R} \cdot 1) \cap \Gamma_n = \mathbb{R}^* \cdot 1$.

**Remark:** If $\Phi$ is any nondegenerate quadratic form, we know (for instance, see Artin [10], Chapter 7, or Gallier [72], Chapter 6, Problem 6.14) that there is an orthogonal basis $(e_1, \ldots, e_n)$ with respect to $\varphi$ (i.e. $\varphi(e_j, e_k) = 0$ for all $j \neq k$). Thus, the commutation relations
\[ e_j^2 = \Phi(e_j) \cdot 1, \quad \text{with } \Phi(e_j) \neq 0, \quad 1 \leq j \leq n, \quad \text{and} \]
\[ e_j e_k = -e_k e_j, \quad 1 \leq j, k \leq n, \quad j \neq k \]
hold, and since the proof only rests on these facts, Lemma 30.7 holds for any nondegenerate quadratic form.

However, Lemma 30.7 may fail for degenerate quadratic forms. For example, if $\Phi \equiv 0$, then $\text{Cl}(V, 0) = \bigwedge^n V$. Consider the element $x = 1 + e_1 e_2$. Clearly, $x^{-1} = 1 - e_1 e_2$. But now, for any $v \in V$, we have
\[ \alpha(1 + e_1 e_2) v (1 + e_1 e_2)^{-1} = (1 + e_1 e_2) v (1 - e_1 e_2) = v. \]
Yet, $1 + e_1 e_2$ is not a scalar multiple of 1.
The following proposition shows that the notion of norm is well-behaved.

**Proposition 30.8.** If \( x \in \Gamma_n \), then \( N(x) \in \mathbb{R}^* \cdot 1 \).

**Proof.** The trick is to show that \( N(x) \) is in the kernel of \( \rho \). To say that \( x \in \Gamma_n \) means that \( \alpha(x)vx^{-1} \in \mathbb{R}^n \) for all \( v \in \mathbb{R}^n \).

Applying \( t \), we get
\[
(t(x)^{-1}vt(\alpha(x))) = \alpha(x)vx^{-1},
\]

since \( t \) is the identity on \( \mathbb{R}^n \). Thus, we have
\[
v = t(x)\alpha(x)v(t(\alpha(x))x)^{-1} = \alpha(\bar{x})v(\bar{x}x)^{-1},
\]

so \( \bar{x} \in \text{Ker}(\rho) \). By Proposition 30.5, we have \( \bar{x} \in \Gamma_n \), and so, \( x\bar{x} = \bar{\bar{x}}x \in \text{Ker}(\rho) \).

**Remark:** Again, the proof also holds for the Clifford group \( \Gamma(\Phi) \) associated with any nondegenerate quadratic form \( \Phi \). When \( \Phi(v) = -\|v\|^2 \), where \( \|v\| \) is the standard Euclidean norm of \( v \), we have \( N(v) = \|v\|^2 \cdot 1 \) for all \( v \in V \). However, for other quadratic forms, it is possible that \( N(x) = \lambda \cdot 1 \) where \( \lambda < 0 \), and this is a difficulty that needs to be overcome.

**Proposition 30.9.** The restriction of the norm \( N \) to \( \Gamma_n \) is a homomorphism \( N : \Gamma_n \to \mathbb{R}^* \cdot 1 \), and \( N(\alpha(x)) = N(x) \) for all \( x \in \Gamma_n \).

**Proof.** We have
\[
N(xy) = xy\bar{y}x = xN(y)\bar{x} = x\bar{x}N(y) = N(x)N(y),
\]

where the third equality holds because \( N(x) \in \mathbb{R}^* \cdot 1 \). We also have
\[
N(\alpha(x)) = \alpha(x)\alpha(\bar{x}) = \alpha(x\bar{x}) = \alpha(N(x)) = N(x).
\]

**Remark:** The proof also holds for the Clifford group \( \Gamma(\Phi) \) associated with any nondegenerate quadratic form \( \Phi \).

**Proposition 30.10.** We have \( \mathbb{R}^n - \{0\} \subseteq \Gamma_n \) and \( \rho(\Gamma_n) \subseteq O(n) \).

**Proof.** Let \( x \in \Gamma_n \) and \( v \in \mathbb{R}^n \), with \( v \neq 0 \). We have
\[
N(\rho(x)(v)) = N(\alpha(x)vx^{-1}) = N(\alpha(x))N(v)N(x^{-1}) = N(x)N(v)N(x)^{-1} = N(v),
\]

since \( N : \Gamma_n \to \mathbb{R}^* \cdot 1 \). However, for \( v \in \mathbb{R}^n \), we know that
\[
N(v) = -\Phi(v) \cdot 1.
\]

Thus, \( \rho(x) \) is norm-preserving, and so, \( \rho(x) \in O(n) \).

**Remark:** The proof that \( \rho(\Gamma(\Phi)) \subseteq O(\Phi) \) also holds for the Clifford group \( \Gamma(\Phi) \) associated with any nondegenerate quadratic form \( \Phi \). The first statement needs to be replaced by the fact that every non-isotropic vector in \( \mathbb{R}^n \) (a vector is non-isotropic if \( \Phi(x) \neq 0 \)) belongs to \( \Gamma(\Phi) \). Indeed, \( x^2 = \Phi(x) \cdot 1 \), which implies that \( x \) is invertible.

We are finally ready for the introduction of the groups \( \text{Pin}(n) \) and \( \text{Spin}(n) \).
30.4 The Groups Pin(n) and Spin(n)

Definition 30.5. We define the pinor group $\text{Pin}(n)$ as the kernel $\text{Ker} (N)$ of the homomorphism $N: \Gamma_n \to \mathbb{R}^* \cdot 1$, and the spinor group $\text{Spin}(n)$ as $\text{Pin}(n) \cap \Gamma^+_n$.

Observe that if $N(x) = 1$, then $x$ is invertible, and $x^{-1} = \overline{x}$ since $x\overline{x} = N(x) = 1$. Thus, we can write

$$\text{Pin}(n) = \{x \in \text{Cl}_n | \alpha(x)vx^{-1} \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ N(x) = 1\}$$

$$= \{x \in \text{Cl}_n | \alpha(x)v \overline{x} \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ x \overline{x} = 1\},$$

and

$$\text{Spin}(n) = \{x \in \text{Cl}_n^0 | xvx^{-1} \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ N(x) = 1\}$$

$$= \{x \in \text{Cl}_n^0 | xv \overline{x} \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ x \overline{x} = 1\}.$$

**Remark:** According to Atiyah, Bott and Shapiro, the use of the name $\text{Pin}(k)$ is a joke due to Jean-Pierre Serre (Atiyah, Bott and Shapiro [14], page 1).

**Theorem 30.11.** The restriction of $\rho: \Gamma_n \to \text{O}(n)$ to the pinor group $\text{Pin}(n)$ is a surjective homomorphism $\rho: \text{Pin}(n) \to \text{O}(n)$ whose kernel is $\{-1, 1\}$, and the restriction of $\rho$ to the spinor group $\text{Spin}(n)$ is a surjective homomorphism $\rho: \text{Spin}(n) \to \text{SO}(n)$ whose kernel is $\{-1, 1\}$.

**Proof.** By Proposition 30.10, we have a map $\rho: \text{Pin}(n) \to \text{O}(n)$. The reader can easily check that $\rho$ is a homomorphism. By the Cartan-Dieudonné theorem (see Bourbaki [25], or Gallier [72], Chapter 7, Theorem 7.2.1), every isometry $f \in \text{SO}(n)$ is the composition $f = s_1 \circ \cdots \circ s_k$ of hyperplane reflections $s_j$. If we assume that $s_j$ is a reflection about the hyperplane $H_j$ orthogonal to the nonzero vector $w_j$, by Proposition 30.6, $\rho(w_j) = s_j$. Since $N(w_j) = \|w_j\|^2 \cdot 1$, we can replace $w_j$ by $w_j/\|w_j\|$, so that $N(w_1 \cdots w_k) = 1$, and then

$$f = \rho(w_1 \cdots w_k),$$

and $\rho$ is surjective. Note that

$$\text{Ker} (\rho | \text{Pin}(n)) = \text{Ker} (\rho) \cap \text{ker}(N) = \{t \in \mathbb{R}^* \cdot 1 | N(t) = 1\} = \{-1, 1\}.$$

As to $\text{Spin}(n)$, we just need to show that the restriction of $\rho$ to $\text{Spin}(n)$ maps $\Gamma_n$ into $\text{SO}(n)$. If this was not the case, there would be some improper isometry $f \in \text{O}(n)$ so that $\rho(x) = f$, where $x \in \Gamma_n \cap \text{Cl}_n^0$. However, we can express $f$ as the composition of an odd number of reflections, say

$$f = \rho(w_1 \cdots w_{2k+1}).$$

Since

$$\rho(w_1 \cdots w_{2k+1}) = \rho(x),$$
we have $x^{-1}w_1 \cdots w_{2k+1} \in \text{Ker}(\rho)$. By Lemma 30.7, we must have

$$x^{-1}w_1 \cdots w_{2k+1} = \lambda \cdot 1$$

for some $\lambda \in \mathbb{R}^*$, and thus

$$w_1 \cdots w_{2k+1} = \lambda \cdot x,$$

where $x$ has even degree and $w_1 \cdots w_{2k+1}$ has odd degree, which is impossible.

