

Introduction

1.1 System Models

This book analyzes mathematical models for systems and explores techniques for optimizing systems described by these models. We use the term system in its broad sense; by a **system** we mean a collection of things which are related in such a way that it makes sense to think of them as a whole. Examples of systems are an electric motor, an automobile, a transportation system, and a city. Each of these systems is part of a larger system. Small systems are usually well understood; large, complex systems are not.

Rational decision making concerning the design and operation of a system is always based upon a model of that system. A **model** of a system is a simpler system that behaves sufficiently like the system of interest to be of use in predicting the behavior of the system. The choice of appropriate model depends upon the complexity of the system, the available resources, and the questions that need to be answered by the model. Many decisions are based upon nothing more than the conceptual model which the decision maker develops by observing the operation of other systems. In this book we concern ourselves with a more quantitative class of models, mathematical models.

Most systems can be thought of (or modeled) as an operation on the system inputs (or independent variables) which produces the system outputs (or dependent variables); we state this input-output relationship symbolically by means of the following mathematical equation:

$$\mathbf{T}\mathbf{x}=\mathbf{y} \quad (1.1)$$

In this equation \mathbf{x} represents the set of inputs to the system and \mathbf{y} the set of outputs of the system.* The symbol \mathbf{T} represents the operation which the system performs on the inputs; thus \mathbf{T} is a mathematical model of the system.

*See Section 2.3 for a more complete discussion of inputs and outputs.

In order for a model of a system to be conceptually simple, it must be abstract. The more details we include explicitly in the model, the more complicated it becomes. The more details we make implicit, the more abstract it becomes. Thus if we seek conceptual simplicity, we cannot avoid abstraction. The model \mathbf{T} of (1.1) epitomizes this simplicity and abstraction.

The generality of the model given in (1.1) allows it to be applied to many different systems. In the simplest of situations \mathbf{T} might represent a simple economic transaction: let p be the unit price of a particular commodity; then (1.1) means $\mathbf{y} = p\mathbf{x}$, where \mathbf{x} is the quantity purchased and \mathbf{y} is the total cost of the purchase. At the other extreme, \mathbf{T} might represent a large city. Figure 1.1 shows the system output \mathbf{y} that might result from a given input \mathbf{x} ; obviously, many pertinent variables are not explicit in Figure 1.1.

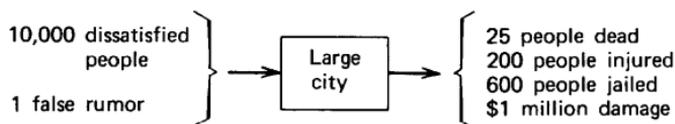


Figure 1.1. A conceptual model of a large city.

Equation (1.1) is the focus of this book. The first five chapters are devoted to a detailed analysis of (1.1) for models \mathbf{T} which are linear.* By decomposing linear models into smaller, simpler pieces we develop an intuitive feel for their properties and determine the practical computational difficulties which can arise in using linear models. Chapter 6 treats the least-square optimization of systems that can be represented by linear models. The analysis and optimization of systems that are described by nonlinear models are considered in Chapters 7-8.

We emphasize linear models because most known analytical results pertain only to linear models. Furthermore, most of the successful techniques for analyzing and optimizing nonlinear systems consist in the repetitive application of linear techniques (Chapter 7-8). We dwell extensively on the two most frequently used linear models—linear algebraic equations and linear differential equations. These models are the most frequently used because they are well understood and relatively easy to deal with. In addition, they are satisfactory models for a large number of practical systems.

Throughout the text we explore the computational implications of the analytical techniques which we develop, but we do not develop computer

*See Section 2.4 for the definition of a linear system.

algorithms. We do not discuss stochastic systems; we treat systems with stochastic inputs only by means of examples.

System Questions

Questions concerning a system usually fall into one of the following categories:

1. *System operation*: in terms of (1.1), given the model \mathbf{T} and the input \mathbf{x} , find the output \mathbf{y} .
2. *System inversion*: given the model \mathbf{T} and output \mathbf{y} , find the input \mathbf{x} .
3. *System synthesis or identification*: given several different choices of input \mathbf{x} and the corresponding output \mathbf{y} for each input, determine a suitable system model \mathbf{T} . (If the system is to be identified, the inputs and outputs are measurements from a real system. If the system is to be synthesized, \mathbf{T} would be chosen to provide some desired input-output relationship.)
4. *System optimization*: pick the input \mathbf{x} , the output \mathbf{y} , or the system \mathbf{T} so that some criterion is optimized.

Note that we have expressed these questions in terms of the system model rather than in terms of the system itself. Although experimentation with actual systems may be appropriate in certain circumstances, these questions are usually explored by means of a model. We discuss the modeling process briefly in Section 1.4. We also examine in Chapter 6 some techniques for making an optimum choice of model parameters once a model structure has been established. However, we do not dwell extensively on techniques for obtaining good models. Rather, we work with the models themselves, assuming that they are good models for the systems they represent. Questions 1 and 2 are treated in Chapters 1, 2, 4, and 5 for linear algebraic equation models and in Chapters 2-5 for linear differential equation models. Question 4 is treated in Chapters 6-8. We do not consider question 3.*

The concepts explored in this book apply directly to any field which uses equations to represent systems or portions of systems. Although we focus on linear algebraic equations and linear differential equations, we also demonstrate the applicability of the concepts to partial differential equations and difference equations; we include equations which are probabilistic, "time-varying," and nonlinear. Our examples pertain to models and optimization in such fields as automatic control, electric power, circuits, statistical communications, coding, heat flow, economics, operations research, etc.

*See Sage [1.10] for a discussion of identification.

1.2 Approach

All students of science and engineering have noticed occasional similarities between the physical laws of different fields. For instance, gravitational attraction, electrostatic attraction, and magnetic attraction all obey an inverse-square law. Electrical resistance to the flow of current has its analogue in the resistance of materials to the conduction of heat. Not only does the physical world tend to repeat itself; it also tends toward simplicity and economy. Most natural phenomena can be explained by simple differential relationships: the net force on a rigid object is proportional to its acceleration; the rate of flow of heat is proportional to the gradient of the temperature distribution.

If we put a number of simple relationships together to describe the motion of a nonrigid object (fuel in a rocket) or the heat flow in an irregular nonhomogeneous object (a nuclear reactor), then nature appears complicated. The human mind is not good at thinking of several things at once. The development of large-scale digital computers has provided the capability for solving complex sets of equations; it has made system study a reality. However, the engineer, the designer of a system, still must conceive of the variables and interactions in the system to such an extent that he can describe for a computer what it is he wants to know. He needs simple conceptual models for systems.

We can simplify models for complex systems by stretching our imagination in a search for analogies. For instance, the multiplication of an electrical current by a resistance to determine a voltage has an analogue in the differentiation of a current and then multiplication by an inductance; both actions are operations on a current to yield a voltage. This analogy suggests that we think of differentiation as analogous to multiplication by a number. By reducing the number of “different concepts” necessary to understand the parts of a system, such analogies help the system designer to achieve greater economy of thought; he can conceive of the system in simpler terms, hopefully gaining insight in the process. William K. Linvill [1.7] has coined the term “portable concept” to describe a concept that is transferable from one setting to another. This book is concerned with *portable mathematical concepts*. The purpose of exploring such concepts is to enhance the ability of the reader to model systems, understand them, synthesize them, and optimize them. Our basic premise is that this ability is enhanced by an intuitive understanding of the models and optimization techniques that have proved useful in many settings in the past. By an intuitive understanding, we mean the type of “intuitive feel” that an engineer obtains by applying and reapplying a concept to many different situations.

It would seem, then, that we must fully absorb most of mathematics. However, much of the mathematical literature is directed toward the

modeling and optimization of pathological cases, those cases for which “standard” models or techniques are insufficient. Because techniques for handling these cases are new, it is appropriate that they be the focus of the current literature. Yet this emphasis on exceptional cases can distort our perspective. In maximizing a function, we should not become so concerned about nondifferentiability of functions that we forget to try setting the derivative equal to zero. Rather than try to explore *all* cases, we focus on well-behaved systems. By making analogies, we organize the most common models and optimization techniques into a framework which contains only a relatively few fundamental concepts. The exceptional cases can be more clearly understood in comparison to this basic framework.

The importance of learning the *structure* of a subject is stressed by Bruner [1.1]: “Grasping the structure of a subject is understanding it in a way that permits many other things to be related to it meaningfully... the transfer of principles is dependent upon mastery of the structure of the subject matter... Perhaps the most basic thing that can be said about human memory, after a century of intensive research, is that unless detail is placed into a structured pattern, it is rapidly forgotten.” In order to simplify and unify the concepts used in model analysis and optimization, we organize fundamental mathematical principles into a mnemonic structure—a structure which draws extensively on geometrical analogies as an aid to the memory. We also develop a mathematical language suitable for communicating these structural concepts.

The first half of this book is concerned with models and their analysis. Mathematically speaking, this is the subject of algebra—the use of symbols to express quantitative concepts and their relations. In the latter half of the book we turn to geometry—the measurement and comparison of quantitative concepts—in order to further analyze models and to optimize their parameters and inputs. Because the bulk of known analytical results are concerned with linear models, these models necessarily dominate our discussions. Our emphasis is on geometrical insight rather than mathematical theorems. We reach deep into the mathematical literature for concepts. We try to be rigorously correct. Yet we develop concepts by means of analogies and simple examples rather than proofs, in order to nurture the intuition of the reader. We concern ourselves with the practical aspects of computation. To engineers the material seems like mathematics; to mathematicians it seems like engineering.

1.3 Portable Concepts

To illustrate the portability of the mathematical model (1.1) we compare the two most common mathematical models: (a) a set of linear algebraic equations; and (b) a linear differential equation. The following algebraic

equations might represent the relationship between the voltages and the currents in a resistive circuit:

$$\begin{aligned} 2\xi_1 + 3\xi_2 &= \eta_1 \\ \xi_1 + \xi_2 &= \eta_2 \end{aligned} \quad (1.2)$$

Such a set of equations is often expressed in the matrix form:

$$\begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \quad (1.3)$$

In the form (1.3), we can interpret the set of equations as an operation (matrix multiplication) on the pair of variables ξ_1 and ξ_2 to obtain the pair of quantities η_1 and η_2 . The relationship (1.2) between the pairs of variables can also be expressed in terms of the “inverse equations”:

$$\begin{aligned} \xi_1 &= -\eta_1 + 3\eta_2 \\ \xi_2 &= \eta_1 - 2\eta_2 \end{aligned} \quad (1.4)$$

Equations (1.4) can be verified by substitution into (1.2). The coefficients in (1.4) indicate what must be done to the “right-hand side” variables in order to determine the solution to (1.2). Equations (1.4) can be expressed in the “inverse matrix” form:

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} -1 & 3 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \quad (1.5)$$

In Section 1.5 we explore in detail the process of solving or inverting equations such as (1.2). In Chapter 2 we begin the discussion of algebraic equation models in a manner which is consistent with the notation of (1.1). Chapters 4 and 5 are, to a great extent, devoted to analyzing these models.

The angular velocity $\omega(t)$ of a particular loaded dc motor, initially at rest, can be expressed in terms of its armature voltage $u(t)$ as

$$\frac{d\omega(t)}{dt} + \omega(t) = u(t), \quad \omega(0) = 0 \quad (1.6)$$

We can think of the differential equation and boundary condition as an abstract operation on ω to obtain u . Equation (1.6) also can be expressed in the inverse form:

$$\omega(t) = \int_0^t e^{-(t-s)} u(s) ds \quad (1.7)$$

That the integral equation (1.7) is, in fact, the solution to (1.6) is easily

verified for a *particular* armature voltage, say, $u(t) = e^{2t}$, by evaluating $\omega(t)$, then substituting it into (1.6). We can think of (1.7) as an abstract “integral” operation on u to determine ω ; this is the “inverse” of the “differential” operation in (1.6). These two abstract operations and techniques for determining the inverse operation are the subject of Chapter 3. The analysis of these abstract operations carries into Chapters 4 and 5.

The algebraic equations (1.2) and the differential equation with its boundary condition (1.6) have much in common. We must not let details cloud the issue; in each case, an “input” is affecting an “output” according to certain (linear) principles. We can think of the pair of variables ξ_1 and ξ_2 and the function ω as each constituting a single “vector” variable. The analogy between these entities is carried further in the comparison of Figure 1.2, wherein the pair of variables ξ_1, ξ_2 is treated as a “discrete” function. This analogy is discussed further in Section 2.1. It seems evident that concepts are more clearly portable if they are abstracted-stripped of their details.

A Portable Optimization Concept

We again employ the analogy between a “discrete vector” variable and a “continuous vector” variable to discuss the portability of an optimization

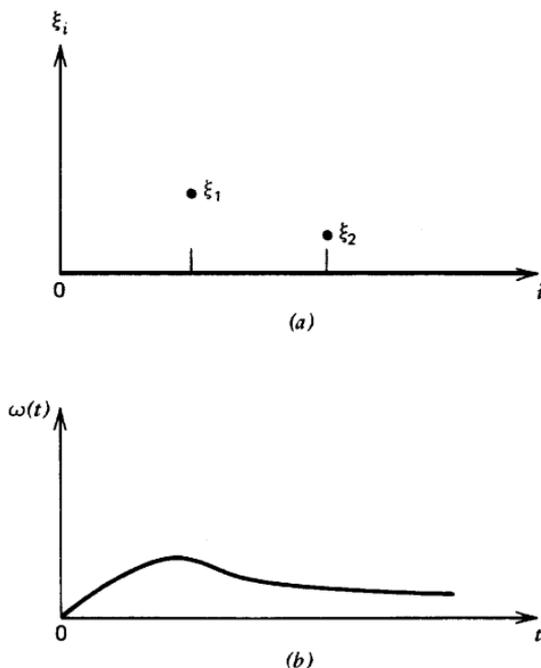


Figure 1.2. Vector variables plotted as functions: (a) discrete variables of (1.2); (b) continuous variable of (1.6).