Let us denote the set of elements $v \in \mathbb{R}^n$ with $N(v) = 1$ (with norm 1) by $S^{n-1}$. We have the following corollary of Theorem 30.11:

**Corollary 30.12.** The group $\text{Pin}(n)$ is generated by $S^{n-1}$, and every element of $\text{Spin}(n)$ can be written as the product of an even number of elements of $S^{n-1}$.

**Example 30.2.** The reader should verify that

$$\text{Pin}(1) \approx \mathbb{Z}/4\mathbb{Z}, \quad \text{Spin}(1) = \{-1, 1\} \approx \mathbb{Z}/2\mathbb{Z},$$

and also that

$$\text{Pin}(2) \approx \{ae_1 + be_2 \mid a^2 + b^2 = 1\} \cup \{c1 + de_1e_2 \mid c^2 + d^2 = 1\}, \quad \text{Spin}(2) = \text{U}(1).$$

We may also write $\text{Pin}(2) = \text{U}(1) + \text{U}(1)$, where $\text{U}(1)$ is the group of complex numbers of modulus 1 (the unit circle in $\mathbb{R}^2$). It can also be shown that $\text{Spin}(3) \approx \text{SU}(2)$ and $\text{Spin}(4) \approx \text{SU}(2) \times \text{SU}(2)$. The group $\text{Spin}(5)$ is isomorphic to the symplectic group $\text{Sp}(2)$, and $\text{Spin}(6)$ is isomorphic to $\text{SU}(4)$ (see Curtis [46] or Porteous [143]).

Let us take a closer look at $\text{Spin}(2)$. The Clifford algebra $\text{Cl}_2$ is generated by the four elements

$$1, \ e_1, \ e_2, \ e_1e_2,$$

and they satisfy the relations

$$e_1^2 = -1, \ e_2^2 = -1, \ e_1e_2 = -e_2e_1.$$

The group $\text{Spin}(2)$ consists of all products

$$\prod_{i=1}^{2k} (a_ie_1 + b_ie_2)$$

consisting of an even number of factors and such that $a_i^2 + b_i^2 = 1$. In view of the above relations, every such element can be written as

$$x = a_1 + be_1e_2,$$
where $x$ satisfies the conditions that $xvx^{-1} \in \mathbb{R}^2$ for all $v \in \mathbb{R}^2$, and $N(x) = 1$. Since

$$X = a1 - be_1e_2,$$

we get

$$N(x) = a^2 + b^2,$$

and the condition $N(x) = 1$ is simply $a^2 + b^2 = 1$.

We claim that if $x \in \text{Cl}_{0}^2$, then $xvx^{-1} \in \mathbb{R}^2$. Indeed, since $x \in \text{Cl}_{0}^2$ and $v \in \text{Cl}_{1}^2$, we have $xvx^{-1} \in \text{Cl}_{1}^2$, which implies that $xvx^{-1} \in \mathbb{R}^2$, since the only elements of $	ext{Cl}_{1}^2$ are those in $\mathbb{R}^2$. Then, $\text{Spin}(2)$ consists of those elements $x = a1 + be_1e_2$ so that $a^2 + b^2 = 1$. If we let $i = e_1e_2$, we observe that

$$i^2 = -1,$$
$$e_1i = -ie_1 = -e_2,$$
$$e_2i = -ie_2 = e_1.$$

Thus, $\text{Spin}(2)$ is isomorphic to $U(1)$. Also note that

$$e_1(a1 + bi) = (a1 - bi)e_1.$$

Let us find out explicitly what is the action of $\text{Spin}(2)$ on $\mathbb{R}^2$. Given $X = a1 + bi$, with $a^2 + b^2 = 1$, for any $v = v_1e_1 + v_2e_2$, we have

$$\alpha(X)vX^{-1} = X(v_1e_1 + v_2e_2)X^{-1} = X(v_1e_1 + v_2e_2)(-e_1e_1)X = X(v_1e_1 + v_2e_2)(-1)(e_1X) = X(v_11 + v_2i)Xe_1 = X^2(v_11 + v_2i)e_1 = ((a^2 - b^2)v_1 - 2abv_2)1 + (a^2 - b^2)v_2 + 2abv_1)i)X = ((a^2 - b^2)v_1 - 2abv_2)e_1 + (a^2 - b^2)v_2 + 2abv_1)e_2.$$

Since $a^2 + b^2 = 1$, we can write $X = a1 + bi = (\cos \theta)1 + (\sin \theta)i$, and the above derivation shows that

$$\alpha(X)vX^{-1} = (\cos 2\theta v_1 - \sin 2\theta v_2)e_1 + (\cos 2\theta v_2 + \sin 2\theta v_1)e_2.$$

This means that the rotation $\rho_X$ induced by $X \in \text{Spin}(2)$ is the rotation of angle $2\theta$ around the origin. Observe that the maps

$$v \mapsto v(-e_1), \quad X \mapsto Xe_1$$

establish bijections between $\mathbb{R}^2$ and $\text{Spin}(2) \simeq U(1)$. Also, note that the action of $X = \cos \theta + i\sin \theta$ viewed as a complex number yields the rotation of angle $\theta$, whereas the action
of $X = (\cos \theta)1 + (\sin \theta)i$ viewed as a member of $\text{Spin}(2)$ yields the rotation of angle $2\theta$. There is nothing wrong. In general, $\text{Spin}(n)$ is a two-to-one cover of $\text{SO}(n)$.

Next, let us take a closer look at $\text{Spin}(3)$. The Clifford algebra $\text{Cl}_3$ is generated by the eight elements 

$$1, e_1, e_2, , e_3, , e_1e_2, e_2e_3, e_3e_1, e_1e_2e_3,$$

and they satisfy the relations

$$e_i^2 = -1, \quad e_je_j = -e_ie_i, \quad 1 \leq i, j \leq 3, \quad i \neq j.$$

The group $\text{Spin}(3)$ consists of all products

$$\prod_{i=1}^{2k} (a_ie_1 + b_ie_2 + c_ie_3)$$

consisting of an even number of factors and such that $a_i^2 + b_i^2 + c_i^2 = 1$. In view of the above relations, every such element can be written as

$$x = a_1 + be_2e_3 + ce_3e_1 + de_1e_2,$$

where $x$ satisfies the conditions that $xvx^{-1} \in \mathbb{R}^3$ for all $v \in \mathbb{R}^3$, and $N(x) = 1$. Since

$$\overline{x} = a_1 - be_2e_3 - ce_3e_1 - de_1e_2,$$

we get

$$N(x) = a^2 + b^2 + c^2 + d^2,$$

and the condition $N(x) = 1$ is simply $a^2 + b^2 + c^2 + d^2 = 1$.

It turns out that the conditions $x \in \text{Cl}_3^0$ and $N(x) = 1$ imply that $xvx^{-1} \in \mathbb{R}^3$ for all $v \in \mathbb{R}^3$. To prove this, first observe that $N(x) = 1$ implies that $x^{-1} = \pm \overline{x}$, and that $\overline{v} = -v$ for any $v \in \mathbb{R}^3$, and so,

$$\overline{xvx^{-1}} = -xvx^{-1}.$$ 

Also, since $x \in \text{Cl}_3^0$ and $v \in \text{Cl}_3^1$, we have $xvx^{-1} \in \text{Cl}_3^1$. Thus, we can write

$$xvx^{-1} = u + \lambda e_1e_2e_3,$$

for some $u \in \mathbb{R}^3$ and some $\lambda \in \mathbb{R}$.

But

$$\overline{e_1e_2e_3} = -e_3e_2e_1 = e_1e_2e_3,$$

and so,

$$\overline{xvx^{-1}} = -u + \lambda e_1e_2e_3 = -xvx^{-1} = -u - \lambda e_1e_2e_3,$$

which implies that $\lambda = 0$. Thus, $xvx^{-1} \in \mathbb{R}^3$, as claimed. Then, $\text{Spin}(3)$ consists of those elements $x = a_1 + be_2e_3 + ce_3e_1 + de_1e_2$ so that $a^2 + b^2 + c^2 + d^2 = 1$. Under the bijection

$$i \mapsto e_2e_3, \quad j \mapsto e_3e_1, \quad k \mapsto e_1e_2,$$
we can check that we have an isomorphism between the group $\mathbf{SU}(2)$ of unit quaternions and $\mathbf{Spin}(3)$. If $X = a1 + be_2e_3 + ce_3e_1 + de_1e_2 \in \mathbf{Spin}(3)$, observe that

$$X^{-1} = X = a1 - be_2e_3 - ce_3e_1 - de_1e_2.$$ 

Now, using the identification

$$i \mapsto e_2e_3, \quad j \mapsto e_3e_1, \quad k \mapsto e_1e_2,$$

we can easily check that

$$(e_1e_2e_3)^2 = 1,$$

$$(e_1e_2e_3)i = i(e_1e_2e_3) = -e_1,$$

$$(e_1e_2e_3)j = j(e_1e_2e_3) = -e_2,$$

$$(e_1e_2e_3)k = k(e_1e_2e_3) = -e_3,$$

$$(e_1e_2e_3)e_1 = -i,$$

$$(e_1e_2e_3)e_2 = -j,$$

$$(e_1e_2e_3)e_3 = -k.$$ 

Then, if $X = a1 + bi + cj + dk \in \mathbf{Spin}(3)$, for every $v = v_1e_1 + v_2e_2 + v_3e_3$, we have

$$\alpha(X)vX^{-1} = X(v_1e_1 + v_2e_2 + v_3e_3)X^{-1} = X(e_1e_2e_3)^2(v_1e_1 + v_2e_2 + v_3e_3)X^{-1} = (e_1e_2e_3)X(e_1e_2e_3)(v_1e_1 + v_2e_2 + v_3e_3)X^{-1} = -(e_1e_2e_3)X(v_1i + v_2j + v_3k)X^{-1}.$$ 

This shows that the rotation $\rho_X \in \mathbf{SO}(3)$ induced by $X \in \mathbf{Spin}(3)$ can be viewed as the rotation induced by the quaternion $a1 + bi + cj + dk$ on the pure quaternions, using the maps

$$v \mapsto -(e_1e_2e_3)v, \quad X \mapsto -(e_1e_2e_3)X$$

to go from a vector $v = v_1e_1 + v_2e_2 + v_3e_3$ to the pure quaternion $v_1i + v_2j + v_3k$, and back.

We close this section by taking a closer look at $\mathbf{Spin}(4)$. The group $\mathbf{Spin}(4)$ consists of all products

$$\prod_{i=1}^{2k}(a_ie_1 + b_ie_2 + c_ie_3 + d_ie_4)$$

consisting of an even number of factors and such that $a_i^2 + b_i^2 + c_i^2 + d_i^2 = 1$. Using the relations

$$e_i^2 = -1, \quad e_je_j = -e_je_i, \quad 1 \leq i, j \leq 4, \quad i \neq j,$$

every element of $\mathbf{Spin}(4)$ can be written as

$$x = a_11 + a_2e_1e_2 + a_3e_2e_3 + a_4e_3e_1 + a_5e_4e_3 + a_6e_4e_1 + a_7e_4e_2 + a_8e_1e_2e_3e_4,$$
where $x$ satisfies the conditions that $xvx^{-1} \in \mathbb{R}^4$ for all $v \in \mathbb{R}^4$, and $N(x) = 1$. Let

$$i = e_1e_2, \quad j = e_2e_3, \quad k = 3e_1, \quad i' = e_4e_3, \quad j' = e_4e_1, \quad k' = e_4e_2,$$

and $I = e_1e_2e_3e_4$. The reader will easily verify that

$$ij = k, \quad jk = i, \quad ki = j,$$
$$i^2 = -1, \quad j^2 = -1, \quad k^2 = -1,$$
$$iI = Ii = i', \quad jI = Ij = j', \quad kI = Ik = k', \quad I^2 = 1, \quad I = I.$$

Then, every $x \in \text{Spin}(4)$ can be written as

$$x = u + Iv, \quad \text{with} \quad u = a1 + bi + cj + dk \quad \text{and} \quad v = a'1 + b'i + c'j + d'k,$$

with the extra conditions stated above. Using the above identities, we have

$$(u + Iv)(u' + Iv') = uu' + vv' + I(uv' + vu').$$

As a consequence,

$$N(u + Iv) = (u + Iv)(u + Iv) = uu + vv + I(uv + vu),$$

and thus, $N(u + Iv) = 1$ is equivalent to

$$uu + vv = 1 \quad \text{and} \quad uv + vu = 0.$$

As in the case $n = 3$, it turns out that the conditions $x \in \text{Cl}^0_4$ and $N(x) = 1$ imply that $xvx^{-1} \in \mathbb{R}^4$ for all $v \in \mathbb{R}^4$. The only change to the proof is that $xvx^{-1} \in \text{Cl}^1_4$ can be written as

$$xvx^{-1} = u + \sum_{i,j,k} \lambda_{i,j,k}e_ie_je_k, \quad \text{for some} \quad u \in \mathbb{R}^4, \quad \text{with} \quad \{i, j, k\} \subseteq \{1, 2, 3, 4\}.$$ 

As in the previous proof, we get $\lambda_{i,j,k} = 0$. Then, $\text{Spin}(4)$ consists of those elements $u + Iv$ so that

$$uu + vv = 1 \quad \text{and} \quad uv + vu = 0,$$

with $u$ and $v$ of the form $a1 + bi + cj + dk$. Finally, we see that $\text{Spin}(4)$ is isomorphic to $\text{Spin}(3) \times \text{Spin}(3)$ under the isomorphism

$$u + vI \mapsto (u + v, u - v).$$
Indeed, we have
\[ N(u + v) = (u + v)(\overline{u} + \overline{v}) = 1, \]
and
\[ N(u - v) = (u - v)(\overline{u} - \overline{v}) = 1, \]
since
\[ u\overline{u} + v\overline{v} = 1 \quad \text{and} \quad u\overline{v} + v\overline{u} = 0, \]
and
\[ (u + v, u - v)(u' + v', u' - v') = (uu' + vv' + uv' + vu', uu' + vv' - (uv' + vu')). \]

Remark: It can be shown that the assertion if \( x \in \text{Cl}_n^0 \) and \( N(x) = 1 \), then \( xvX^{-1} \in \mathbb{R}^n \) for all \( v \in \mathbb{R}^n \), is true up to \( n = 5 \) (see Porteous [143], Chapter 13, Proposition 13.58). However, this is already false for \( n = 6 \). For example, if \( X = 1/\sqrt{2}(1 + e_1e_2e_3e_4e_5e_6) \), it is easy to see that \( N(X) = 1 \), and yet, \( Xe_1X^{-1} \notin \mathbb{R}^6 \).

30.5 The Groups \( \text{Pin}(p, q) \) and \( \text{Spin}(p, q) \)

For every nondegenerate quadratic form \( \Phi \) over \( \mathbb{R} \), there is an orthogonal basis with respect to which \( \Phi \) is given by
\[ \Phi(x_1, \ldots, x_{p+q}) = x_1^2 + \cdots + x_p^2 - (x_{p+1}^2 + \cdots + x_{p+q}^2), \]
where \( p \) and \( q \) only depend on \( \Phi \). The quadratic form corresponding to \( (p, q) \) is denoted \( \Phi_{p,q} \) and we call \( (p, q) \) the signature of \( \Phi_{p,q} \). Let \( n = p + q \). We define the group \( \mathbf{O}(p, q) \) as the group of isometries w.r.t. \( \Phi_{p,q} \), i.e., the group of linear maps \( f \) so that
\[ \Phi_{p,q}(f(v)) = \Phi_{p,q}(v) \quad \text{for all} \quad v \in \mathbb{R}^n \]
and the group \( \mathbf{SO}(p, q) \) as the subgroup of \( \mathbf{O}(p, q) \) consisting of the isometries \( f \in \mathbf{O}(p, q) \) with \( \det(f) = 1 \). We denote the Clifford algebra \( \text{Cl}(\Phi_{p,q}) \), where \( \Phi_{p,q} \) has signature \( (p, q) \) by \( \text{Cl}_{p,q} \), the corresponding Clifford group by \( \Gamma_{p,q} \), and the special Clifford group \( \Gamma_{p,q} \cap \text{Cl}_{0,n}^0 \) by \( \Gamma_{p,q}^+ \). Note that with this new notation, \( \text{Cl}_n = \text{Cl}_{0,n} \).