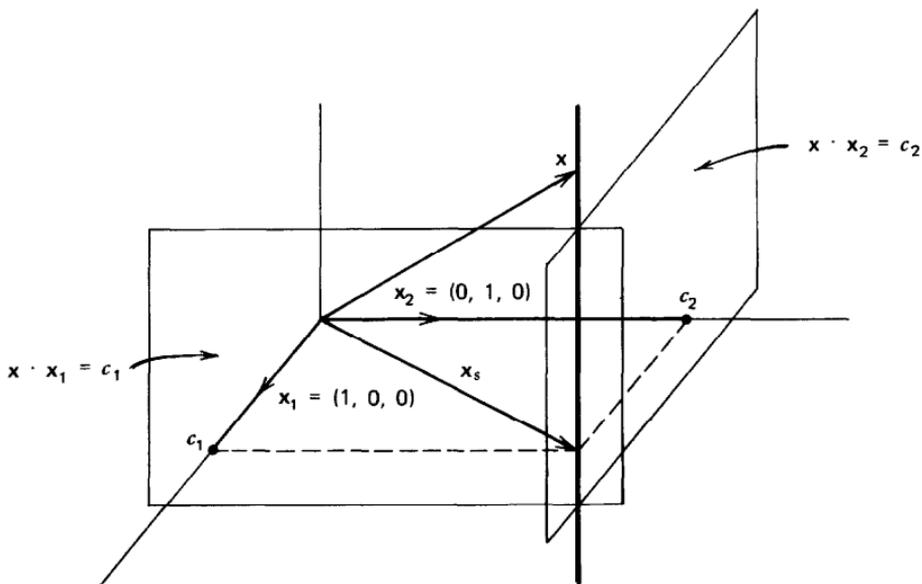


Figure 1.3. A vector of minimum length.

concept. Figure 1.3 shows the locus of all vectors \mathbf{x} in a three-dimensional space which lie in the intersection of two planes. We seek that vector \mathbf{x} which is of minimum length. The solution vector \mathbf{x}_s is perpendicular to the line which constitutes the locus of the candidate vectors \mathbf{x} .

Using the standard notation of analytic geometry, we think of the vector \mathbf{x} as $\mathbf{x} = (\xi_1, \xi_2, \xi_3)$. The plane that is perpendicular to the vector \mathbf{x}_1 can be expressed mathematically in terms of the dot product of vectors as $\mathbf{x} \cdot \mathbf{x}_1 = \xi_1 = c_1$. Similarly, the second plane consists in vectors \mathbf{x} which satisfy $\mathbf{x} \cdot \mathbf{x}_2 = c_2$. Since \mathbf{x}_s must be perpendicular to the intersection of the planes, it must be some combination of the vectors \mathbf{x}_1 and \mathbf{x}_2 that determine the planes; that is, $\mathbf{x}_s = d_1 \mathbf{x}_1 + d_2 \mathbf{x}_2$ for some constants d_1 and d_2 . Substituting \mathbf{x}_s into the equations that determine the planes, we obtain a pair of algebraic equations in d_1 and d_2 :

$$\begin{aligned} \mathbf{x}_s \cdot \mathbf{x}_1 &= d_1 \mathbf{x}_1 \cdot \mathbf{x}_1 + d_2 \mathbf{x}_2 \cdot \mathbf{x}_1 = c_1 \\ \mathbf{x}_s \cdot \mathbf{x}_2 &= d_1 \mathbf{x}_1 \cdot \mathbf{x}_2 + d_2 \mathbf{x}_2 \cdot \mathbf{x}_2 = c_2 \end{aligned} \quad (1.8)$$

Since the vectors \mathbf{x}_1 and \mathbf{x}_2 are perpendicular and of unit length, then

$$\mathbf{x}_1 \cdot \mathbf{x}_1 = \mathbf{x}_2 \cdot \mathbf{x}_2 = 1, \quad \mathbf{x}_1 \cdot \mathbf{x}_2 = \mathbf{x}_2 \cdot \mathbf{x}_1 = 0, \quad d_i = c_i,$$

and

$$\mathbf{x}_s = c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 = (c_1, c_2, 0)$$

The geometric minimization problem described above is simple. By using geometric notions, we have found the vector \mathbf{x} which satisfies two linear equations and for which the quantity $\xi_1^2 + \xi_2^2 + \xi_3^2$ (the length of \mathbf{x} squared) is minimum. The same geometric principles can be used to solve other, more complicated, problems wherein linear equations must be satisfied and a quadratic quantity minimized. For instance, the angular position $\phi(t)$ of the shaft of the dc motor of (1.6) satisfies

$$\frac{d^2\phi(t)}{dt^2} + \frac{d\phi(t)}{dt} = u(t) \quad (1.9)$$

Suppose we seek that armature voltage function $u(t)$ that will drive the motor shaft from one position to another in a fixed time, while consuming a minimum amount of energy; that is, let $\phi(0) = \dot{\phi}(0) = 0$, $\phi(1) = 1$, $\dot{\phi}(1) = 0$, and pick u to minimize $\int_0^1 u^2(t) dt$. In our search for a technique for solving this problem we should not cloud the issue by thinking about techniques for solving differential equations. Equation (1.9) is linear; the quantity to be minimized is quadratic. Chapter 6 is devoted to solving such problems by using analogues of the planes and perpendicular vectors of Figure 1.3.

1.4 System Modeling

The rationale for modeling a system is a desire to determine how to design and/or operate a system without experimenting with actual systems. If a system is large, experimenting is usually very time consuming, extremely expensive, and often socially unacceptable. A designer uses models to predict the performance characteristics of a system or to aid in modifying the design of the system so that it meets a desired set of specifications. He will probably be interested in the degree of stability of the system, its accuracy, and its speed of response to commands. The designer also uses models to predict the nature of the interaction of the system with other systems. For example, he may wish to predict the effect of the system or of a particular system operating policy on the environment or on a related energy distribution system. Or he may wish to predict the performance of the system in the presence of extraneous inputs (noise) or sudden changes in load. The reliability of the system and the sensitivity of the system performance to changes in the environment are also important.

Types of Models

A single system has many models. One or more models of the system pertain to its electrical behavior, others to its thermal behavior, still others

to its mechanical behavior. An investigation of the social or economic characteristics of the system requires additional models.

Physical models are appropriate in many situations. One example of such a model is a scale model of a building or bridge. The conceptual representation of a rocket by a solid cylinder is another example. In most system studies, a *mathematical model* for the system (or part of the system) facilitates analysis. An appropriate mathematical model usually can be derived more easily from a simplified physical model than from the original system. The resulting mathematical model usually consists of a set of algebraic and/or differential equations. Often these equations can be solved (for given system inputs) on a digital, analogue, or hybrid computer.[†] In some instances, the distributed nature of the system requires a mathematical model consisting of partial differential equations, and computer solutions are difficult to obtain even if the equations are linear.

The behavior of some systems fluctuates randomly with time. For such systems (or portions of systems) it is common to build a discrete-event *simulation model*.^t Rather than predicting the precise behavior of the system, such a model simulates the behavior numerically in a manner that is statistically correct. For instance, we might be interested in the flow of customers through a set of checkout counters. A simple physical model of such a customer service system consists of a single checkout counter, where customers arrive, wait for service, are served, then leave; arrival times and service times are random with known statistics. By means of a digital computer, we would generate a random sequence of arrivals (with correct statistical properties). We would also determine a service time for each customer by an appropriate random number generation process. Then we would observe the simulated flow of customers over time. The simulation would predict not only the average flow through the system, but also the frequency of occurrence of various queue lengths and waiting times. Thus the dynamic performance of certain types of systems can be predicted by digital simulation.

As a practical matter, a model should contain no more detail than is necessary to accomplish the purposes of the model. One is seldom sure of the accuracy of a model. Yet if a model is accurate enough to improve one's decision-making capability, it serves a useful purpose. Generally speaking, the more complex the model is, the more expensive will be the process of developing and using the model. In the extreme, the most accurate model is a copy of the system itself.

*Special computer programs have been developed to facilitate the solving of certain classes of equations. One example is MATLAB[®]; it is effective in solving linear algebraic and linear differential equations.

[†]Specialized computer languages have been developed to facilitate discrete-event simulation. Examples are ARENA[®], SIMSCRIPT[®], and GPSS[®].

Unfortunately, it is probable that some complex systems will never be represented in sufficient detail by manageable mathematical models. Yet a **conceptual model** can be applied in situations where it is difficult to obtain meaningful quantitative models; for example, the principle of negative feedback (with its beneficial effects on stability and sensitivity) often is applied successfully without the use of a mathematical model. The system concepts that are associated with mathematical models serve as a guide to the exploration of complex systems. By the use of specific models for small subsystems, by computer analysis of the combined subsystem models, and by the application of model concepts (such as feedback) to the whole system, we can better understand large systems.

The Modeling Process

The process of modeling can be divided into two closely related steps: (1) establishing the model structure and (2) supplying the data. We focus primarily on the first step. However, we cannot ignore the second; it is seldom useful to establish a model structure for which we cannot obtain data.

We begin the modeling process by examining the system of interest. In many complex systems, even the boundaries of the system are not clear. The motivation for modeling such a system is usually a desire to solve a problem, to improve an unsatisfactory situation, or to satisfy a felt need. We must describe the system and the manner in which it performs in a simple fashion, omitting unnecessary detail. As we begin to understand better the relationship between the system and the problem which motivates study of the system, we will be able to establish suitable boundaries for the system.

Suppose a housing official of a large city is concerned because the number of vacant apartments in his city cycles badly, some times being so high as to seriously depress rental rates, other times being so low as to make it difficult for people to find or afford housing.* What is the reason for the cycling? To answer this question, we need to explore the "housing system." Should we include in "the system" the financial institutions which provide capital? The construction industry and labor unions which affect new construction? The welfare system which supports a significant fraction of low-income housing? Initially, we would be likely to concern ourselves only with the direct mechanisms by which vacant apartments are generated (new construction, people moving out, etc.) and eliminated (new renters).

Should the model account for different sizes of apartments? Different styles? Different locations? Seasonal variations in the number of vacan-

*The idea for this example was obtained from Truxal [1.1], Chapter 21.

ties? A model that accounts for all these factors would require detailed data (as a function of time) for each factor. These data are not likely to be readily available. Rather, obtaining the data would require the cooperation of many apartment managers and an extensive data-taking operation over at least a 1-year period. A more likely approach, at least initially, would be to develop a simple model which predicts the average number of vacancies (of any type) in the city in a 1-year period. Data concerning this quantity are probably available for at least a large fraction of the large apartment complexes in the city.

Once the approximate extent of the system and the approximate degree of detail of the model have been determined, the course of model development usually progresses through the following steps:

1. Development of a simple physical model.
2. Derivation of a mathematical model of the physical model.
3. Obtaining of data from which model parameters are determined.
4. Validation of the model.

In deriving a model for a system it usually helps to visualize the behavior of the unfamiliar system in terms of the behavior of familiar systems which are similar. It is for this reason that we start with a simple physical model. The physical model of the system is likely to be conceptual rather than actual. It is a simple abstraction which retains only the essential characteristics of the original system. In the case of the apartment vacancy model introduced above, a simple physical model might consist of a set of identical empty boxes (vacant apartments). At 1-year intervals some number of boxes is added by construction or renters moving out; another number of boxes is removed by new renters. See Figure 1.4.

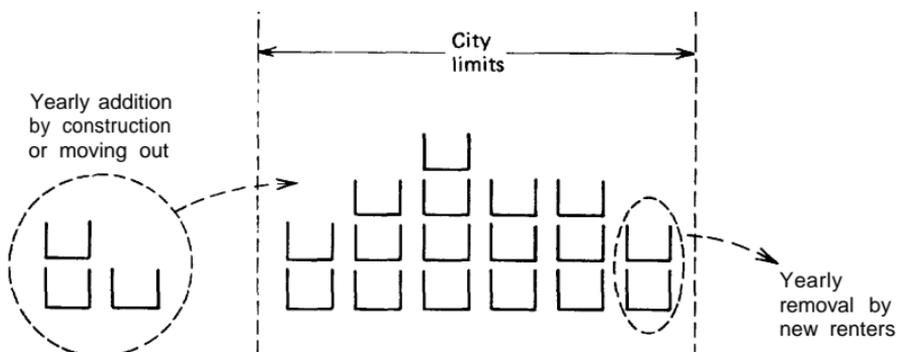


Figure 1.4. Simple physical model of apartment vacancies.

A mathematical model of a system is usually easier to derive from a simple physical model than from the system itself. In most instances the mathematical model consists of algebraic and/or differential equations. The mathematical model must be kept simple in order that it be solvable analytically or by means of practical computer techniques. Generally, the model simplifications that reduce data requirements also reduce the complexity of the mathematical model. For example, in the housing system described above, the aggregation of the various types of apartments into a single type greatly reduces the number of variables in the mathematical model. Other simplifying approximations which may be appropriate in some situations are (1) ignoring interaction between the system and its environment; (2) neglecting uncertainty and noise; (3) lumping distributed characteristics; and (4) assuming linearity and time invariance. Sage [1.10] describes some techniques that are useful in identifying the structure and parameter values of those systems that act in a linear fashion.

Mathematical Model of Apartment Vacancies

In order to demonstrate the logical thought process entailed in the derivation of a mathematical model, we derive a mathematical model of the physical model of apartment vacancies illustrated in Figure 1.4.

We expect that the number of “apartment construction starts” in a given year is approximately equal to the apparent need for new apartments. We formalize this statement by postulating the following relationship:

$$\begin{aligned}
 S(n) &= \alpha(V_d - V(n)), & V(n) &\leq V_d \\
 &= 0, & V(n) &\geq V_d
 \end{aligned}
 \tag{1.10}$$

where $S(n)$ = number of apartment construction starts in year n ;

$V(n)$ = average number of vacant apartments during the 1-year period centered at the beginning of year n .

Underlying (1.10) is the assumption that the people who build apartments feel that the city should have approximately V_d vacancies. The proportionality factor α and the number of vacancies V_d should be selected in such a manner that (1.10) most nearly describes recent historical data for the city.

Of course, actual apartment completions lag behind the starts by an appreciable time. We formalize this statement by the equation

$$C(n) = S(n - l)
 \tag{1.11}$$

where $C(n)$ is the number of completions in year n , and l is the average

construction time. A suitable value for the lag l should be determined from historical data.

Let $R(n)$ denote the number of new apartments rented during year n . We can include in $R(n)$ the families who move out of apartments during the year [$R(n)$ can be negative]. From Figure 1.4, it is apparent that

$$\Delta V(n) = C(n) - R(n) \quad (1.12)$$

where $\Delta V(n) = V(n+1) - V(n)$, the increase in vacant apartments during the 1-year period.

The empirical relations (1.10)-(1.11) and the logical statement (1.12) can be related pictorially by means of a **block diagram**. A block diagram is a conceptual tool which is useful for clarifying the structure of a model or for portraying sequences of events. It dramatizes cause and effect relationships. A block diagram of the mathematical model (1.10)-(1.12) is shown in Figure 1.5. Each block in the diagram displays one of the relationships in the mathematical model. *

Figure 1.5 establishes the model structure. In order to determine the values of the model parameters and to validate the model, we need historical data for each variable in the model. The data that we need in order to pick appropriate values for the parameters α , V_d , and l are historical values of yearly starts $S(n)$, yearly completions $C(n)$, and yearly average vacancies $V(n)$. We would probably pick the values of α , V_d , and l by the least-square data-fitting process known as *linear regression* (see Section 6.1).