As we mentioned earlier, since Lawson and Michelsohn [115] adopt the opposite of our sign convention in defining Clifford algebras; their \( \text{Cl}(p, q) \) is our \( \text{Cl}(q, p) \).

As we mentioned in Section 30.3, we have the problem that \( N(v) = -\Phi(v) \cdot 1 \), but \( -\Phi(v) \) is not necessarily positive (where \( v \in \mathbb{R}^n \)). The fix is simple: Allow elements \( x \in \Gamma_{p,q} \) with \( N(x) = \pm 1 \).

Definition 30.6. We define the pinor group \( \text{Pin}(p, q) \) as the group
\[ \text{Pin}(p, q) = \{ x \in \Gamma_{p,q} \mid N(x) = \pm 1 \}, \]
and the spinor group \( \text{Spin}(p, q) \) as \( \text{Pin}(p, q) \cap \Gamma_{p,q}^+ \).
Remarks:

(1) It is easily checked that the group \( \text{Spin}(p, q) \) is also given by
\[
\text{Spin}(p, q) = \{ x \in \text{Cl}_{p,q}^0 \mid xv\overline{x} \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ N(x) = 1 \}.
\]
This is because \( \text{Spin}(p, q) \) consists of elements of even degree.

(2) One can check that if \( N(x) \neq 0 \), then
\[
\alpha(x)vx^{-1} = xvt(x)/N(x).
\]
Thus, we have
\[
\text{Pin}(p, q) = \{ x \in \text{Cl}_{p,q} \mid xv(x)N(x) \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ N(x) = \pm 1 \}.
\]
When \( \Phi(x) = -\|x\|^2 \), we have \( N(x) = \|x\|^2 \), and
\[
\text{Pin}(n) = \{ x \in \text{Cl}_n \mid xv(x) \in \mathbb{R}^n \text{ for all } v \in \mathbb{R}^n, \ N(x) = 1 \}.
\]

Theorem 30.11 generalizes as follows:

\textbf{Theorem 30.13.} The restriction of \( \rho: \Gamma_{p,q} \to \text{GL}(n) \) to the pinor group \( \text{Pin}(p, q) \) is a surjective homomorphism \( \rho: \text{Pin}(p, q) \to \text{O}(p, q) \) whose kernel is \( \{-1, 1\} \), and the restriction of \( \rho \) to the spinor group \( \text{Spin}(p, q) \) is a surjective homomorphism \( \rho: \text{Spin}(p, q) \to \text{SO}(p, q) \) whose kernel is \( \{-1, 1\} \).

\textbf{Proof.} The Cartan-Dieudonné also holds for any nondegenerate quadratic form \( \Phi \), in the sense that every isometry in \( \text{O}(\Phi) \) is the composition of reflections defined by hyperplanes orthogonal to non-isotropic vectors (see Dieudonné [50], Chevalley [42], Bourbaki [25], or Gallier [72], Chapter 7, Problem 7.14). Thus, Theorem 30.11 also holds for any nondegenerate quadratic form \( \Phi \). The only change to the proof is the following: Since \( N(w_j) = -\Phi(w_j) \cdot 1 \), we can replace \( w_j \) by \( w_j/\sqrt{\Phi(w_j)} \), so that \( N(w_1 \cdots w_k) = \pm 1 \), and then
\[
f = \rho(w_1 \cdots w_k),
\]
and \( \rho \) is surjective.

If we consider \( \mathbb{R}^n \) equipped with the quadratic form \( \Phi_{p,q} \) (with \( n = p + q \)), we denote the set of elements \( v \in \mathbb{R}^n \) with \( N(v) = 1 \) by \( S_{p,q}^{n-1} \). We have the following corollary of Theorem 30.13 (generalizing Corollary 30.14):

\textbf{Corollary 30.14.} The group \( \text{Pin}(p, q) \) is generated by \( S_{p,q}^{n-1} \), and every element of \( \text{Spin}(p, q) \) can be written as the product of an even number of elements of \( S_{p,q}^{n-1} \).
Example 30.3. The reader should check that 
\[ \mathcal{C}l_{0,1} \approx \mathbb{C}, \quad \mathcal{C}l_{1,0} \approx \mathbb{R} \oplus \mathbb{R}. \]

We also have 
\[ \text{Pin}(0, 1) \approx \mathbb{Z}/4\mathbb{Z}, \quad \text{Pin}(1, 0) \approx \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, \]
from which we get \( \mathcal{S}pin(0, 1) = \mathcal{S}pin(1, 0) \approx \mathbb{Z}/2\mathbb{Z} \). Also, show that \( \text{Pin}(2, 0), \text{Pin}(1, 1), \text{and Pin}(0, 2); \) see Choquet-Bruhat [43], Chapter I, Section 7, page 26. Show that 
\[ \mathcal{S}pin(0, 2) = \mathcal{S}pin(2, 0) \approx U(1), \]
and 
\[ \mathcal{S}pin(1, 1) = \{ a1 + be_1e_2 | a^2 - b^2 = 1 \}. \]

Observe that \( \mathcal{S}pin(1, 1) \) is not connected.

More generally, it can be shown that \( \mathcal{C}l_{p,q}^0 \) and \( \mathcal{C}l_{q,p}^0 \) are isomorphic, from which it follows that \( \mathcal{S}pin(p, q) \) and \( \mathcal{S}pin(q, p) \) are isomorphic, but \( \text{Pin}(p, q) \) and \( \text{Pin}(q, p) \) are not isomorphic in general, and in particular, \( \text{Pin}(p, 0) \) and \( \text{Pin}(0, p) \) are not isomorphic in general (see Choquet-Bruhat [43], Chapter I, Section 7). However, due to the “8-periodicity” of the Clifford algebras (to be discussed in the next section), it follows that \( \mathcal{C}l_{p,q} \) and \( \mathcal{C}l_{q,p} \) are isomorphic when \( |p - q| = 0 \mod 4 \).

### 30.6 Periodicity of the Clifford Algebras \( \mathcal{C}l_{p,q} \)

It turns out that the real algebras \( \mathcal{C}l_{p,q} \) can be build up as tensor products of the basic algebras \( \mathbb{R}, \mathbb{C}, \) and \( \mathbb{H} \). As pointed out by Lounesto (Section 23.16 [118]), the description of the real algebras \( \mathcal{C}l_{p,q} \) as matrix algebras and the 8-periodicity was first observed by Elie Cartan in 1908; see Cartan’s article, \textit{Nombres Complexes}, based on the original article in German by E. Study, in Molk [131], article I-5 (fasc. 3), pages 329-468. These algebras are defined in Section 36 under the name “Systems of Clifford and Lipschitz,” page 463-466. Of course, Cartan used a very different notation; see page 464 in the article cited above. These facts were rediscovered independently by Raoul Bott in the 1960’s (see Raoul Bott’s comments in Volume 2 of his Collected papers.).

We will use the notation \( \mathbb{R}(n) \) (resp. \( \mathbb{C}(n) \)) for the algebra \( M_n(\mathbb{R}) \) of all \( n \times n \) real matrices (resp. the algebra \( M_n(\mathbb{C}) \) of all \( n \times n \) complex matrices). As mentioned in Example 30.3, it is not hard to show that 
\[ \begin{align*}
\mathcal{C}l_{0,1} &= \mathbb{C} \quad \mathcal{C}l_{1,0} = \mathbb{R} \oplus \mathbb{R} \\
\mathcal{C}l_{0,2} &= \mathbb{H} \quad \mathcal{C}l_{2,0} = \mathbb{R}(2),
\end{align*} \]
and
\[ \text{Cl}_{1,1} = \mathbb{R}(2). \]
The key to the classification is the following lemma:

**Lemma 30.15.** We have the isomorphisms

\[
\begin{align*}
\text{Cl}_{0,n+2} & \cong \text{Cl}_{n,0} \otimes \text{Cl}_{0,2} \\
\text{Cl}_{n+2,0} & \cong \text{Cl}_{0,n} \otimes \text{Cl}_{2,0} \\
\text{Cl}_{p+1,q+1} & \cong \text{Cl}_{p,q} \otimes \text{Cl}_{1,1},
\end{align*}
\]
for all \( n, p, q \geq 0 \).

**Proof.** Let \( \Phi_{0,n}(x) = -\|x\|^2 \), where \( \|x\| \) is the standard Euclidean norm on \( \mathbb{R}^{n+2} \), and let \((e_1, \ldots, e_{n+2})\) be an orthonormal basis for \( \mathbb{R}^{n+2} \) under the standard Euclidean inner product. We also let \((e'_1, \ldots, e'_n)\) be a set of generators for \( \text{Cl}_{n,0} \) and \((e''_1, e''_2)\) be a set of generators for \( \text{Cl}_{0,2} \). We can define a linear map \( f : \mathbb{R}^{n+2} \rightarrow \text{Cl}_{0,0} \otimes \text{Cl}_{0,2} \) by its action on the basis \((e_1, \ldots, e_{n+2})\) as follows:

\[
f(e_i) = \begin{cases} 
  e'_i \otimes e''_2 & \text{for } 1 \leq i \leq n \\
  1 \otimes e''_{i-n} & \text{for } n + 1 \leq i \leq n + 2.
\end{cases}
\]

Observe that for \( 1 \leq i, j \leq n \), we have

\[
f(e_i)f(e_j) + f(e_j)f(e_i) = (e'_i e'_j + e'_j e'_i) \otimes (e''_1 e''_2)^2 = -2\delta_{ij} 1 \otimes 1,
\]

since \( e'_i e'_j = -e'_j e'_i \), \( (e''_1)^2 = -1 \), and \( (e''_2)^2 = -1 \), and \( e'_i e'_j = -e'_j e'_i \), for all \( i \neq j \), and \( (e'_i)^2 = 1 \), for all \( i \) with \( 1 \leq i \leq n \). Also, for \( n + 1 \leq i, j \leq n + 2 \), we have

\[
f(e_i)f(e_j) + f(e_j)f(e_i) = 1 \otimes (e''_{i-n} e''_{j-n} + e''_{j-n} e''_{i-n}) = -2\delta_{ij} 1 \otimes 1,
\]

and

\[
f(e_i)f(e_k) + f(e_k)f(e_i) = 2e'_i \otimes (e''_1 e''_{n-k} + e''_{n-k} e''_1 e''_2) = 0,
\]

for \( 1 \leq i, j \leq n \) and \( n + 1 \leq k \leq n + 2 \) (since \( e''_{n-k} = e''_1 \) or \( e''_{n-k} = e''_2 \)). Thus, we have

\[
f(x)^2 = -\|x\|^2 \cdot 1 \otimes 1 \quad \text{for all } x \in \mathbb{R}^{n+2},
\]

and by the universal mapping property of \( \text{Cl}_{0,n+2} \), we get an algebra map

\[
\tilde{f} : \text{Cl}_{0,n+2} \rightarrow \text{Cl}_{n,0} \otimes \text{Cl}_{0,2}.
\]

Since \( \tilde{f} \) maps onto a set of generators, it is surjective. However

\[
\dim(\text{Cl}_{0,n+2}) = 2^{n+2} = 2^n \cdot 2 = \dim(\text{Cl}_{n,0})\dim(\text{Cl}_{0,2}) = \dim(\text{Cl}_{n,0} \otimes \text{Cl}_{0,2}),
\]

and \( \tilde{f} \) is an isomorphism.
The proof of the second identity is analogous. For the third identity, we have

\[ \Phi_{p,q}(x_1, \ldots, x_{p+q}) = x_1^2 + \cdots + x_p^2 - (x_{p+1}^2 + \cdots + x_{p+q}^2), \]

and let \((e_1, \ldots, e_{p+1}, e_1, \ldots, e_{q+1})\) be an orthogonal basis for \(\mathbb{R}^{p+q+2}\) so that \(\Phi_{p+1,q+1}(e_i) = +1\) and \(\Phi_{p+1,q+1}(e_j) = -1\) for \(i = 1, \ldots, p+1\) and \(j = 1, \ldots, q+1\). Also, let \((e'_1, \ldots, e'_p, e'_1, \ldots, e'_q)\) be a set of generators for \(\text{Cl}_{p,q}\) and \((e''_1, e''_1)\) be a set of generators for \(\text{Cl}_{1,1}\). We define a linear map \(f: \mathbb{R}^{p+q+2} \to \text{Cl}_{p,q} \otimes \text{Cl}_{1,1}\) by its action on the basis as follows:

\[
f(e_i) = \begin{cases} 
  e'_i \otimes e''_1 & \text{for } 1 \leq i \leq p \\
  1 \otimes e''_1 & \text{for } i = p+1,
\end{cases}
\]

and

\[
f(\epsilon_j) = \begin{cases} 
  e'_j \otimes e''_1 & \text{for } 1 \leq j \leq q \\
  1 \otimes e''_1 & \text{for } j = q+1.
\end{cases}
\]

We can check that

\[ f(x)^2 = \Phi_{p+1,q+1}(x) \cdot 1 \otimes 1 \quad \text{for all } x \in \mathbb{R}^{p+q+2}, \]

and we finish the proof as in the first case.

To apply this lemma, we need some further isomorphisms among various matrix algebras.

**Proposition 30.16.** The following isomorphisms hold:

\[
\begin{align*}
\mathbb{R}(m) \otimes \mathbb{R}(n) & \approx \mathbb{R}(mn) \quad \text{for all } m, n \geq 0 \\
\mathbb{R}(n) \otimes_{\mathbb{R}} K & \approx K(n) \quad \text{for } K = \mathbb{C} \text{ or } K = \mathbb{H} \text{ and all } n \geq 0 \\
\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C} & \approx \mathbb{C} \oplus \mathbb{C} \\
\mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} & \approx \mathbb{C}(2) \\
\mathbb{H} \otimes_{\mathbb{R}} \mathbb{H} & \approx \mathbb{R}(4).
\end{align*}
\]

**Proof.** Details can be found in Lawson and Michelsohn [115]. The first two isomorphisms are quite obvious. The third isomorphism \(\mathbb{C} \oplus \mathbb{C} \to \mathbb{C} \otimes \mathbb{C}\) is obtained by sending

\[ (1,0) \mapsto \frac{1}{2}(1 \otimes 1 + i \otimes i), \quad (0,1) \mapsto \frac{1}{2}(1 \otimes 1 - i \otimes i). \]

The field \(\mathbb{C}\) is isomorphic to the subring of \(\mathbb{H}\) generated by \(i\). Thus, we can view \(\mathbb{H}\) as a \(\mathbb{C}\)-vector space under left scalar multiplication. Consider the \(\mathbb{R}\)-bilinear map

\[ \pi: \mathbb{C} \times \mathbb{H} \to \text{Hom}_\mathbb{C}(\mathbb{H}, \mathbb{H}) \]

given by

\[ \pi_{y,z}(x) = yx \overline{z}, \]

where \(y \in \mathbb{C}\) and \(x, z \in \mathbb{H}\). Thus, we get an \(\mathbb{R}\)-linear map \(\pi: \mathbb{C} \otimes_{\mathbb{R}} \mathbb{H} \to \text{Hom}_\mathbb{C}(\mathbb{H}, \mathbb{H})\). However, we have \(\text{Hom}_\mathbb{C}(\mathbb{H}, \mathbb{H}) \approx \mathbb{C}(2)\). Furthermore, since

\[ \pi_{y,z} \circ \pi_{y',z'} = \pi_{yy',zz'}, \]
the map $\pi$ is an algebra homomorphism

$$\pi: \mathbb{C} \times \mathbb{H} \to \mathbb{C}(2).$$

We can check on a basis that $\pi$ is injective, and since

$$\dim_{\mathbb{R}}(\mathbb{C} \times \mathbb{H}) = \dim_{\mathbb{R}}(\mathbb{C}(2)) = 8,$$

the map $\pi$ is an isomorphism. The last isomorphism is proved in a similar fashion.

We now have the main periodicity theorem.