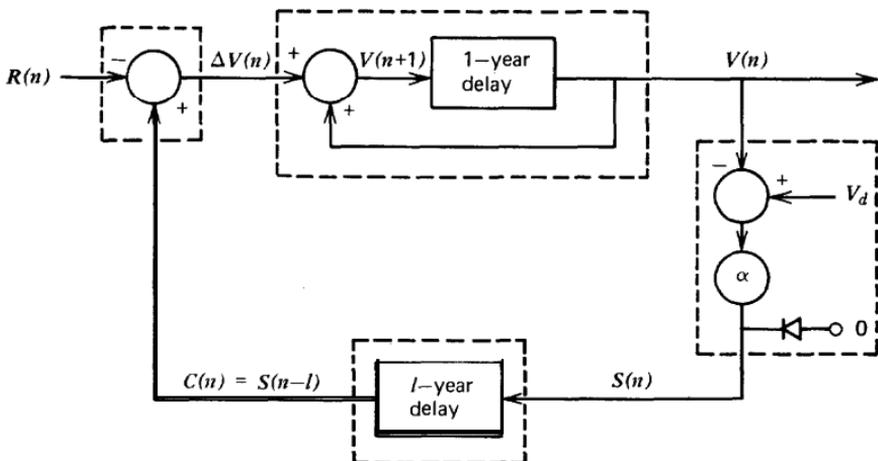


Figure 1.5. Block diagram model of apartment vacancies.

*See Cannon [1.2] for a detailed discussion of block diagrams and their use.

After parameter values have been determined, we need to verify that the mathematical model is a sufficiently good representation of the actual apartment vacancy system. In order to validate the model, we need historical values of the model input $R(n)$ and output $V(n)$. Since we required data for $V(n)$ previously, the only additional data needed are a corresponding set of yearly rentals $R(n)$ (new rentals minus renters moving out). We use the input data $R(n)$ for a sequence of years together with the mathematical model to obtain a predicted sequence of values of $V(n)$. The model is validated if the predicted values of $V(n)$ agree sufficiently with the corresponding historical values of $V(n)$. If the model were verified to be accurate to a certain precision for historical data, we would feel confident that it would exhibit approximately the same accuracy in predicting future apartment vacancies. A housing official would probably be satisfied if the predicted vacancies were within 10% of the actual average vacancies. Of course, predictions of future values of $V(n)$ have to be based on assumed future values of $R(n)$. If future values of $R(n)$ cannot be predicted with reasonable confidence, then another model must be developed to relate the demand for apartments $R(n)$ to those variables which affect demand.

If the data do not validate the model to a sufficient degree, then the model structure must be modified; additional factors must be accounted for. Specifically, the number of apartment construction starts $S(n)$ is likely to depend not only on the demand for housing $R(n)$, but also on the number of uncompleted housing starts (starts from the previous $l-1$ years). The number of starts $S(n)$ is also likely to depend on the availability of capital at a favorable interest rate. Thus an improved apartment vacancy model would probably have more than one input variable.

Once a validated model has been obtained, it can be used to aid city officials in determining an appropriate housing policy. City officials can affect the number of apartment vacancies by modifying the variables which are inputs to the model. Demand for apartments $R(n)$ can be affected by adjusting tax rates, rent subsidies, urban renewal plans, etc. If the final model includes interest rate as an input, this interest rate can be affected by means of interest rate subsidies.

Suppose that low interest capital has been plentiful, and there has been an overabundance of housing. Specifically, suppose recent historical data indicate that the best values for the parameters of the model in Figure 1.5 are $V_d = 1000$ apartments, $\alpha = 0.5$, and $l = 2$ years, and that reasonable initial conditions are $V(0) = 1500$ vacancies, and $S(-2) = S(-1) = 0$ apartments. Suppose that as a result of a new rent subsidy program we expect the future demand to be $R(n) = 500$ apartments, $n = 0, 1, 2, \dots$. According to the mathematical model of (1.10)-(1.12) and Figure 1.5, the new rent subsidy program will cause the apartment vacancies in the city to exhibit the behavior shown in Table 1.1 and Figure 1.6.

Table 1.1 Apartment Vacancies Predicted by Figure 1.5

n	$V(n)$	$S(n)$	$C(n)$	$R(n)$	$\Delta V(n)$	$V(n+1)$
0	1500	0	0	500	-500	1000
1	1000	0	0	500	-500	500
2	500	250	0	500	-500	0
3	0	500	0	500	-500	-500
4	-500	750	250	500	-250	-750
5	-750	875	500	500	0	-750
6	-750	875	750	500	250	-500
7	-500	750	875	500	375	-125
8	-125	563	875	500	375	250
9	250	375	750	500	250	500
10	500	250	563	500	63	563
11	563	219	375	500	-125	438
12	438	281	250	500	-250	188

According to Figure 1.6, the model predicts that severe housing shortages will result from the new housing policy. If the model is correct, and if social pressures make the rent subsidy program mandatory, then the city officials must compensate for the policy by encouraging builders to expand the available housing. (Perhaps this expansion could be encouraged by publicizing the predicted housing shortage, or by having the city assume some of the risk of investment in new construction.)

If the model has not been carefully validated, however, the predictions that result from the model should be used with caution. The fact that builders themselves might predict future housing shortages is ignored in

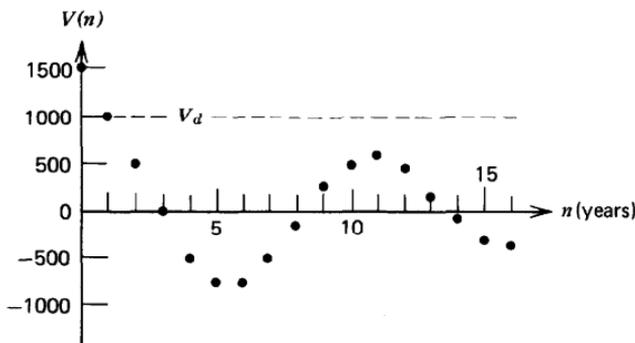


Figure 1.6. Apartment vacancies predicted by Figure 1.5.

(1.10). Thus this simple model of the relationship between vacancies and construction starts should probably be modified to more accurately describe the process by which builders decide to start new construction. Then the determination of model parameter values and the validation of the model should be repeated for the new model before it is used to predict the effect of housing policies.

The modeling process we have described has been used extensively to describe such situations as the flow of electric power in large transmission line networks and the growth of competing species in ecosystems. It is apparent that the same modeling process can be used to describe the relationships among the variables in many other types of systems. For example, it is suitable for describing the response of an eye pupil to variations in light intensity, the response of a banking system to market fluctuations, or the response of the people of a given country to variations in the world price of oil. It is in the social, economic, and biological fields that system modeling is likely to have its greatest impact in the future.

1.5 Solution of Linear Algebraic Equations

To this point our discussion has been of an introductory nature. The development of vector space concepts and the vector space language begins in Chapter 2. We now explore briefly, in a matrix format, the process of solving sets of linear algebraic equations, in order that we be able to use such sets of equations in the examples of Chapter 2 and later chapters. In this discussion we emphasize practical techniques for computing solutions to sets of linear algebraic equations and for computing the inverses of square matrices.

Models of most systems eventually lead to the formation and solution of sets of linear algebraic equations. For example, it is common practice to replace the derivatives in a differential equation by finite differences, thereby producing a set of linear algebraic equations which can be solved by a digital computer. The solution of nonlinear equations almost always requires linearization and, again, involves solution of linear algebraic equations (Chapter 8). Thus simultaneous algebraic equations are fundamental to practical analysis.

There is a wide variety of methods for solving a set of linear algebraic equations.* The design of *practical* computer algorithms which will obtain accurate solutions in an efficient manner calls upon most of the concepts of this book: spectral analysis, least-square optimization, orthogonalization, iteration, etc. Frequently, the sets of equations that arise in practice

*See Forsythe [1.6].

are nearly degenerate; that is, they border on being unsolvable by computers which have finite accuracy. Furthermore, the number of equations can be large; finite-difference approximations for partial differential equations sometimes involve more than 100,000 equations (P&C 2.17). Thus the solution of linear algebraic equations constitutes one of the easiest, and yet one of the most difficult problems.

Any set of linear algebraic equations can be written in the form

$$\begin{aligned} a_{11}\xi_1 + a_{12}\xi_2 + \cdots + a_{1n}\xi_n &= \eta_1 \\ \vdots & \\ a_{m1}\xi_1 + a_{m2}\xi_2 + \cdots + a_{mn}\xi_n &= \eta_m \end{aligned} \quad (1.13)$$

Equation (1.13) easily fits the symbolic structure of the basic system model (1.1). Suppose we define $\mathbf{x} \triangleq \{\xi_1, \xi_2, \dots, \xi_n\}$ and $\mathbf{y} \triangleq \{\eta_1, \eta_2, \dots, \eta_m\}$ as the unknown inputs and known outputs, respectively, of the model, \mathbf{T} . Our immediate goal is to clarify the manner in which \mathbf{T} , by way of the coefficients a_{ij} , relates \mathbf{x} to \mathbf{y} . Associated with (1.13) are three basic questions:

1. Do the equations possess a solution \mathbf{x} for each given \mathbf{y} ; that is, are the equations consistent?
2. Is the solution unique; that is, are there enough independent equations to determine \mathbf{x} ?
3. What is the solution (or solutions)?

It is appropriate to ask the same questions concerning (1.1). Although the third question may appear to be the most pertinent for a specific problem, the answers to the other two give valuable insight into the structure of the model and its applicability to the situation it is supposed to represent. Such insight is generally the real reason for solving the equations, and certainly the prime purpose of our present analysis.

We rephrase the problem in matrix notation in order to separate the information about the system $\{a_{ij}\}$ from the information about the "state" or "condition" of the system (the variables $\{\xi_i\}$, $\{\eta_j\}$).

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_m \end{pmatrix} \quad (1.14)$$

Matrix multiplication is defined in such a way that (1.13) and (1.14) are equivalent.* The notation of (1.14) is close to the abstract symbolism of (1.1). In order to be more direct concerning the meaning of \mathbf{T} , we redefine \mathbf{x} and \mathbf{y} as the column matrices:

$$\mathbf{x} \triangleq \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \quad \mathbf{y} \triangleq \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_m \end{pmatrix}$$

Then (1.14) states

$$\mathbf{Ax} = \mathbf{y} \tag{1.15}$$

where \mathbf{A} is the $m \times n$ matrix of equation coefficients. The system \mathbf{T} can be defined explicitly by $\mathbf{T}\mathbf{x} \triangleq \mathbf{Ax}$; that is, the abstract operation of the system model \mathbf{T} on the "vector" \mathbf{x} is multiplication of \mathbf{x} by the matrix \mathbf{A} .

Typical of the classical methods of solution of (1.15) is Cramer's formula (Appendix 1):

$$\xi_i = \frac{\det(\mathbf{A}(i))}{\det(\mathbf{A})}$$

where $\mathbf{A}(i)$ is the matrix \mathbf{A} with its i th column replaced by \mathbf{y} . The formula applies only when \mathbf{A} is square ($m = n$) and $\det(\mathbf{A}) \neq 0$. The method indicates that for square \mathbf{A} , $\det(\mathbf{A}) \neq 0$ is a necessary and sufficient condition to guarantee a unique solution \mathbf{x} to (1.15).

The most efficient scheme for evaluating a determinant requires approximately $n^3/3$ multiplications (Appendix 1 and P&C 1.3). Thus solution for \mathbf{x} using Cramer's formula requires $(n + 1)n^3/3$ multiplications. Compared with other techniques, Cramer's formula is not a practical tool for analyzing linear equations.

Row Reduction

Ordinary elimination of variables forms the basis for an efficient method of solution to (1.15). In point of fact, it is the basis for most computer algorithms for solving sets of linear algebraic equations. In essence, the method consists in successively adding some multiple of one equation to another until only one variable remains in each equation; then we obtain

*See Appendix 1 for a brief introduction to matrices and determinants.

the unknowns by inspection. For example:

$$\begin{array}{rcl} \xi_1 + 2\xi_2 = 2 & \longrightarrow & \xi_1 + 2\xi_2 = 2 \\ 3\xi_1 + 4\xi_2 = 6 & & -2\xi_2 = 0 \end{array} \longrightarrow$$

$$\begin{array}{rcl} \xi_1 + 2\xi_2 = 2 & \longrightarrow & \xi_1 = 2 \\ \xi_2 = 0 & & \xi_2 = 0 \end{array}$$

The elimination method reduces to an automatable procedure (or algorithm) which requires no creative decision making by the user. Since the unknowns are unaffected by the procedure, they need not be written down; the above elimination process is expressed in matrix notation by

$$\begin{pmatrix} 1 & 2 & \vdots & 2 \\ 3 & 4 & \vdots & 6 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & \vdots & 2 \\ 0 & -2 & \vdots & 0 \end{pmatrix} \longrightarrow$$

$$\begin{pmatrix} 1 & 2 & \vdots & 2 \\ 0 & 1 & \vdots & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & \vdots & 2 \\ 0 & 1 & \vdots & 0 \end{pmatrix}$$

The first matrix in this elimination process is $(\mathbf{A} : \mathbf{y})$; we call it the **augmented matrix** (we augmented \mathbf{A} with \mathbf{y}). We refer to the matrix version of this elimination process as row reduction of the matrix $(\mathbf{A} : \mathbf{y})$. Specifically, **row reduction of a matrix \mathbf{B}** consists in systematically operating on the rows of \mathbf{B} as if they were equations until (a) the first nonzero element in each row is 1; (b) each column which contains the leading 1 for some row has all its other entries 0; and (c) the leading 1's are in an order which descends from the left, with all zero rows at the bottom. We need the last requirement only to make the row-reduced matrix unique. We call the row-reduced matrix the **echelon form** (or Hermite normal form) of \mathbf{B} .

There are two basic techniques for row reducing a matrix. In **Gauss-Jordan elimination** we complete the operations on each column, obtaining a single 1 with all other elements 0, before concerning ourselves with succeeding columns (Example 1). In **Gaussian elimination** we first eliminate all elements below the main diagonal, one column at a time, thereby making the matrix "upper triangular." We then eliminate elements above the diagonal by a process commonly called "back substitution." In Example 2 the first three steps demonstrate the triangularization, the last two the back substitution. Although the two methods are similar, Gaussian elimination is 33% more efficient than Gauss-Jordan elimination for large sets of equations (say, $n > 5$); Gaussian elimination requires about $n^3/3$ multiplications to row reduce $(\mathbf{A} : \mathbf{y})$ for an $n \times n$ matrix \mathbf{A} . Gauss-Jordan

elimination requires about $n^3/2$ multiplications. Both methods are far superior to Cramer's formula for solving linear algebraic equations (P&C 1.3).