**Theorem 30.17. (Cartan/Bott)** For all $n \geq 0$, we have the following isomorphisms:

$$\begin{align*}
\text{Cl}_{0,n+8} & \approx \text{Cl}_{0,n} \otimes \text{Cl}_{0,8} \\
\text{Cl}_{n+8,0} & \approx \text{Cl}_{n,0} \otimes \text{Cl}_{8,0}.
\end{align*}$$

Furthermore,

$$\text{Cl}_{0,8} = \text{Cl}_{8,0} = \mathbb{R}(16).$$

**Proof.** By Lemma 30.15 we have the isomorphisms

$$\begin{align*}
\text{Cl}_{0,n+2} & \approx \text{Cl}_{n,0} \otimes \text{Cl}_{0,2} \\
\text{Cl}_{n+2,0} & \approx \text{Cl}_{0,n} \otimes \text{Cl}_{2,0},
\end{align*}$$

and thus,

$$\text{Cl}_{0,n+8} \approx \text{Cl}_{n+6,0} \otimes \text{Cl}_{0,2} \approx \text{Cl}_{0,n+4} \otimes \text{Cl}_{2,0} \otimes \text{Cl}_{0,2} \approx \cdots \approx \text{Cl}_{0,n} \otimes \text{Cl}_{2,0} \otimes \text{Cl}_{0,2} \otimes \text{Cl}_{2,0} \otimes \text{Cl}_{0,2}.$$

Since $\text{Cl}_{0,2} = \mathbb{H}$ and $\text{Cl}_{2,0} = \mathbb{R}(2)$, by Proposition 30.16, we get

$$\text{Cl}_{2,0} \otimes \text{Cl}_{0,2} \otimes \text{Cl}_{2,0} \otimes \text{Cl}_{0,2} \approx \mathbb{H} \otimes \mathbb{H} \otimes \mathbb{R}(2) \otimes \mathbb{R}(2) \approx \mathbb{R}(4) \otimes \mathbb{R}(4) \approx \mathbb{R}(16).$$

The second isomorphism is proved in a similar fashion.

From all this, we can deduce the following table:

<table>
<thead>
<tr>
<th>$n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Cl}_{0,n}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{C}$</td>
<td>$\mathbb{H}$</td>
<td>$\mathbb{H} \oplus \mathbb{H}$</td>
<td>$\mathbb{H}(2)$</td>
<td>$\mathbb{C}(4)$</td>
<td>$\mathbb{R}(8)$</td>
<td>$\mathbb{R}(8) \oplus \mathbb{R}(8)$</td>
<td>$\mathbb{R}(16)$</td>
</tr>
<tr>
<td>$\text{Cl}_{n,0}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R} \oplus \mathbb{R}$</td>
<td>$\mathbb{R}(2)$</td>
<td>$\mathbb{C}(2)$</td>
<td>$\mathbb{H}(2)$</td>
<td>$\mathbb{H}(2) \oplus \mathbb{H}(2)$</td>
<td>$\mathbb{H}(4)$</td>
<td>$\mathbb{C}(8)$</td>
<td>$\mathbb{R}(16)$</td>
</tr>
</tbody>
</table>

A table of the Clifford groups $\text{Cl}_{p,q}$ for $0 \leq p, q \leq 7$ can be found in Kirillov [101], and for $0 \leq p, q \leq 8$, in Lawson and Michelsohn [115] (but beware that their $\text{Cl}_{p,q}$ is our $\text{Cl}_{q,p}$). It can also be shown that

$$\begin{align*}
\text{Cl}_{p+1,q} & \approx \text{Cl}_{q+1,p} \\
\text{Cl}_{p,q} & \approx \text{Cl}_{p-4,q+4}.
\end{align*}$$
with \( p \geq 4 \) in the second identity (see Lounesto [118], Chapter 16, Sections 16.3 and 16.4). Using the second identity, if \( |p-q| = 4k \), it is easily shown by induction on \( k \) that \( \text{Cl}_{p,q} \approx \text{Cl}_{q,p} \), as claimed in the previous section.

We also have the isomorphisms

\[
\text{Cl}_{p,q} \approx \text{Cl}_{p,q+1}^0,
\]

from which it follows that

\[
\text{Spin}(p,q) \approx \text{Spin}(q,p)
\]

(see Choquet-Bruhat [43], Chapter I, Sections 4 and 7). However, in general, \( \text{Pin}(p,q) \) and \( \text{Pin}(q,p) \) are not isomorphic. In fact, \( \text{Pin}(0,n) \) and \( \text{Pin}(n,0) \) are not isomorphic if \( n \neq 4k \), with \( k \in \mathbb{N} \) (see Choquet-Bruhat [43], Chapter I, Section 7, page 27).

### 30.7 The Complex Clifford Algebras \( \text{Cl}(n, \mathbb{C}) \)

One can also consider Clifford algebras over the complex field \( \mathbb{C} \). In this case, it is well-known that every nondegenerate quadratic form can be expressed by

\[
\Phi_n^C(x_1, \ldots, x_n) = x_1^2 + \cdots + x_n^2
\]

in some orthonormal basis. Also, it is easily shown that the complexification \( \mathbb{C} \otimes \mathbb{R} \text{Cl}_{p,q} \) of the real Clifford algebra \( \text{Cl}_{p,q} \) is isomorphic to \( \text{Cl}(\Phi_n^C) \). Thus, all these complex algebras are isomorphic for \( p + q = n \), and we denote them by \( \text{Cl}(n, \mathbb{C}) \). Theorem 30.15 yields the following periodicity theorem:

**Theorem 30.18.** The following isomorphisms hold:

\[
\text{Cl}(n+2, \mathbb{C}) \approx \text{Cl}(n, \mathbb{C}) \otimes \mathbb{C} \text{Cl}(2, \mathbb{C}),
\]

with \( \text{Cl}(2, \mathbb{C}) = \mathbb{C}(2) \).

**Proof.** Since \( \text{Cl}(n, \mathbb{C}) = \mathbb{C} \otimes \mathbb{R} \text{Cl}_{0,n} = \mathbb{C} \otimes \mathbb{R} \text{Cl}_{n,0} \), by Lemma 30.15, we have

\[
\text{Cl}(n+2, \mathbb{C}) = \mathbb{C} \otimes \mathbb{R} \text{Cl}_{0,n+2} \approx \mathbb{C} \otimes \mathbb{R} (\text{Cl}_{n,0} \otimes \mathbb{R} \text{Cl}_{0,2}) \approx (\mathbb{C} \otimes \mathbb{R} \text{Cl}_{n,0}) \otimes \mathbb{C} (\mathbb{C} \otimes \mathbb{R} \text{Cl}_{0,2}).
\]

However, \( \text{Cl}_{0,2} = \mathbb{H} \), \( \text{Cl}(n, \mathbb{C}) = \mathbb{C} \otimes \mathbb{R} \text{Cl}_{n,0} \), and \( \mathbb{C} \otimes \mathbb{R} \mathbb{H} \approx \mathbb{C}(2) \), so we get \( \text{Cl}(2, \mathbb{C}) = \mathbb{C}(2) \) and

\[
\text{Cl}(n+2, \mathbb{C}) \approx \text{Cl}(n, \mathbb{C}) \otimes \mathbb{C} \mathbb{C}(2),
\]

and the theorem is proved. \( \square \)

As a corollary of Theorem 30.18, we obtain the fact that

\[
\text{Cl}(2k, \mathbb{C}) \approx \mathbb{C}(2^k) \quad \text{and} \quad \text{Cl}(2k+1, \mathbb{C}) \approx \mathbb{C}(2^k) \oplus \mathbb{C}(2^k).
\]
The table of the previous section can also be completed as follows:

<table>
<thead>
<tr>
<th>$n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Cl}_{0,n}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{C}$</td>
<td>$\mathbb{H}$</td>
<td>$\mathbb{H} \oplus \mathbb{H}$</td>
<td>$\mathbb{H}(2)$</td>
<td>$\mathbb{C}(4)$</td>
<td>$\mathbb{R}(8)$</td>
<td>$\mathbb{R}(8) \oplus \mathbb{R}(8)$</td>
<td>$\mathbb{R}(16)$</td>
</tr>
<tr>
<td>$\text{Cl}_{n,0}$</td>
<td>$\mathbb{R}$</td>
<td>$\mathbb{R} \oplus \mathbb{R}$</td>
<td>$\mathbb{R}(2)$</td>
<td>$\mathbb{C}(2)$</td>
<td>$\mathbb{H}(2)$</td>
<td>$\mathbb{H}(2) \oplus \mathbb{H}(2)$</td>
<td>$\mathbb{H}(4)$</td>
<td>$\mathbb{C}(8)$</td>
<td>$\mathbb{R}(16)$</td>
</tr>
<tr>
<td>$\text{Cl}(n, \mathbb{C})$</td>
<td>$\mathbb{C}$</td>
<td>$2\mathbb{C}$</td>
<td>$\mathbb{C}(2)$</td>
<td>$2\mathbb{C}(2)$</td>
<td>$\mathbb{C}(4)$</td>
<td>$2\mathbb{C}(4)$</td>
<td>$\mathbb{C}(8)$</td>
<td>$2\mathbb{C}(8)$</td>
<td>$\mathbb{C}(16)$</td>
</tr>
</tbody>
</table>

where $2\mathbb{C}(k)$ is an abbreviation for $\mathbb{C}(k) \oplus \mathbb{C}(k)$.

30.8 The Groups $\text{Pin}(p, q)$ and $\text{Spin}(p, q)$ as double covers of $\text{O}(p, q)$ and $\text{SO}(p, q)$

It turns out that the groups $\text{Pin}(p, q)$ and $\text{Spin}(p, q)$ have nice topological properties w.r.t. the groups $\text{O}(p, q)$ and $\text{SO}(p, q)$. To explain this, we review the definition of covering maps and covering spaces (for details, see Fulton [69], Chapter 11). Another interesting source is Chevalley [41], where it is proved that $\text{Spin}(n)$ is a universal double cover of $\text{SO}(n)$ for all $n \geq 3$.

Since $C_{p,q}$ is an algebra of dimension $2^{p+q}$, it is a topological space as a vector space isomorphic to $V = \mathbb{R}^{2^{p+q}}$. Now, the group $C^*_{p,q}$ of units of $C_{p,q}$ is open in $C_{p,q}$, because $x \in C_{p,q}$ is a unit if the linear map $y \mapsto xy$ is an isomorphism, and $\text{GL}(V)$ is open in $\text{End}(V)$, the space of endomorphisms of $V$. Thus, $C^*_{p,q}$ is a Lie group, and since $\text{Pin}(p, q)$ and $\text{Spin}(p, q)$ are clearly closed subgroups of $C^*_{p,q}$, they are also Lie groups.

The definition below is analogous to the definition of a covering map given in Section 9.2 (Definition 9.5) except that now, we are only dealing with topological spaces and not manifolds.

**Definition 30.7.** Given two topological spaces $X$ and $Y$, a covering map is a continuous surjective map $p: Y \to X$ with the property that for every $x \in X$, there is some open subset $U \subseteq X$ with $x \in U$, so that $p^{-1}(U)$ is the disjoint union of open subsets $V_\alpha \subseteq Y$, and the restriction of $p$ to each $V_\alpha$ is a homeomorphism onto $U$. We say that $U$ is evenly covered by $p$. We also say that $Y$ is a covering space of $X$. A covering map $p: Y \to X$ is called trivial if $X$ itself is evenly covered by $p$ (i.e., $Y$ is the disjoint union of open subsets $Y_\alpha$ each homeomorphic to $X$), and nontrivial otherwise. When each fiber $p^{-1}(x)$ has the same finite cardinality $n$ for all $x \in X$, we say that $p$ is an $n$-covering (or $n$-sheeted covering).

Note that a covering map $p: Y \to X$ is not always trivial, but always locally trivial (i.e., for every $x \in X$, it is trivial in some open neighborhood of $x$). A covering is trivial iff $Y$ is isomorphic to a product space of $X \times T$, where $T$ is any set with the discrete topology. Also, if $Y$ is connected, then the covering map is nontrivial.

**Definition 30.8.** An isomorphism $\varphi$ between covering maps $p: Y \to X$ and $p': Y' \to X$ is a homeomorphism $\varphi: Y \to Y'$ so that $p = p' \circ \varphi$. 
Typically, the space \( X \) is connected, in which case it can be shown that all the fibers \( p^{-1}(x) \) have the same cardinality.

One of the most important properties of covering spaces is the path–lifting property, a property that we will use to show that \( \text{Spin}(n) \) is path-connected. The proposition below is the analog of Proposition 9.11 for topological spaces and continuous curves.

**Proposition 30.19.** (Path lifting) Let \( p: Y \to X \) be a covering map, and let \( \gamma: [a, b] \to X \) be any continuous curve from \( x_a = \gamma(a) \) to \( x_b = \gamma(b) \) in \( X \). If \( y \in Y \) is any point so that \( p(y) = x_a \), then there is a unique curve \( \tilde{\gamma}: [a, b] \to Y \) so that \( y = \tilde{\gamma}(a) \) and \( p \circ \tilde{\gamma}(t) = \gamma(t) \) for all \( t \in [a, b] \).

**Proof.** See Fulton [70], Chapter 11, Lemma 11.6.

Many important covering maps arise from the action of a group \( G \) on a space \( Y \). If \( Y \) is a topological space, an action (on the left) of a group \( G \) on \( Y \) is a map \( \alpha: G \times Y \to Y \) satisfying the following conditions, where for simplicity of notation, we denote \( \alpha(g, y) \) by \( g \cdot y \):

1. \( g \cdot (h \cdot y) = (gh) \cdot y \), for all \( g, h \in G \) and \( y \in Y \);
2. \( 1 \cdot y = y \), for all \( y \in Y \), where 1 is the identity of the group \( G \);
3. The map \( y \mapsto g \cdot y \) is a homeomorphism of \( Y \) for every \( g \in G \).

We define an equivalence relation on \( Y \) as follows: \( x \equiv y \) iff \( y = g \cdot x \) for some \( g \in G \) (check that this is an equivalence relation). The equivalence class \( G \cdot x = \{ g \cdot x \mid g \in G \} \) of any \( x \in Y \) is called the orbit of \( x \). We obtain the quotient space \( Y/G \) and the projection map \( p: Y \to Y/G \) sending every \( y \in Y \) to its orbit. The space \( Y/G \) is given the quotient topology (a subset \( U \) of \( Y/G \) is open iff \( p^{-1}(U) \) is open in \( Y \)).

Given a subset \( V \) of \( Y \) and any \( g \in G \), we let

\[
g \cdot V = \{ g \cdot y \mid y \in V \}.
\]

We say that \( G \) acts evenly on \( Y \) if for every \( y \in Y \), there is an open subset \( V \) containing \( y \) so that \( g \cdot V \) and \( h \cdot V \) are disjoint for any two distinct elements \( g, h \in G \).

The importance of the notion a group acting evenly is that such actions induce a covering map.

**Proposition 30.20.** If \( G \) is a group acting evenly on a space \( Y \), then the projection map \( p: Y \to Y/G \) is a covering map.

**Proof.** See Fulton [70], Chapter 11, Lemma 11.17.
The following proposition shows that $\text{Pin}(p, q)$ and $\text{Spin}(p, q)$ are nontrivial covering spaces, unless $p = q = 1$.

**Proposition 30.21.** For all $p, q \geq 0$, the groups $\text{Pin}(p, q)$ and $\text{Spin}(p, q)$ are double covers of $\text{O}(p, q)$ and $\text{SO}(p, q)$, respectively. Furthermore, they are nontrivial covers unless $p = q = 1$.

**Proof.** We know that kernel of the homomorphism $\rho: \text{Pin}(p, q) \rightarrow \text{O}(p, q)$ is $\mathbb{Z}_2 = \{-1, 1\}$. If we let $\mathbb{Z}_2$ act on $\text{Pin}(p, q)$ in the natural way, then $\text{O}(p, q) \approx \text{Pin}(p, q)/\mathbb{Z}_2$, and the reader can easily check that $\mathbb{Z}_2$ acts evenly. By Proposition 30.20, we get a double cover. The argument for $\rho: \text{Spin}(p, q) \rightarrow \text{SO}(p, q)$ is similar.