Example 1. Gauss Jordan Elimination

$$\begin{pmatrix} 1 & 2 & 2 & 1 \\ 2 & 3 & 5 & 1 \\ 3 & 2 & 5 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} \textcircled{1} & 2 & 2 & 1 \\ 0 & -1 & 1 & -1 \\ 0 & -4 & -1 & -2 \end{pmatrix} \\ \rightarrow \begin{pmatrix} 1 & \textcircled{0} & 4 & -1 \\ 0 & 1 & -1 & 1 \\ 0 & \textcircled{0} & -5 & 2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & \textcircled{0} & \frac{3}{5} \\ 0 & 1 & 0 & \frac{3}{5} \\ 0 & 0 & 1 & -\frac{2}{5} \end{pmatrix}$$

Example 2. Gaussian Elimination

$$\begin{pmatrix} 1 & 2 & 2 & 1 \\ 2 & 3 & 5 & 1 \\ 3 & 2 & 5 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} \textcircled{1} & 2 & 2 & 1 \\ 0 & -1 & 1 & -1 \\ 0 & -4 & -1 & -2 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 2 & 1 \\ 0 & \textcircled{1} & -1 & 1 \\ 0 & \textcircled{0} & -5 & 2 \end{pmatrix} \\ \rightarrow \begin{pmatrix} 1 & 2 & 2 & 1 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & \textcircled{1} & -\frac{2}{5} \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & \textcircled{0} & \frac{2}{5} \\ 0 & 1 & \textcircled{0} & \frac{3}{5} \\ 0 & 0 & 1 & -\frac{2}{5} \end{pmatrix} \rightarrow \begin{pmatrix} 1 & \textcircled{0} & 0 & \frac{3}{5} \\ 0 & 1 & 0 & \frac{3}{5} \\ 0 & 0 & 1 & -\frac{2}{5} \end{pmatrix}$$

In the row reduction of small matrices by hand, the number of multiplications is of less concern than is accuracy. To guard against errors during row reduction of a matrix \mathbf{B} , we can add a "check" column whose i th element is the sum of the elements in the i th row of \mathbf{B} . Throughout the row-reduction process the i th element in the check column should remain equal to the sum of all other elements in the i th row; wherever it is not equal to that sum, one of the elements in that row is in error. Because adding fractions by hand is complicated, we can avoid fractions by not forcing nonzero elements to be 1 until the last step in the row-reduction process.

Example 3. Row Reduction by Hand

$$(\mathbf{B} \text{ : check column}) \stackrel{\Delta}{=} \begin{pmatrix} 3 & 1 & 2 & \vdots & 6 \\ 4 & 2 & 1 & \vdots & 7 \end{pmatrix} \rightarrow \begin{pmatrix} 12 & 4 & 8 & \vdots & 24 \\ 12 & 6 & 3 & \vdots & 21 \end{pmatrix} \\ \rightarrow \begin{pmatrix} 12 & 4 & 8 & \vdots & 24 \\ 0 & 2 & -5 & \vdots & -3 \end{pmatrix} \rightarrow \begin{pmatrix} 6 & 2 & 4 & \vdots & 12 \\ 0 & 2 & -5 & \vdots & -3 \end{pmatrix} \rightarrow \begin{pmatrix} 6 & 0 & 9 & \vdots & 15 \\ 0 & 2 & -5 & \vdots & -3 \end{pmatrix} \\ \rightarrow \begin{pmatrix} 1 & 0 & \frac{3}{2} & \vdots & \frac{5}{2} \\ 0 & 1 & -\frac{5}{2} & \vdots & -\frac{3}{2} \end{pmatrix}$$

If we are interested in the solution to a set of equations $\mathbf{Ax} = \mathbf{y}$ as a function of \mathbf{y} , we can carry an unspecified \mathbf{y} through the row-reduction process.

Example 4. Row Reduction with an Unspecified Column

$$\begin{aligned}
 (\mathbf{A} : \mathbf{y}) &\stackrel{\Delta}{=} \begin{pmatrix} 1 & 2 & 2 & \vdots & \eta_1 \\ 2 & 3 & 5 & \vdots & \eta_2 \\ 3 & 2 & 5 & \vdots & \eta_3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 2 & \vdots & \eta_1 \\ 0 & -1 & 1 & \vdots & \eta_2 - 2\eta_1 \\ 0 & -4 & -1 & \vdots & \eta_3 - 3\eta_1 \end{pmatrix} \\
 &\rightarrow \begin{pmatrix} 1 & 0 & 4 & \vdots & -3\eta_1 + 2\eta_2 \\ 0 & 1 & -1 & \vdots & 2\eta_1 - \eta_2 \\ 0 & 0 & -5 & \vdots & 5\eta_1 - 4\eta_2 + \eta_3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & \vdots & \eta_1 - \frac{6}{5}\eta_2 + \frac{4}{5}\eta_3 \\ 0 & 1 & 0 & \vdots & \eta_1 - \frac{1}{5}\eta_2 - \frac{1}{5}\eta_3 \\ 0 & 0 & 1 & \vdots & -\eta_1 + \frac{4}{5}\eta_2 - \frac{1}{5}\eta_3 \end{pmatrix}
 \end{aligned}$$

The solution to the equations represented by the matrix $(\mathbf{A} : \mathbf{y})$ of Example 4 can be expressed

$$\mathbf{x} = \begin{pmatrix} \eta_1 - \frac{6}{5}\eta_2 + \frac{4}{5}\eta_3 \\ \eta_1 - \frac{1}{5}\eta_2 - \frac{1}{5}\eta_3 \\ -\eta_1 + \frac{4}{5}\eta_2 - \frac{1}{5}\eta_3 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{6}{5} & \frac{4}{5} \\ 1 & -\frac{1}{5} & -\frac{1}{5} \\ -1 & \frac{4}{5} & -\frac{1}{5} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}$$

Clearly, the final coefficients on the variables $\{\eta_i\}$ constitute the inverse matrix \mathbf{A}^{-1} . The coefficients which multiply these variables during the row reduction keep a record of the elimination operations on the rows of \mathbf{A} . The variables $\{\eta_i\}$ merely serve to keep the coefficients separated. The row reduction of Example 4 was, in effect, performed on $(\mathbf{A} : \mathbf{I})$ to obtain $(\mathbf{I} : \mathbf{A}^{-1})$, where \mathbf{I} is the identity matrix; that is,*

$$\begin{pmatrix} 1 & 2 & 2 & \vdots & 1 & 0 & 0 \\ 2 & 3 & 5 & \vdots & 0 & 1 & 0 \\ 3 & 2 & 5 & \vdots & 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & \vdots & 1 & -\frac{6}{5} & \frac{4}{5} \\ 0 & 1 & 0 & \vdots & 1 & -\frac{1}{5} & -\frac{1}{5} \\ 0 & 0 & 1 & \vdots & -1 & \frac{4}{5} & -\frac{1}{5} \end{pmatrix}$$

Row reduction is an efficient method for computing \mathbf{A}^{-1} . Yet in most instances, computation of \mathbf{A}^{-1} is, in itself, inefficient. Computing \mathbf{A}^{-1} by using Gaussian elimination on $(\mathbf{A} : \mathbf{I})$ requires $\frac{4}{3}n^3$ multiplications for an $n \times n$ matrix \mathbf{A} (P&C 1.3). Since this is four times the number of multiplications needed to find the solution \mathbf{x} for a given \mathbf{y} , we find the inverse only when we actually need it—when we are interested in the properties of the system model (the set of equations) and of the matrix \mathbf{A} which represents it.

*In Appendix 1, \mathbf{A}^{-1} is defined as a matrix which satisfies $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. In P&C 1.4 we find that if such a matrix exists, the row reduction of $(\mathbf{A} : \mathbf{I})$ will produce it.

Many system models lead to matrices which are not square; there can be more equations than unknowns; there can be fewer. Even if the matrix is square, its inverse need not exist. Yet for any $m \times n$ matrix \mathbf{A} , row reduction of $(\mathbf{A} : \mathbf{I})$ yields complete information about the equation $\mathbf{Ax} = \mathbf{y}$, including answers to the questions of existence and uniqueness of the solutions (P&C 1.1, 1.2).

Example 5. Solution by Row Reduction-a Nonsquare Matrix. Suppose we obtain the following equations from three independent measurements of some quantity

$$\xi_1 + \xi_2 = 1.2$$

$$\xi_1 + \xi_2 = 1.3$$

$$\xi_1 + \xi_2 = 1.2$$

Then

$$(\mathbf{A} : \mathbf{I}) = \begin{pmatrix} 1 & 1 & \vdots & 1 & 0 & 0 \\ 1 & 1 & \vdots & 0 & 1 & 0 \\ 1 & 1 & \vdots & 0 & 0 & 1 \end{pmatrix}$$

which we row reduce to

$$\begin{pmatrix} 1 & 1 & \vdots & 1 & 0 & 0 \\ 0 & 0 & \vdots & -1 & 1 & 0 \\ 0 & 0 & \vdots & -1 & 0 & 1 \end{pmatrix}$$

We interpret the row reduced matrix to mean

$$\xi_1 + \xi_2 = \eta_1$$

$$0 = \eta_2 - \eta_1$$

$$0 = \eta_3 - \eta_1$$

Unless $\eta_1 = \eta_2 = \eta_3$, the equations allow no solution. In our example the equations are not consistent; $\eta_1 = \eta_3 = 1.2$, but $\eta_2 = 1.3$. If the equations were consistent, the row-reduced equations indicate that the solution would not be unique; for example, if η_2 were 1.2, the solution would be

$$\xi_1 + \xi_2 = \eta_1$$

Row and Column Interpretations

We have, to this point, viewed the matrix multiplication in (1.14) as the operation of the system on \mathbf{x} to produce \mathbf{y} . This interpretation is expressed in (1.15). We now suggest two more interpretations that will be useful

throughout our discussions of modeling. It is apparent from (1.14) and (1.15) that the columns of the matrix \mathbf{A} are in some sense similar to \mathbf{y} ; they both contain the same number (m) of elements. We call them **column vectors of \mathbf{A}** , and denote the j th column vector by $\mathbf{A}_{(j)}$. Again, the rows of \mathbf{A} are similar to \mathbf{x} , both containing n elements; we denote the i th **row vector of \mathbf{A}** by $\mathbf{A}^{(i)}$. If we focus on the column vectors of \mathbf{A} , (1.14) becomes

$$\xi_1 \mathbf{A}_{(1)} + \xi_2 \mathbf{A}_{(2)} + \cdots + \xi_n \mathbf{A}_{(n)} = \mathbf{y} \quad (1.16)$$

That is, \mathbf{y} is a simple combination of the column vectors of \mathbf{A} ; the elements of \mathbf{x} specify the combination. We will make use of this column vector interpretation in Section 2.2 and thereafter.

Changing our focus to the row vectors of \mathbf{A} , (1.14) becomes

$$\begin{aligned} \mathbf{A}^{(1)} \mathbf{x} &= \eta_1 \\ \mathbf{A}^{(2)} \mathbf{x} &= \eta_2 \\ &\vdots \\ \mathbf{A}^{(m)} \mathbf{x} &= \eta_m \end{aligned} \quad (1.17)$$

Each element of \mathbf{y} is determined by the corresponding row vector of \mathbf{A} . By this interpretation, we are merely focusing separately on each of the equations of (1.13). We can use the geometrical pictures of analytic geometry to help develop a physical feel for the individual algebraic equations of (1.17). Suppose

$$\mathbf{A} \mathbf{x} = \begin{pmatrix} 2 & 1 \\ 2 & 1 + \epsilon \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad (1.18)$$

where ϵ is some constant. The 2×1 matrix \mathbf{x} and the 1×2 matrices $\mathbf{A}^{(i)}$ are each equivalent to a vector (or arrow) in a plane. We simply pick coordinate axes and associate with each element of \mathbf{x} or $\mathbf{A}^{(i)}$ a component along one of the axes. Thus we can represent (1.18) geometrically as in Figure 1.7. The vectors \mathbf{x} such that

$$\mathbf{A}^{(1)} \mathbf{x} = \text{a constant}$$

terminate on a line perpendicular to the vector $\mathbf{A}^{(1)}$. The solution \mathbf{x} to the pair of equations lies at the intersection of the lines $\mathbf{A}^{(1)} \mathbf{x} = 2$ and $\mathbf{A}^{(2)} \mathbf{x} = 3$. Since the lines in Figure 1.7 have a well-defined intersection, the equations of (1.18) possess a well-defined (unique) solution. However, if $\epsilon \rightarrow 0$, $\mathbf{A}^{(2)} \rightarrow \mathbf{A}^{(1)}$ and the system becomes degenerate; the lines become parallel, the equations become inconsistent, and there is no solution (intersection). If

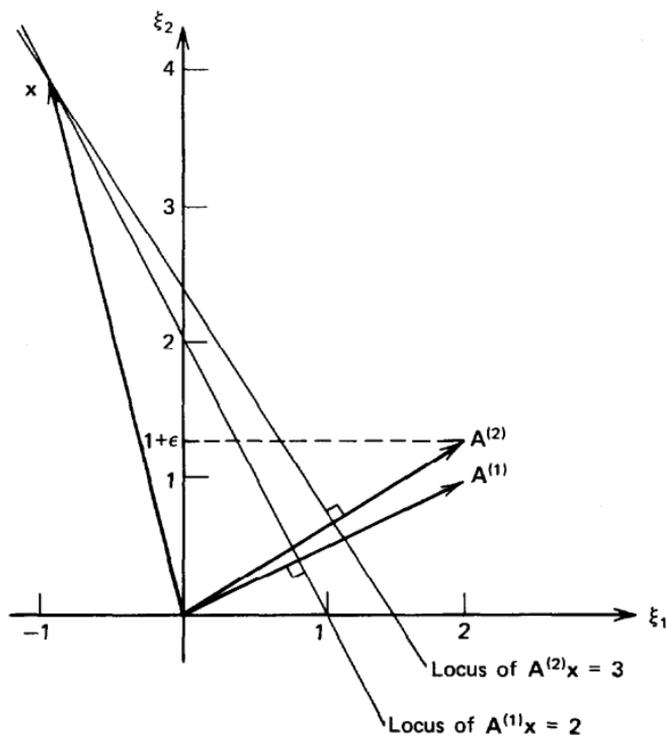


Figure 1.7. Row vector interpretation of (1.18) for $\epsilon = 0.25$.

the numbers on the right side of (1.18) were equal, the lines would overlap, the equations would be consistent, but the solution would not be unique—any \mathbf{x} terminating on the common line would satisfy both equations.