Let us now assume that $p \neq 1$ or $q \neq 1$. In order to prove that we have nontrivial covers, it is enough to show that $-1$ and $1$ are connected by a path in $\text{Pin}(p, q)$ (If we had $\text{Pin}(p, q) = U_1 \cup U_2$ with $U_1$ and $U_2$ open, disjoint, and homeomorphic to $\text{O}(p, q)$, then $-1$ and $1$ would not be in the same $U_i$, and so, they would be in disjoint connected components. Thus, $-1$ and $1$ can’t be path-connected, and similarly with $\text{Spin}(p, q)$ and $\text{SO}(p, q)$.) Since $(p, q) \neq (1, 1)$, we can find two orthogonal vectors $e_1$ and $e_2$ so that $\Phi_{p,q}(e_1) = \Phi_{p,q}(e_2) = \pm 1$. Then,

$$\gamma(t) = \pm \cos(2t) 1 + \sin(2t) e_1 e_2 = (\cos t e_1 + \sin t e_2)(\sin t e_2 - \cos t e_1),$$

for $0 \leq t \leq \pi$, defines a path in $\text{Spin}(p, q)$, since

$$(\pm \cos(2t) 1 + \sin(2t) e_1 e_2)^{-1} = \pm \cos(2t) 1 - \sin(2t) e_1 e_2,$$

as desired. \hfill \square

In particular, if $n \geq 2$, since the group $\text{SO}(n)$ is path-connected, the group $\text{Spin}(n)$ is also path-connected. Indeed, given any two points $x_a$ and $x_b$ in $\text{Spin}(n)$, there is a path $\gamma$ from $\rho(x_a)$ to $\rho(x_b)$ in $\text{SO}(n)$ (where $\rho: \text{Spin}(n) \rightarrow \text{SO}(n)$ is the covering map). By Proposition 30.19, there is a path $\tilde{\gamma}$ in $\text{Spin}(n)$ with origin $x_a$ and some origin $\tilde{x}_b$ so that $\rho(\tilde{x}_b) = \rho(x_b)$. However, $\rho^{-1}(\rho(x_b)) = \{-x_b, x_b\}$, and so $\tilde{x}_b = \pm x_b$. The argument used in the proof of Proposition 30.21 shows that $x_b$ and $-x_b$ are path-connected, and so, there is a path from $x_a$ to $x_b$, and $\text{Spin}(n)$ is path-connected.

In fact, for $n \geq 3$, it turns out that $\text{Spin}(n)$ is simply connected. Such a covering space is called a universal cover (for instance, see Chevalley [41]).

This last fact requires more algebraic topology than we are willing to explain in detail, and we only sketch the proof. The notions of fibre bundle, fibration, and homotopy sequence associated with a fibration are needed in the proof. We refer the perseverant readers to Bott and Tu [24] (Chapter 1 and Chapter 3, Sections 16–17) or Rotman [147] (Chapter 11) for a detailed explanation of these concepts.

Recall that a topological space is simply connected if it is path connected and if $\pi_1(X) = (0)$, which means that every closed path in $X$ is homotopic to a point. Since we just proved that $\text{Spin}(n)$ is path connected for $n \geq 2$, we just need to prove that $\pi_1(\text{Spin}(n)) = (0)$ for all $n \geq 3$. The following facts are needed to prove the above assertion:
30.8. THE GROUPS PIN(P, Q) AND SPIN(P, Q) AS DOUBLE COVERS

(1) The sphere $S^{n-1}$ is simply connected for all $n \geq 3$.

(2) The group $\text{Spin}(3) \simeq \text{SU}(2)$ is homeomorphic to $S^3$, and thus, $\text{Spin}(3)$ is simply connected.

(3) The group $\text{Spin}(n)$ acts on $S^{n-1}$ in such a way that we have a fibre bundle with fibre $\text{Spin}(n-1)$:

$$\text{Spin}(n-1) \to \text{Spin}(n) \to S^{n-1}.$$ 

Fact (1) is a standard proposition of algebraic topology, and a proof can found in many books. A particularly elegant and yet simple argument consists in showing that any closed curve on $S^{n-1}$ is homotopic to a curve that omits some point. First, it is easy to see that in $\mathbb{R}^n$, any closed curve is homotopic to a piecewise linear curve (a polygonal curve), and the radial projection of such a curve on $S^{n-1}$ provides the desired curve. Then, we use the stereographic projection of $S^{n-1}$ from any point omitted by that curve to get another closed curve in $\mathbb{R}^{n-1}$. Since $\mathbb{R}^{n-1}$ is simply connected, that curve is homotopic to a point, and so is its preimage curve on $S^{n-1}$. Another simple proof uses a special version of the Seifert—van Kampen’s theorem (see Gramain [77]).

Fact (2) is easy to establish directly, using (1).

To prove (3), we let $\text{Spin}(n)$ act on $S^{n-1}$ via the standard action: $x \cdot v = xvx^{-1}$. Because $\text{SO}(n)$ acts transitively on $S^{n-1}$ and there is a surjection $\text{Spin}(n) \to \text{SO}(n)$, the group $\text{Spin}(n)$ also acts transitively on $S^{n-1}$. Now, we have to show that the stabilizer of any element of $S^{n-1}$ is $\text{Spin}(n-1)$. For example, we can do this for $e_1$. This amounts to some simple calculations taking into account the identities among basis elements. Details of this proof can be found in Mneimné and Testard [130], Chapter 4. It is still necessary to prove that $\text{Spin}(n)$ is a fibre bundle over $S^{n-1}$ with fibre $\text{Spin}(n-1)$. For this, we use the following results whose proof can be found in Mneimné and Testard [130], Chapter 4:

**Lemma 30.22.** Given any topological group $G$, if $H$ is a closed subgroup of $G$ and the projection $\pi: G \to G/H$ has a local section at every point of $G/H$, then

$$H \to G \to G/H$$

is a fibre bundle with fibre $H$.

Lemma 30.22 implies the following key proposition:

**Proposition 30.23.** Given any linear Lie group $G$, if $H$ is a closed subgroup of $G$, then

$$H \to G \to G/H$$

is a fibre bundle with fibre $H$. 
Now, a fibre bundle is a fibration (as defined in Bott and Tu [24], Chapter 3, Section 16, or in Rotman [147], Chapter 11). For a proof of this fact, see Rotman [147], Chapter 11, or Mneimné and Testard [130], Chapter 4. So, there is a homotopy sequence associated with the fibration (Bott and Tu [24], Chapter 3, Section 17, or Rotman [147], Chapter 11, Theorem 11.48), and in particular, we have the exact sequence

$$\pi_1(\text{Spin}(n-1)) \to \pi_1(\text{Spin}(n)) \to \pi_1(S^{n-1}).$$

Since $\pi_1(S^{n-1}) = (0)$ for $n \geq 3$, we get a surjection

$$\pi_1(\text{Spin}(n-1)) \to \pi_1(\text{Spin}(n)),$$

and so, by induction and (2), we get

$$\pi_1(\text{Spin}(n)) \cong \pi_1(\text{Spin}(3)) = (0),$$

proving that $\text{Spin}(n)$ is simply connected for $n \geq 3$.

We can also show that $\pi_1(\text{SO}(n)) = \mathbb{Z}/2\mathbb{Z}$ for all $n \geq 3$. For this, we use Theorem 30.11 and Proposition 30.21, which imply that $\text{Spin}(n)$ is a fibre bundle over $\text{SO}(n)$ with fibre $\{-1, 1\}$, for $n \geq 2$:

$$\{-1, 1\} \to \text{Spin}(n) \to \text{SO}(n).$$

Again, the homotopy sequence of the fibration exists, and in particular we get the exact sequence

$$\pi_1(\text{Spin}(n)) \to \pi_1(\text{SO}(n)) \to \pi_0(\{-1, +1\}) \to \pi_0(\text{SO}(n)).$$

Since $\pi_0(\{-1, +1\}) = \mathbb{Z}/2\mathbb{Z}$, $\pi_0(\text{SO}(n)) = (0)$, and $\pi_1(\text{Spin}(n)) = (0)$, when $n \geq 3$, we get the exact sequence

$$(0) \to \pi_1(\text{SO}(n)) \to \mathbb{Z}/2\mathbb{Z} \to (0),$$

and so, $\pi_1(\text{SO}(n)) = \mathbb{Z}/2\mathbb{Z}$. Therefore, $\text{SO}(n)$ is not simply connected for $n \geq 3$.

**Remark:** Of course, we have been rather cavalier in our presentation. Given a topological space $X$, the group $\pi_1(X)$ is the *fundamental group of $X$*, i.e. the group of homotopy classes of closed paths in $X$ (under composition of loops). But $\pi_0(X)$ is generally *not* a group! Instead, $\pi_0(X)$ is the set of path-connected components of $X$. However, when $X$ is a Lie group, $\pi_0(X)$ is indeed a group. Also, we have to make sense of what it means for the sequence to be exact. All this can be made rigorous (see Bott and Tu [24], Chapter 3, Section 17, or Rotman [147], Chapter 11).
Bibliography