The geometrical example of (1.18) and Figure 1.7 introduces a significant computational difficulty which exists in nearly degenerate systems of equations. Slight changes in the numbers on the right side of (1.18) result in slight shifts in the positions of the lines in Figure 1.7. Slight changes in the equation coefficients cause slight tilts in these lines. If ϵ is nearly zero, the lines are nearly parallel, and slight perturbations in the line positions or angles cause large swings in the intersection (or solution) \mathbf{x} . A solution to a matrix equation which is very sensitive to small changes (or errors) in the data is called an **unstable solution**. A matrix (or the corresponding set of equations) which leads to an unstable solution is said to be **ill-conditioned**. Assume the matrix is normalized so that the magnitude of its largest element is approximately one. Then the magnitudes of the elements of the inverse matrix indicate the degree of sensitivity of the solution \mathbf{x} of (1.14) to errors in the data, $\{a_{ij}\}$ or $\{\eta_i\}$. In Section 6.6 we define a condition number which indicates the size of the largest elements of the inverse. A

very large condition number implies that the matrix is ill-conditioned. The size of $\det(\mathbf{A})$ is another indication of the ill-conditioning of the equations; as the equations become more degenerate, $\det(\mathbf{A})$ must approach zero (P&C 1.6). However, $\det(\mathbf{A})$ is not an absolute measure of ill-conditioning as is the condition number.

Numerical Error

There are two fundamental sources of error in the solution to a set of linear algebraic equations, measurement error and computer roundoff. When the data that are used to make up a set of equations come from physical measurements, these data usually contain empirical error. Even if the data are exact, however, the numbers are rounded by the computer; the data can be represented only to a finite number of significant digits. Thus inaccuracies in the equation data are the rule, not the exception. As computations are carried out, further rounding occurs. Although individual inaccuracies are slight, their cumulative effect can be disastrous if handled carelessly.

The following example demonstrates that slight errors in the data can be vastly magnified by straightforward use of row-reduction techniques. Let

$$(\mathbf{A} : \mathbf{y}) = \begin{pmatrix} 2 & 1 & 3 & \vdots & 1 \\ 2 & 1.01 & 1 & \vdots & 2 \\ 2 & 3 & 2 & \vdots & 3 \end{pmatrix} \quad (1.19)$$

Suppose the element a_{22} is in error by 0.5%; that is, $a_{22} = 1.01 \pm 0.005$. Elimination operations on the first column reduces (1.19) to

$$\begin{pmatrix} 2 & 1 & 3 & \vdots & 1 \\ 0 & 0.01 & -2 & \vdots & 1 \\ 0 & 2 & -1 & \vdots & 2 \end{pmatrix} \quad (1.20)$$

where the subtraction of two nearly equal numbers has magnified the error at the element in question to about 50%, that is, the new element in row 2, column 2, is 0.01 ± 0.005 . Were we to use this element to eliminate the other elements in column 2, we would propagate this 50% error throughout the matrix; that is, we would obtain

$$\begin{pmatrix} 2 & 0 \mp 0.5 & 203 \pm 100 & \vdots & -99 \mp 50 \\ 0 & 1 \pm 0.5 & -200 \mp 100 & \vdots & 100 \pm 50 \\ 0 & 0 \mp 1 & 399 \pm 200 & \vdots & -198 \mp 100 \end{pmatrix} \quad (1.21)$$

Further computations would be meaningless. Fortunately, we do not need to divide by the inaccurate element. We merely interchange rows 2 and 3

in (1.20) to obtain

$$\begin{pmatrix} 2 & 1 & 3 & \vdots & 1 \\ 0 & 2 & -1 & \vdots & 2 \\ 0 & 0.01 & -2 & \vdots & 1 \end{pmatrix} \quad (1.22)$$

This interchange is equivalent to writing the equations in a different order. We now use the larger and more accurate element “2” of row 2, column 2 to eliminate the other elements in column 2:

$$\begin{pmatrix} 4 & 0 & 7 & \vdots & 0 \\ 0 & 2 & -1 & \vdots & 2 \\ 0 & 0 \pm 0.005 & -1.995 & \vdots & 0.99 \end{pmatrix} \quad (1.23)$$

The element moved into position for elimination of other elements in its column is called a **pivot**. The process of interchanging rows to avoid division by relatively small (and therefore inaccurate) numbers is called **pivoting** or **positioning for size**. We also can move the inaccurate element from row 2, column 2 of (1.20) by interchanging *columns* 2 and 3 if we change the order of the variables ξ_2 and ξ_3 which multiply these columns. This column interchange is also used in pivoting. All good computer algorithms for solving sets of linear algebraic equations or for inverting square matrices use some form of pivoting to minimize the magnification and propagation of errors in the data. Scaling of the equations is also an important part of these algorithms.

The matrix of (1.19) is not ill-conditioned. It is apparent, therefore, that we must compute solutions carefully, regardless of the conditioning of the equations, if we are to avoid magnification of errors. If the equations are ill-conditioned, however, careful computing (scaling and pivoting) and the use of double precision arithmetic (additional significant digits) are crucial. Furthermore, division by small numbers is inevitable at some point in the process of solving ill-conditioned equations, and errors *will* be magnified. An iterative technique for improving the computed solution to a set of ill-conditioned equations is described in P&C 1.5.

If a set of equations is very ill-conditioned, it may be that the underlying system is degenerate. Perhaps the matrix would be singular, were it not for empirical error in the data. (That is, perhaps ϵ should be zero in (1.18) and Figure 1.7.) Then in order to completely solve the set of equations, we not only need to compute a particular solution \mathbf{x} as described above, but we also need to estimate the full set of “near solutions” (the locus of the “nearly-overlapping” lines of Figure 1.7 for $\epsilon = 0$). We describe a technique for computing this set of “near solutions” in Section 2.4. Further informa-

tion on the solution of linear algebraic equations is contained in Forsythe and Moler [1.4] and Forsythe [1.5].

1.6 Problems and Comments

*1.1 *Exploring matrix equations by row reduction:* let \mathbf{A} be an $m \times n$ matrix. Row reduction of $(\mathbf{A} \vdots \mathbf{y})$ for an unspecified column vector $\mathbf{y} = (\eta_1 \cdots \eta_m)^T$, or the equivalent row reduction of $(\mathbf{A} \vdots \mathbf{I})$ for an $m \times m$ matrix \mathbf{I} , determines the conditions which must be satisfied by \mathbf{y} in order for the equation $\mathbf{Ax} = \mathbf{y}$ to have a solution; the set of vectors \mathbf{y} for which a solution \mathbf{x} exists is called the **range of \mathbf{A}** . The same row reduction determines the set of solutions \mathbf{x} for $\mathbf{y} = (0 \cdots 0)^T$; this set of solutions is referred to as the **nullspace of \mathbf{A}** . If the nullspace of \mathbf{A} contains nonzero vectors, the solutions to $\mathbf{Ax} = \mathbf{y}$ cannot be unique. Let the matrix equation be

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 2 & 1 & 1 & 3 \\ 4 & 5 & 3 & 9 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}$$

- Row reduce $(\mathbf{A} \vdots \mathbf{I})$.
 - Determine the range of \mathbf{A} ; that is, determine the relationships that must exist among the elements $\{\eta_i\}$ of \mathbf{y} in order for the matrix equation $\mathbf{Ax} = \mathbf{y}$ to have a solution.
 - Determine the nullspace of \mathbf{A} .
 - Determine the solutions \mathbf{x} for the specified right-hand side \mathbf{y} .
 - Give an example of a matrix equation that is both inconsistent and underdetermined; that is, an equation for which \mathbf{y} is not in the range of \mathbf{A} and for which the nullspace of \mathbf{A} is nonzero.
- 1.2 Use the row-reduction technique to determine the solutions to the following sets of equations:

$$\begin{aligned} (a) \quad & \xi_1 + 6\xi_2 - 18\xi_3 = 0 \\ & -4\xi_1 \quad \quad + 5\xi_3 = 0 \\ & -3\xi_1 + 6\xi_2 - 13\xi_3 = 0 \\ & -7\xi_1 + 6\xi_2 - 8\xi_3 = 0 \end{aligned}$$

$$(b) \quad 2\xi_1 + 3\xi_2 + 4\xi_3 = 9$$

$$\xi_1 + \xi_2 + \xi_3 = 3$$

$$3\xi_1 + 2\xi_2 + 2\xi_3 = 7$$

$$(c) \quad 2\xi_1 + 3\xi_2 + 4\xi_3 = 9$$

$$3\xi_1 + 4\xi_2 + 5\xi_3 = 12$$

$$4\xi_1 + 3\xi_2 + 3\xi_3 = 10$$

$$5\xi_1 + 5\xi_2 + 6\xi_3 = 10$$

$$(d) \quad \xi_1 - 2\xi_3 = \eta_1$$

$$2\xi_1 + 2\xi_2 = \eta_2$$

$$2\xi_1 - 4\xi_3 = \eta_3$$

$$\xi_1 + \xi_2 + 3\xi_3 = \eta_4$$

1.3 *Efficiency of computations:* the number of multiplications performed during a computation is a measure of the efficiency of a computational technique. Let \mathbf{A} be an invertible $n \times n$ matrix. Determine the number of multiplications required:

(a) To compute \mathbf{A}^{-1} by Gaussian elimination;

(b) To compute \mathbf{A}^{-1} by Gauss-Jordan elimination;

(c) To compute $\det(\mathbf{A})$, using Gaussian elimination to triangularize \mathbf{A} (Example 2, Appendix 1).

Determine the number of multiplications required to solve $\mathbf{Ax} = \mathbf{y}$ for a specific vector \mathbf{y} by:

(d) Cramer's rule [Hint: use the answer to (c)].

(e) The computation in (a) and the multiplication $\mathbf{A}^{-1}\mathbf{y}$;

(f) Direct row reduction of $(\mathbf{A} : \mathbf{y})$.

1.4 *Elementary matrices:* the row reduction of an $m \times n$ matrix \mathbf{A} consists in performing elementary operations on the rows of \mathbf{A} . Each such operation is equivalent to the multiplication of \mathbf{A} by a simple $m \times m$ matrix which we refer to as an **elementary matrix**.

(a) For $m = 5$, find the elementary matrices corresponding to the following:

(1) the multiplication of row 3 by a constant c ;

(2) the addition of row 4 to row 1;

(3) the interchange of row 3 with row 5.

(b) Every elementary matrix is invertible. Find the inverses of the elementary matrices determined in (a).

(c) The row reduction of $(\mathbf{A} : \mathbf{I})$ is equivalent to multiplication of $(\mathbf{A} : \mathbf{I})$ by an invertible matrix \mathbf{B} (a product of elementary matrices). Show that if \mathbf{A} is square and $(\mathbf{A} : \mathbf{I})$ can be row reduced to the form $(\mathbf{I} : \mathbf{B})$, then $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, and therefore $\mathbf{B} = \mathbf{A}^{-1}$.

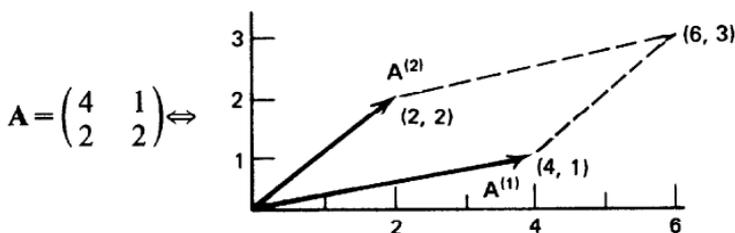
- 1.5 *Iterative improvement of solutions:* the solution to the matrix equation $\mathbf{Ax} = \mathbf{y}$ can be obtained by Gaussian elimination. As a result of roundoff, the computed solution \mathbf{x}_1 is usually in error. Denote the error by $\mathbf{x} - \mathbf{x}_1$, where \mathbf{x} is the exact solution. A computable measure of the error is the residual $\mathbf{r}_1 \stackrel{\Delta}{=} \mathbf{y} - \mathbf{Ax}_1$. If we could solve exactly for $(\mathbf{x} - \mathbf{x}_1)$ in the equation $\mathbf{A}(\mathbf{x} - \mathbf{x}_1) = \mathbf{y} - \mathbf{Ax}_1 = \mathbf{r}_1$, we could obtain the exact solution. We solve $\mathbf{Az}_1 = \mathbf{r}_1$ by Gaussian elimination to obtain a correction \mathbf{z}_1 ; $\mathbf{x}_2 \stackrel{\Delta}{=} \mathbf{x}_1 + \mathbf{z}_1$ is an improved solution. By repeating the improvement process iteratively, we obtain an approximate solution which is accurate to the number of significant digits used in the computation. However, the residuals $\mathbf{r}_k = \mathbf{y} - \mathbf{Ax}_k$ must be computed to double precision; otherwise the corrections, \mathbf{z}_k , will not be improvements. See Forsythe and Moler [1.4, p. 49]. Let

$$\mathbf{A} = \begin{pmatrix} 2.1 & 1.9 \\ 1.9 & 2.0 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 1.2 \\ 1.3 \end{pmatrix}$$

To five figures, the solution to $\mathbf{Ax} = \mathbf{y}$ is $\mathbf{x} = (-0.11864 \ 0.76271)^T$.

- (a) Compute an approximate solution \mathbf{x}_1 by Gaussian elimination, rounding all computations to three significant digits (slide rule accuracy).
- (b) Find the residual \mathbf{r}_1 by hand computation to *full* accuracy.
- (c) Round \mathbf{r}_1 to three significant digits, if necessary, and compute the correction \mathbf{z}_1 . Find $\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{z}_1$.

- 1.6 *Determinants and volumes:* using a natural correspondence between row vectors and arrows in a plane, we associate a parallelogram with the rows of every real 2×2 matrix \mathbf{A} . For example,



- (a) Show that the area of the above parallelogram is equal to the determinant of the matrix \mathbf{A} which is associated with it.
- (b) For the right-hand coordinate system shown above, we define the *sign of the area* to be positive if $\mathbf{A}^{(1)}$ turns counterclockwise inside the parallelogram in order to reach $\mathbf{A}^{(2)}$; if $\mathbf{A}^{(1)}$ turns clockwise, the area is negative. Show graphically that the area of the above parallelogram obeys the following properties of determinants:
- (1) The value of $\det(\mathbf{A})$ is not changed if we add to one row of \mathbf{A} a multiple of another row of \mathbf{A} ;
 - (2) The sign of $\det(\mathbf{A})$ is reversed if we interchange two rows of \mathbf{A} ;
 - (3) If we multiply one row of \mathbf{A} by c , then $\det(\mathbf{A})$ is multiplied by c ;
 - (4) If the rows of \mathbf{A} are dependent (i.e., one is a multiple of the other), then $\det(\mathbf{A}) = 0$.
- (c) The geometrical interpretation of $\det(\mathbf{A})$ can be extended to $n \times n$ matrices by defining n -dimensional spaces, n -dimensional parallelepipeds, and signed volumes. See Martin and Mizel [1.9]. Since $\det(\mathbf{A}^T) = \det(\mathbf{A})$, the volume of the parallelepiped described by the columns of \mathbf{A} equals the volume described by the rows of \mathbf{A} . Verify graphically that the geometrical interpretation of determinants extends to 3×3 matrices.

1.7 *Partitioned matrices*: it is sometimes useful to partition a matrix into an array of submatrices. If \mathbf{P} and \mathbf{Q} are conformable, we can form the partitions

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix}$$

in a manner which allows us to express \mathbf{PQ} as

$$\mathbf{PQ} = \begin{pmatrix} \mathbf{P}_{11}\mathbf{Q}_{11} + \mathbf{P}_{12}\mathbf{Q}_{21} & \mathbf{P}_{11}\mathbf{Q}_{12} + \mathbf{P}_{12}\mathbf{Q}_{22} \\ \mathbf{P}_{21}\mathbf{Q}_{11} + \mathbf{P}_{22}\mathbf{Q}_{21} & \mathbf{P}_{21}\mathbf{Q}_{12} + \mathbf{P}_{22}\mathbf{Q}_{22} \end{pmatrix}$$

(a) Assume that \mathbf{A} is an invertible matrix. The following factorization can be verified by the block multiplication described above:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{C}\mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix}$$

(b) Show that for any submatrix \mathbf{P} of appropriate dimensions,

$$\begin{vmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{P} & \mathbf{I} \end{vmatrix} = 1$$

Use this result with (a) to show that

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{A}| |\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}|$$

(c) Use (a) to show that

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{O} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{O} \\ \mathbf{O} & (\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{O} \\ -\mathbf{CA}^{-1} & \mathbf{I} \end{pmatrix}$$

The number of multiplications required to compute the determinant or the inverse of an $n \times n$ matrix can be reduced by a factor of eight (if n is large) by use of the partitioning schemes in (b) or (c), respectively.

1.7 References

- [1.1] Bruner, Jerome S., *The Process of Education*, Harvard University Press, Cambridge, Mass., 1960.
- [1.2] Cannon, Robert H., Jr., *Dynamics of Physical Systems*, McGraw-Hill, New York, 1969.
- [1.3] Forrester, Jay W., *Urban Dynamics*, M.I.T. Press, Cambridge, Mass., 1969.
- *[1.4] Forsythe, George E. and Cleve B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, N.J., 1967.
- [1.5] Forsythe, George E., "Today's Computational Methods of Linear Algebra," *SIAM Rev.*, **9**, 3 (July 1967), 489-515.
- [1.6] Forsythe, George E., "Solving Linear Algebraic Equations Can be Interesting," *Bull. Am. Math. Soc.*, **59** (1953), 299-329.
- [1.7] Linvill, William K., "Models and Model Construction," *IRE Trans. Educ.*, **E-5**, 2 (June 1962), 64-67.
- [1.8] Meier, Robert C., William T. Newell, and Harold L. Pazer, *Simulation in Business and Economics*, Prentice-Hall, Englewood Cliffs, N.J., 1969.
- [1.9] Martin, Allan D. and Victor J. Mizel, *Introduction to Linear Algebra*, McGraw-Hill, New York, 1966.
- [1.10] Sage, Andrew P. and James L. Melsa, *System Identification*, Academic Press, New York, 1971.
- [1.11] Truxal, John G., *Introduction to Systems Engineering*, McGraw-Hill, New York, 1972.

System Models : Transformations on Vector Spaces

The fundamental purpose in modeling a system is to develop a mechanism for predicting the condition or change in condition of the system. In the abstract model $\mathbf{T}\mathbf{x} = \mathbf{y}$ of (1.1), \mathbf{T} represents (or is a model of) the system, whereas \mathbf{x} and \mathbf{y} have to do with the condition of the system. We explore first some familiar models for the condition or changes in condition of systems. These examples lead us to use a generalization of the usual notion of a vector as a model for the condition of a system. We then develop the concept of a transformation of vectors as a model of the system itself. The rest of the chapter is devoted to examination of the most commonly used models-linear models-and their matrix representations.

2.1 The Condition of a System

The physical condition (or change in condition) of many simple systems has been found to possess a magnitude and a direction in our physical three-dimensional space. It is natural, therefore, that a mathematical concept of condition (or change in condition) has developed over time which has these two properties; this concept is the vector. Probably the most obvious example of the use of this concept is the use of arrows in a two-dimensional plane to represent changes in the position of an object on the two-dimensional surface of the earth (see Figure 2.1). Using the usual techniques of analytic geometry, we can represent each such arrow by a pair of numbers that indicates the components of that arrow along each of a pair of coordinate axes. Thus pairs of numbers serve as an equivalent model for changes in position.

An ordinary road map is another model for the two-dimensional surface of the earth. It is equivalent to the arrow diagram; points on the map are

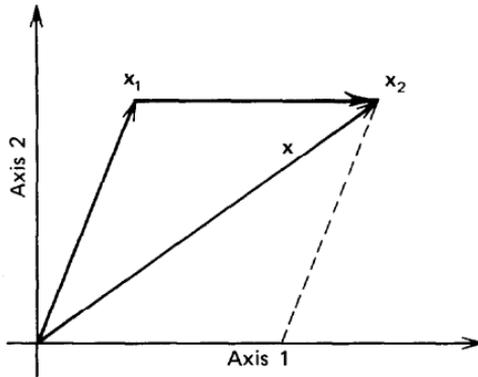


Figure 2.1. An "arrow vector" diagram.

equivalent to the arrow tips of Figure 2.1. The only significant difference between these two models is that the map emphasizes the position (or condition) of an object on the earth, whereas the arrow diagram stresses the changes in position and the manner in which intermediate changes in position add to yield a total change in position. We can also interpret a position on the map as a change from some reference position. The manner in which we combine arrows or changes in position (the parallelogram rule) is the most significant characteristic of either model. Consequently we focus on the arrow model which emphasizes the combination process.

Reference arrows (coordinate axes) are used to tie the arrow model to the physical world. By means of a reference position and a pair of reference "position changes" on the surface of the earth, we relate the positions and changes in position on the earth to positions and arrows in the arrow diagram. However, there are no inherent reference axes on either the physical earth or the two-dimensional plane of arrows.

The same vector model that we use to represent changes in position can be used to represent the forces acting at a point on a physical object. The reason we can use the same model is that the magnitudes and directions of forces also combine according to the parallelogram rule. The physical natures of the reference vectors are different in these three situations: in one case they are changes in position on the earth, in another they are arrows, in the third, forces. Yet once reference vectors are chosen in each, all three situations become in some sense equivalent; corresponding to each vector in one situation is a vector in the other two; corresponding to each sum of vectors in one is a corresponding sum in the other two. We use the set of arrows as a model for the other two situations because it is the most convenient of the three to work with.

The set of complex numbers is one more example of a set of objects which is equivalent to the set of arrows. We usually choose as references in

the set of complex numbers the two numbers 1 and i . Based on these reference numbers and two reference arrows, we interpret every arrow as a complex number. Here we have one set of mathematical (or geometrical) objects serving as a model for another set of mathematical objects.

Consider now a physical system which is more complicated than the two physical systems discussed above. Imagine a flat metal sheet exposed to the sun and partly submerged in a stream. (The sheet is representative of any object subject to heat sources and coolants.) The thermal condition of the sheet is described by the temperature distribution over the surface of the sheet. A change in the cloud cover in the sky will change the pattern in which the sun falls on the sheet. As a result, the temperature distribution will change. Assuming the temperature distribution reaches a new steady state, the new distribution equals the old distribution plus the change in the distribution. We model this situation as follows. Let (s, t) denote a position in some two-dimensional coordinate system on the surface of the sheet. Let $\mathbf{f}(s, t)$ be the temperature at the point (s, t) , measured in degrees centigrade, for all points (s, t) on the sheet. We model a change in the thermal condition of the sheet by

$$\mathbf{f}_{\text{new}}(s, t) = \mathbf{f}_{\text{old}}(s, t) + \mathbf{f}_{\text{change}}(s, t) \quad (2.1)$$

for all (s, t) on the sheet. In effect, (2.1) defines $\mathbf{f}_{\text{change}}$. However, we hope to use a model of the system to *predict* $\mathbf{f}_{\text{change}}$. Then (2.1) will determine \mathbf{f}_{new} . Equation (2.1) is a “distributed” equivalent of the arrow diagram in Figure 2.1; each of these models illustrates the manner in which changes in condition combine to yield a net condition of the system in question. Once again, references have been chosen in both the physical system and the model (mathematical system) in order to equate the two systems; choosing physical units of measurement (degrees centigrade) amounts to fixing the relationship between the physical and mathematical systems.

The most significant difference between a system modeled by Figure 2.1 and a system modeled by (2.1) consists in the nature of the conditions in each system. In one case we have a quantity with magnitude and direction (e.g., force); in the other, a quantity without magnitude and direction—a quantity that is distributed over a two-dimensional region. Yet there are important similarities between the two systems. The changes in condition of the system are under scrutiny; also, several changes in condition combine by simple rules to yield a total or net condition.

Vector Spaces

By expressing various types of problems in a common framework, we learn to use concepts derived from one type of problem in understanding other types of problems. In particular, we are able to draw useful analogies

between algebraic equations and differential equations by expressing both types of equations as “vector” equations. Therefore, we now generalize the common notion of a vector to include all the examples discussed in the previous section.

Definition. A **linear space** (or **vector space**) \mathcal{V} is a set of elements \mathbf{x} , \mathbf{y} , \mathbf{z}, \dots , called vectors, together with definitions of *vector addition* and *scalar multiplication*.

- a. The definition of vector addition is such that:
 1. To every pair, \mathbf{x} and \mathbf{y} , of vectors in \mathcal{V} there corresponds a unique vector $\mathbf{x} + \mathbf{y}$ in \mathcal{V} , called the **sum** of \mathbf{x} and \mathbf{y} .
 2. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
 3. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
 4. There is a unique vector $\boldsymbol{\theta}$ in \mathcal{V} , called the **zero vector** (or origin), such that $\mathbf{x} + \boldsymbol{\theta} = \mathbf{x}$ for all \mathbf{x} in \mathcal{V} .
 5. Corresponding to each \mathbf{x} in \mathcal{V} there is a unique vector “ $-\mathbf{x}$ ” in \mathcal{V} such that $\mathbf{x} + (-\mathbf{x}) = \boldsymbol{\theta}$.
- b. The definition of scalar multiplication is such that:
 1. To every vector \mathbf{x} in \mathcal{V} and every scalar a there corresponds a unique vector $a\mathbf{x}$ in \mathcal{V} , called the **scalar multiple** of \mathbf{x} .*
 2. $a(b\mathbf{x}) = (ab)\mathbf{x}$.
 3. $1(\mathbf{x}) = \mathbf{x}$ (where 1 is the unit scalar).
 4. $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$.
 5. $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$.

Notice that a vector space includes not only a set of elements (vectors) but also “valid” definitions of vector addition and scalar multiplication. Also inherent in the definition is the fact that the vector space \mathcal{V} contains all “combinations” of its own vectors: if \mathbf{x} and \mathbf{y} are in \mathcal{V} , then $a\mathbf{x} + b\mathbf{y}$ is also in \mathcal{V} . The rules of algebra are so much a part of us that some of the requirements may at first appear above definition; however, they are necessary. A few more vector space properties which may be deduced from the above definition are as follows:

1. $0\mathbf{x} = \boldsymbol{\theta}$ (where “0” is the zero scalar).
2. $a\boldsymbol{\theta} = \boldsymbol{\theta}$.
3. $(-1)\mathbf{x} = -\mathbf{x}$.

Example 1. The **Real 3-tuple Space** \mathcal{R}^3 . The space \mathcal{R}^3 consists in the set of all

*The scalars are any set of elements which obey the usual rules of algebra. A set of elements which obeys these rules constitutes a field (see Hoffman and Kunze [2.6]). We usually use as scalars either the real numbers or the complex numbers. There are other useful fields, however (P&C 2.4).

real 3-tuples (all ordered sequences of three real numbers), $\mathbf{x} = (\xi_1, \eta_1, \zeta_1)$, $\mathbf{y} = (\eta_1, \eta_2, \eta_3)$, with the following definitions of addition and scalar multiplication:

$$\begin{aligned} \mathbf{x} + \mathbf{y} &\stackrel{\Delta}{=} (\xi_1 + \eta_1, \xi_2 + \eta_2, \xi_3 + \eta_3) \\ a\mathbf{x} &\stackrel{\Delta}{=} (a\xi_1, a\xi_2, a\xi_3) \end{aligned} \tag{2.2}$$

It is clear that the zero vector for this 3-tuple space, $\mathbf{0} = (0,0,0)$, satisfies $\mathbf{x} + \mathbf{0} = \mathbf{x}$. We show that $\mathbf{0}$ is unique by assuming another vector \mathbf{y} also satisfies $\mathbf{x} + \mathbf{y} = \mathbf{x}$; that is,

$$(\xi_1 + \eta_1, \xi_2 + \eta_2, \xi_3 + \eta_3) = (\xi_1, \xi_2, \xi_3)$$

or $\xi_i + \eta_i = \xi_i$. The properties of scalars then require $\eta_i = 0$ (or $\mathbf{y} = \mathbf{0}$). It is easy to prove that \mathcal{R}^3 , as defined above, satisfies the other requirements for a linear space. In each instance, questions about vectors are reduced to questions about scalars.

We emphasize that the definition of \mathcal{R}^3 says nothing about coordinates. Coordinates are multipliers for reference vectors (reference arrows, for instance). The 3-tuples are vectors in their own right. However, there is a commonly used correspondence between \mathcal{R}^3 and the set of vectors (arrows) in the usual three-dimensional space which makes it difficult not to think of the 3-tuples as coordinates. The two sets of vectors are certainly equivalent. We will, in fact, use this natural correspondence to help illustrate vector concepts graphically.

Example 2. The Two-Dimensional Space of Points (or Arrows). This space consists in the set of all points in a plane. Addition is defined by the parallelogram rule using a fixed reference point (see Figure 2.2). Scalar multiplication is defined as "length" multiplication using the reference point. The zero vector is obviously the

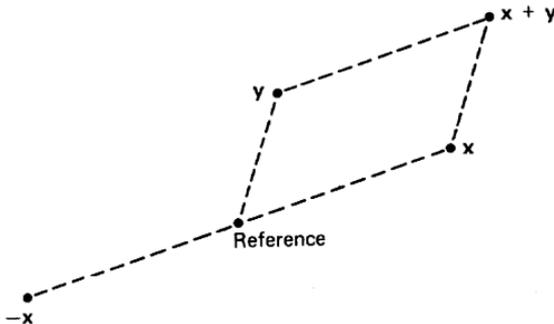


Figure 2.2. The two-dimensional space of points.

reference point. Each of the requirements can be verified by geometrical arguments.

An equivalent (but not identical) space is one where the vectors are not the points, but rather, arrows to the points from the reference point. We distinguish only the magnitude and direction of each arrow; *two parallel arrows of the same length are considered identical*.

Both the arrow space and the point space are easily visualized: we often use the arrow space in two or three dimensions to demonstrate concepts graphically. Although the arrow space contains no *inherent* reference arrows, we sometimes *specify* reference arrows in order to equate the arrows to vectors in \mathcal{R}^3 . Because of the equivalence between vectors in \mathcal{R}^3 and vectors in the three-dimensional space of points, we occasionally refer to vectors in \mathcal{R}^3 and in other spaces as *points*.

Example 3. The Space of Column Vectors $\mathcal{N}^{3 \times 1}$. The space $\mathcal{N}^{3 \times 1}$ consists in the set of all real 3x1 column matrices (or column vectors), denoted by

$$\mathbf{x} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}$$

with the following definitions of addition and scalar multiplication:

$$\mathbf{x} + \mathbf{y} \triangleq \begin{pmatrix} \xi_1 + \eta_1 \\ \xi_2 + \eta_2 \\ \xi_3 + \eta_3 \end{pmatrix} \quad a\mathbf{x} \triangleq \begin{pmatrix} a\xi_1 \\ a\xi_2 \\ a\xi_3 \end{pmatrix} \quad (2.3)$$

In order to save space in writing, we occasionally write vectors from $\mathcal{N}^{3 \times 1}$ in the transposed form $\mathbf{x} = (\xi_1 \ \xi_2 \ \xi_3)^T$. The equivalence between $\mathcal{N}^{3 \times 1}$ and \mathcal{R}^3 is obvious. The only difference between the two vector spaces is in the nature of their vectors. Vectors in $\mathcal{N}^{3 \times 1}$ can be multiplied by $m \times 3$ matrices (as in Section 1.5), whereas vectors in \mathcal{R}^3 cannot.

Example 4. The Space of Real Square-Summable Sequences, l_2 . The space l_2 consists in the set of all infinite sequences of real numbers, $\mathbf{x} = (\xi_1, \xi_2, \xi_3, \dots)$, $\mathbf{y} = (\eta_1, \eta_2, \eta_3, \dots)$ which are square summable; that is, for which $\sum_{i=1}^{\infty} \xi_i^2 < \infty$. Addition and scalar multiplication in l_2 are defined by

$$\mathbf{x} + \mathbf{y} \triangleq (\xi_1 + \eta_1, \xi_2 + \eta_2, \xi_3 + \eta_3, \dots) \\ a\mathbf{x} \triangleq (a\xi_1, a\xi_2, a\xi_3, \dots) \quad (2.4)$$

Most of the properties required by the definition of a linear space are easily verified for l_2 ; for instance, the zero vector is obviously $\mathbf{0} = (0,0,0, \dots)$. However, there is one subtle difference between l_2 and the space \mathcal{R}^3 of Example 1. Because

the sequences in l_2 are infinite, it is not obvious that if \mathbf{x} and \mathbf{y} are in l_2 , $\mathbf{x} + \mathbf{y}$ is also in l_2 . It can be shown that

$$\sqrt{\sum_{i=1}^{\infty} (\xi_i + \eta_i)^2} \leq \sqrt{\sum_{i=1}^{\infty} \xi_i^2} + \sqrt{\sum_{i=1}^{\infty} \eta_i^2}$$

[This fact is known as the triangle inequality (P&C 5.4)]. Therefore,

$$\sum_{i=1}^{\infty} (\xi_i + \eta_i)^2 < \infty$$

and $\mathbf{x} + \mathbf{y}$ is square-summable. The requirement of square summability is a definite restriction on the elements of l_2 ; the simple sequence (1, 1, 1, . . .), for instance, is not in l_2 .

The definition of \mathcal{R}^3 extends easily to \mathcal{R}^n , the space of n -tuples of real numbers (where n is a positive integer). The space $\mathcal{N}^{n \times 1}$ is a similar extension of $\mathcal{N}^3 \times 1$. Mathematically these “ n -dimensional” spaces are no more complicated than their three-dimensional counterparts. Yet we are not able to draw arrow-space equivalents because our physical world is three-dimensional. Visualization of an abstract vector space is most easily accomplished by thinking in terms of its three-dimensional counterpart.

The spaces \mathcal{R}^n , $\mathcal{N}^{n \times 1}$, and l_2 can also be redefined using complex numbers, rather than real numbers, for scalars. We denote by \mathcal{R}_c^n the complex n -tuple space. We use the symbol $\mathcal{N}_c^{n \times 1}$ for the space of complex $n \times 1$ column vectors. Let l_2^c represent the space of complex square-summable sequences. (We need a slightly different definition of square summability for the space $l_2^c: \sum_{i=1}^{\infty} |\xi_i|^2 < \infty$). In most vector space definitions, either set of scalars can be used. A notable exception to interchangeability of scalars is the arrow space in two or three dimensions. The primary value of the arrow space is in graphical illustration. We have already discussed the equivalence of the set of complex scalars to the two-dimensional space of arrows. Therefore, substituting complex scalars in the real two-dimensional arrow space would require four-dimensional graphical illustration.

We eventually find it useful to combine simple vector spaces to form more complicated spaces.

Definition. Suppose \mathcal{V} and \mathcal{W} are vector spaces. We define the **Cartesian product** $\mathcal{V} \times \mathcal{W}$ of the spaces \mathcal{V} and \mathcal{W} to be the set of pairs of vectors $\mathbf{z} \stackrel{\Delta}{=} (\mathbf{x}, \mathbf{y})$, with \mathbf{x} in \mathcal{V} and \mathbf{y} in \mathcal{W} . We define addition and scalar multiplication of vectors in $\mathcal{V} \times \mathcal{W}$ in terms of the corresponding operations in \mathcal{V} and in \mathcal{W} : if $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{y}_1)$ and $\mathbf{z}_2 = (\mathbf{x}_2, \mathbf{y}_2)$, then

$$\mathbf{z}_1 + \mathbf{z}_2 \stackrel{\Delta}{=} (\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}_1 + \mathbf{y}_2)$$

$$a\mathbf{z}_1 \stackrel{\Delta}{=} (a\mathbf{x}_1, a\mathbf{y}_1)$$

Example 5. A Cartesian Product. Let $\mathbf{x} = (\xi_1, \xi_2)$, a vector in \mathcal{R}^2 . Let $\mathbf{y} = (\eta_1)$, a vector in \mathcal{R}^1 . Then $\mathbf{z} \triangleq ((\xi_1, \xi_2), (\eta_1))$ is a typical vector in $\mathcal{R}^2 \times \mathcal{R}^1$. This Cartesian product space is clearly equivalent to \mathcal{R}^3 . Strictly speaking, however, \mathbf{z} is not in \mathcal{R}^3 . It is not a 3-tuple, but rather a 2-tuple followed by a 1-tuple. Yet we have no need to distinguish between \mathcal{R}^3 and $\mathcal{R}^2 \times \mathcal{R}^1$.

Function Spaces

Each vector in the above examples has discrete elements. It is a small conceptual step from the notion of an infinite sequence of discrete numbers (a vector in l_2) to the usual notion of a function—a “continuum” of numbers. Yet vectors and functions are seldom related in the thinking of engineers. We will find that vectors and functions can be viewed as essentially equivalent objects; functions can be treated as vectors, and vectors can be treated as functions. A **function space** is a linear space whose elements are functions. We usually think of a function as a rule or graph which associates with each scalar in its domain a single scalar value. We do not confuse the graph with particular values of the function. Our notation should also keep this distinction. Let \mathbf{f} denote a **function**; that is, the symbol \mathbf{f} recalls to mind a particular rule or graph. Let $\mathbf{f}(t)$ denote the **value of the function at t** . By $\mathbf{f} = \mathbf{g}$, we mean that the scalars $\mathbf{f}(t)$ and $\mathbf{g}(t)$ are equal for each t of interest.

Example 6. \mathcal{P}^n , The Polynomials of Degree Less Than n . The space \mathcal{P}^n consists in all real-valued polynomial functions of degree less than n : $\mathbf{f}(t) = \xi_1 + \xi_2 t + \cdots + \xi_n t^{n-1}$ for all real t . Addition and scalar multiplication of vectors (functions) in \mathcal{P}^n are defined by

$$\begin{aligned}(\mathbf{f} + \mathbf{g})(t) &\triangleq \mathbf{f}(t) + \mathbf{g}(t) \\ (\mathbf{a}\mathbf{f})(t) &\triangleq a(\mathbf{f}(t))\end{aligned}\tag{2.5}$$

for all t . The zero function is $\mathbf{0}(t) = 0$ for all t . This zero function is unique; if the function \mathbf{g} also satisfied $\mathbf{f} + \mathbf{g} = \mathbf{f}$, then the values of \mathbf{f} and \mathbf{g} would satisfy

$$(\mathbf{f} + \mathbf{g})(t) = \mathbf{f}(t) + \mathbf{g}(t) = \mathbf{f}(t)$$

It would follow that $\mathbf{g}(t) = 0$ for all t , or $\mathbf{g} = \mathbf{0}$. The other requirements for a vector space are easily verified for \mathcal{P}^n .

We emphasize that the vector \mathbf{f} in Example 6 is the entire portrait of the function \mathbf{f} . The scalar variable t is a “dummy” variable. The only purpose of this variable is to order the values of the function in precisely the same way that the subscript i orders the elements in the following vector from l_2 :

$$\mathbf{x} = (\xi_1, \xi_2, \dots, \xi_i, \dots)$$

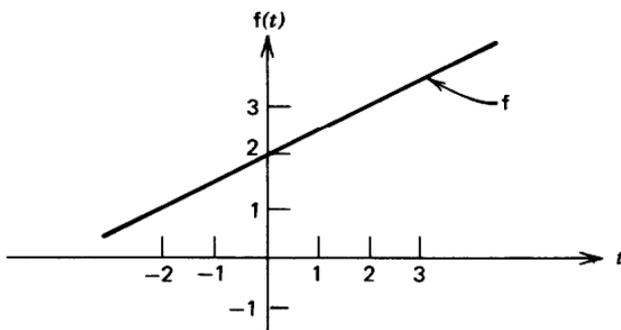


Figure 2.3. A function \mathbf{f} and its values $\mathbf{f}(t)$.

Figure 2.3 distinguishes graphically between the vector \mathbf{f} and its value at t for the specific function \mathbf{f} defined by $\mathbf{f}(t) = 2 + 0.5t$. Figure 2.4 distinguishes in a similar manner between an infinite sequence \mathbf{x} and its i th element.

It is evident that the vector \mathbf{x} from l_2 is just as much a function as is the polynomial \mathbf{f} from \mathcal{P}^n . In the space of polynomials, the index t is continuous; in the space of infinite sequences the index i is discrete—it takes on only positive integral values. In the latter case, we could as well refer to the i th element ξ_i as the value of \mathbf{x} at i . In point of fact, most vector spaces can be interpreted as spaces of functions; the terms vector space and function space are somewhat interchangeable. However, it is common practice to use the term function space only for a space in which the index t varies continuously over an interval.

It is unfortunate that the symbol $\mathbf{f}(t)$ is commonly used to represent both a function and the value of that function at t . This blurring of the meaning of symbols is particularly true of the sinusoidal and exponential functions. We will try to be explicit in our distinction between the two concepts. As discussed in the preface, boldface type is used to *emphasize* the interpretation of a function as a vector. However, to avoid overuse of boldface type, it is not used where emphasis on the vector interpretation appears un-

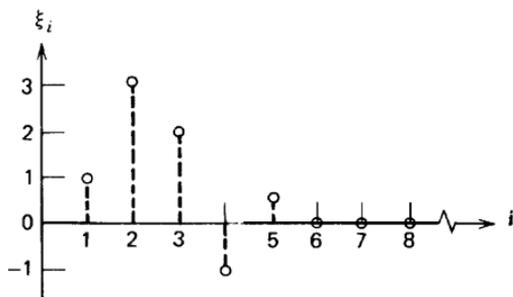


Figure 2.4. The elements ξ_i of an infinite sequence \mathbf{x} .

necessary; thus the value of a function \mathbf{f} at t may appear either as $\mathbf{f}(t)$ or as $f(t)$. Furthermore, where confusion is unlikely, we sometimes use standard mathematical shorthand; for example, we use $\int_a^b \mathbf{f} \mathbf{g} dt$ to mean $\int_a^b \mathbf{f}(t) \mathbf{g}(t) dt$.

It is difficult to describe or discuss functions in any detail except in terms of their scalar values. In Example 6, for instance, the definitions of addition and scalar multiplication were given in terms of function values. Furthermore, we resorted again to function values to verify that the vector space requirements were met. We will find ourselves continually reducing questions about functions to questions about the scalar values of those functions. Why then do we emphasize the function \mathbf{f} rather than the value $\mathbf{f}(t)$? Because system models act on the whole vector \mathbf{f} rather than on its individual values. As an example, we turn to the one system model we have explored thus far—the matrix equation $\mathbf{Ax} = \mathbf{y}$ which was introduced in Section 1.5. If \mathbf{A} is an $m \times n$ matrix, the vector \mathbf{x} is a column matrix in $\mathfrak{N}^{n \times 1}$; \mathbf{y} is in $\mathfrak{N}^{m \times 1}$. Even though the matrix multiplication requires manipulation of the individual elements (or values) of \mathbf{x} , it is impossible to determine *any* element of \mathbf{y} without operating on *all* elements of \mathbf{x} . Thus it is natural to think in terms of \mathbf{A} operating on the whole vector \mathbf{x} . Similarly, equations involving functions require operations on the whole function (e.g., integration), as we shall see in Section 2.3.

Example 7. The Space $\mathcal{C}(\mathbf{a}, \mathbf{b})$ of Continuous Functions. The vectors in \mathcal{C} are those real functions which are defined and continuous on the interval $[\mathbf{a}, \mathbf{b}]$. Addition and scalar multiplication of functions in $\mathcal{C}(\mathbf{a}, \mathbf{b})$ are defined by the standard function space definitions (2.5) for all t in $[\mathbf{a}, \mathbf{b}]$. It is clear that the sums and scalar multiples of continuous functions are also continuous functions.

Example 8. $\mathcal{L}_2(\mathbf{a}, \mathbf{b})$ The Real Square-integrable Functions. The space $\mathcal{L}_2(\mathbf{a}, \mathbf{b})$ consists in all real functions which are defined and square integrable on the interval $[\mathbf{a}, \mathbf{b}]$; that is, functions \mathbf{f} for which*

$$\int_a^b \mathbf{f}^2(t) dt < \infty$$

Addition and scalar multiplication of functions in $\mathcal{L}_2(\mathbf{a}, \mathbf{b})$ are defined by (2.5) for all t in $[\mathbf{a}, \mathbf{b}]$. The space $\mathcal{L}_2(\mathbf{a}, \mathbf{b})$ is analogous to l_2 . It is not clear that the sum of two square-integrable functions is itself square integrable. As in Example 4, we must rely on P&C 5.4 and the concepts of Chapter 5 to find that

$$\sqrt{\int_a^b [\mathbf{f}(t) + \mathbf{g}(t)]^2 dt} < \sqrt{\int_a^b \mathbf{f}^2(t) dt} + \sqrt{\int_a^b \mathbf{g}^2(t) dt}$$

*The integral used in the definition of $\mathcal{L}_2(\mathbf{a}, \mathbf{b})$ is the Lebesgue integral. For all practical purposes, Lebesgue integration can be considered the same as the usual Riemann integration. Whenever the Riemann integral exists, it yields the same result as the Lebesgue integral. (See Royden [2.1].)

It follows that if \mathbf{f} and \mathbf{g} are square integrable, then $\mathbf{f} + \mathbf{g}$ is square integrable.

Example 9. A Set of Functions. The set of positive real functions [together with the definitions of addition and scalar multiplication in (2.5)] does *not* form a vector space. This set contains a positive valued function \mathbf{f} , but not the negative valued function $-\mathbf{f}$; therefore, this set does not include all sums and multiples of its members.

Example 10. Functions of a Complex Variable. Let \mathcal{V} be the space of all complex functions \mathbf{w} of the complex variable z which are defined and analytic on some region Ω of the complex z plane.* For instance, Ω might be the circle $|z| < 1$. We define addition and scalar multiplication of functions in \mathcal{V} by

$$\begin{aligned} (\mathbf{w}_1 + \mathbf{w}_2)(z) &\triangleq \mathbf{w}_1(z) + \mathbf{w}_2(z) \\ (a\mathbf{w})(z) &\triangleq a(\mathbf{w}(z)) \end{aligned} \tag{2.6}$$

for all z in Ω . In this example, the zero vector θ is defined by $\theta(z) = \mathbf{0}$ for all z in Ω . (We do not care about the values of the functions θ and \mathbf{w} outside of Ω .)

Exercise 1. Show that if \mathbf{w}_1 and \mathbf{w}_2 are in the space \mathcal{V} of Example 10, then $\mathbf{w}_1 + \mathbf{w}_2$ is also in \mathcal{V} .

Example 11. A Vector Space of Random Variables † A **random variable** \mathbf{x} is a numerical-valued function whose domain consists in the possible outcomes of an experiment or phenomenon. Associated with the experiment is a probability distribution. Therefore, there is a probability distribution associated with the values of the random variable. For example, the throwing of a single die is an experiment. We define the random variable \mathbf{x} in terms of the possible outcomes σ by

$$\begin{aligned} \mathbf{x}(\sigma) &\triangleq \mathbf{0} \quad \text{for } \sigma = 2,4,6 \text{ (the die is even)} \\ &\triangleq \mathbf{1} \quad \text{for } \sigma = 1,3,5 \text{ (the die is odd)} \end{aligned}$$

The probability mass function ω associated with the outcome σ of the experiment is given by

$$\omega(\sigma) = \frac{1}{6} \quad \text{for } \sigma = 1,2,3,4,5,6$$

*Express the complex variable z in the form $s + it$, where s and t are real. Let the complex function \mathbf{w} be written as $\mathbf{u} + i\mathbf{v}$, where $\mathbf{u}(z)$ and $\mathbf{v}(z)$ are real. Then \mathbf{w} is analytic in Ω if and only if the partial derivatives of \mathbf{u} and \mathbf{v} are continuous and satisfy the Cauchy-Riemann conditions in Ω :

$$\frac{\partial \mathbf{u}(z)}{\partial s} = \frac{\partial \mathbf{v}(z)}{\partial t}, \quad \frac{\partial \mathbf{v}(z)}{\partial s} = -\frac{\partial \mathbf{u}(z)}{\partial t}$$

For instance, $\mathbf{w}(z) \triangleq z^2$ is analytic in the whole z plane. See Wylie [2.11].

† See Papoulis [2.7], or Cramér and Leadbetter [2.2].

Then the probability mass function $\omega_{\mathbf{x}}$ associated with the values of the random variable \mathbf{x} is

$$\omega_{\mathbf{x}}(x) = \frac{1}{2} \quad \text{for } x = 0, 1$$

We can define many other random variables (functions) for the same die-throwing experiment. One other random variable is

$$\begin{aligned} \mathbf{y}(\sigma) &\stackrel{\Delta}{=} 1 \quad \text{for } \sigma = 1 \text{ (the die is 1)} \\ &\stackrel{\Delta}{=} 0 \quad \text{for } \sigma = 2, 3, 4, 5, 6 \text{ (the die is not 1)} \end{aligned}$$

where

$$\begin{aligned} \omega_{\mathbf{y}}(y) &= \frac{5}{6} \quad \text{for } y = 0 \\ &= \frac{1}{6} \quad \text{for } y = 1 \end{aligned}$$

Two random variables \mathbf{x}_1 and \mathbf{x}_2 are equal if and only if their values $\mathbf{x}_1(\sigma)$ and $\mathbf{x}_2(\sigma)$ are identical for all possible outcomes σ of the experiment.

A vector **space of random variables** defined on a given experiment consists in a set of functions defined on the possible outcomes of the experiment, together with the following definitions of addition and scalar multiplication*:

$$(a\mathbf{x})(\sigma) \stackrel{\Delta}{=} a(\mathbf{x}(\sigma)) \quad (\mathbf{x} + \mathbf{y})(\sigma) \stackrel{\Delta}{=} \mathbf{x}(\sigma) + \mathbf{y}(\sigma)$$

for all possible outcomes σ of the experiment. Let \mathfrak{V} be the space of all possible random variables defined on the above die-throwing experiment. If \mathbf{x} and \mathbf{y} are the particular vectors described above, then $\mathbf{x} + \mathbf{y}$ is given by

$$\begin{aligned} (\mathbf{x} + \mathbf{y})(\sigma) &\stackrel{\Delta}{=} 2 \quad \text{for } \sigma = 1 \\ &\stackrel{\Delta}{=} 1 \quad \text{for } \sigma = 3, 5 \\ &\stackrel{\Delta}{=} 0 \quad \text{for } \sigma = 2, 4, 6 \end{aligned}$$

and

$$\begin{aligned} \omega_{\mathbf{x} + \mathbf{y}}(z) &= \frac{1}{2} \quad \text{for } z = 0 \\ &= \frac{1}{3} \quad \text{for } z = 1 \\ &= \frac{1}{6} \quad \text{for } z = 2 \end{aligned}$$

What is the zero random variable for the vector space \mathfrak{V} ? It is $\mathbf{0}(\sigma) = 0$ for $\sigma = 1, \dots, 6$.

*We note that the set of functions must be such that it includes all sums and scalar multiples of its members.

2.2 Relations Among Vectors

Combining Vectors

Assuming a vector represents the condition or change in condition of a system, we can use the definitions of addition and scalar multiplication of vectors to find the net result of several successive changes in condition of the system.

Definition. A vector \mathbf{x} is said to be a **linear combination** of the vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ if it can be expressed as

$$\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \dots + c_n\mathbf{x}_n \tag{2.7}$$

for some set of scalars c_1, \dots, c_n . This concept is illustrated in Figure 2.5 where $\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3$.

A vector space \mathcal{V} is simply a set of elements and a definition of linear combination (addition and scalar multiplication); the space \mathcal{V} includes all linear combinations of its own elements. If S is a subset of \mathcal{V} , the set of all linear combinations of vectors from S , using the same definition of linear combination, is also a vector space. We call it a subspace of \mathcal{V} . A line or plane through the origin of the three-dimensional vector space is an example of a subspace.

Definition. A subset \mathcal{W} of a vector space \mathcal{V} is a **linear subspace** (or **linear manifold**) of \mathcal{V} if along with every pair, \mathbf{x}_1 and \mathbf{x}_2 , of vectors in \mathcal{W} , every linear combination $c_1\mathbf{x}_1 + c_2\mathbf{x}_2$ is also in \mathcal{W} .* We call \mathcal{W} a *proper subspace* if it is smaller than \mathcal{V} ; that is if \mathcal{W} is not \mathcal{V} itself.

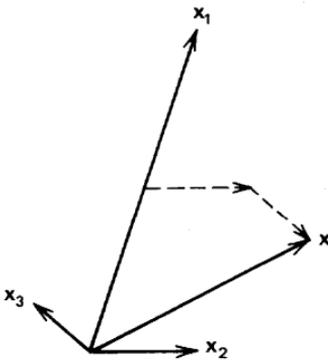


Figure 2.5. A linear combination of arrows.

*In the discussion of infinite-dimensional Hilbert spaces (Section 5.3), we distinguish between a linear subspace and a linear manifold. Linear manifold is the correct term to use in this definition. Yet because a finite-dimensional linear manifold is a linear subspace as well, we emphasize the physically motivated term subspace.

Example 1. A Linear Subspace. The set of vectors from \mathfrak{R}^3 which are of the form $(c_1, c_2, c_1 + c_2)$ forms a subspace of \mathfrak{R}^3 . It is, in fact, the set of all linear combinations of the two vectors $(1, 0, 1)$ and $(0, 1, 1)$.

Example 2. A Solution Space. The set \mathfrak{W} of all solutions to the matrix equation

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 2 & 3 & 3 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

is a subspace of $\mathfrak{N}^{3 \times 1}$. By elimination (Section 1.5), we find that \mathfrak{W} contains all vectors of the form $(0 \ \xi_2 \ -\xi_2)^T$. Clearly, \mathfrak{W} consists in all linear combinations of the single vector $(0 \ 1 \ -1)^T$. This example extends to general matrices. Let A be an $m \times n$ matrix. Let \mathbf{x} be in $\mathfrak{N}^{n \times 1}$. Using the rules of matrix multiplication (Appendix 1) it can be shown that if \mathbf{x}_1 and \mathbf{x}_2 are solutions to $A\mathbf{x} = \boldsymbol{\theta}$, then an arbitrary linear combination $c_1\mathbf{x}_1 + c_2\mathbf{x}_2$ is also a solution. Thus the space of solutions is a subspace of $\mathfrak{N}^{n \times 1}$.

Example 3. Subspaces (Linear Manifolds) of Functions. Let $\mathcal{C}^2(\Omega)$ be the space of all real-valued functions which are defined and have continuous second partial derivatives in the two-dimensional region Ω . (This region could be the square $0 < s < 1, 0 < t < 1$, for instance.) Let Γ denote the boundary of the region Ω . Linear combination in $\mathcal{C}^2(\Omega)$ is defined by

$$\begin{aligned} (\mathbf{f} + \mathbf{g})(s, t) &\triangleq \mathbf{f}(s, t) + \mathbf{g}(s, t) \\ (a\mathbf{f})(s, t) &\triangleq a(\mathbf{f}(s, t)) \end{aligned} \tag{2.8}$$

for all (s, t) in Ω . The functions \mathbf{f} in $\mathcal{C}^2(\Omega)$ which satisfy the homogeneous boundary condition $\mathbf{f}(s, t) = 0$ for (s, t) on Γ constitute a linear manifold of $\mathcal{C}^2(\Omega)$. For if \mathbf{f}_1 and \mathbf{f}_2 satisfy the boundary condition, then $(c_1\mathbf{f}_1 + c_2\mathbf{f}_2)(s, t) = c_1\mathbf{f}_1(s, t) + c_2\mathbf{f}_2(s, t) = 0$, and the arbitrary linear combination $c_1\mathbf{f}_1 + c_2\mathbf{f}_2$ also satisfies the boundary condition.

The set of solutions to Laplace's equation,

$$\frac{\partial^2 \mathbf{f}(s, t)}{\partial s^2} + \frac{\partial^2 \mathbf{f}(s, t)}{\partial t^2} = 0 \tag{2.9}$$

for all (s, t) in Ω , also forms a linear manifold of $\mathcal{C}^2(\Omega)$. For if \mathbf{f}_1 and \mathbf{f}_2 both satisfy (2.9), then

$$\frac{\partial^2 [c_1\mathbf{f}_1(s, t) + c_2\mathbf{f}_2(s, t)]}{\partial s^2} + \frac{\partial^2 [c_1\mathbf{f}_1(s, t) + c_2\mathbf{f}_2(s, t)]}{\partial t^2} = 0$$

and the arbitrary linear combination $c_1\mathbf{f}_1 + c_2\mathbf{f}_2$ also satisfies (2.9). Equation (2.9) is phrased in terms of the values of \mathbf{f} . Laplace's equation can also be expressed in the

vector notation

$$\nabla^2 \mathbf{f} = \boldsymbol{\theta} \quad (2.10)$$

The domain of definition Ω is implicit in (2.10). The vector $\boldsymbol{\theta}$ is defined by $\boldsymbol{\theta}(s, t) = 0$ for all (s, t) in Ω .

In using vector diagrams to analyze physical problems, we often resolve a vector into a linear combination of component vectors. We usually do this in a unique manner. In Figure 2.5, \mathbf{x} is not a unique linear combination of \mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3 ; $\mathbf{x} = 0\mathbf{x}_1 + 3\mathbf{x}_2 + 2\mathbf{x}_3$ is a second resolution of \mathbf{x} ; the number of possible resolutions is infinite. In point of fact, \mathbf{x} can be represented as a linear combination of any two of the other vectors; the three vectors \mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3 are redundant as far as representation of \mathbf{x} is concerned.

Definition. The vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are **linearly dependent** (or coplanar) if at least one of them can be written as a linear combination of the others. Otherwise they are **linearly independent**. (We often refer to sets of vectors as simply “dependent” or “independent.”)

In Figure 2.5 the set $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is dependent. Any two of the vectors form an independent set. In any vector space, a set which contains the $\boldsymbol{\theta}$ vector is dependent, for $\boldsymbol{\theta}$ can be written as zero times any other vector in the set. We define the $\boldsymbol{\theta}$ vector by itself as a dependent set.

The following statement is equivalent to the above definition of independence: the vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are linearly independent if and only if

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n = \boldsymbol{\theta} \Rightarrow c_1 = \cdots = c_n = 0 \quad (2.11)$$

Equation (2.11) says the “zero combination” is the only combination that equals $\boldsymbol{\theta}$. For if c_i were not 0, we could simply divide by c_i to find \mathbf{x}_i as a linear combination of the other vectors, and the set $\{\mathbf{x}_j\}$ would be dependent. If $c_i = 0$, \mathbf{x}_i cannot be a linear combination of the other vectors. Equation (2.11) is a practical tool for determining independence of vectors.

Exercise 1. Explore graphically and by means of (2.11) the following set of vectors from \mathcal{R}^3 : $\{\mathbf{x}_1 = (1, 0, 0), \mathbf{x}_2 = (0, 1, 0), \mathbf{x}_3 = (1, 1, 0), \mathbf{x}_4 = (0, 0, 1)\}$.

Example 4. Determining Independence In the space \mathcal{R}^3 let $\mathbf{x}_1 = (1, 2, 1)$, $\mathbf{x}_2 = (2, 3, 1)$, and $\mathbf{x}_3 = (4, 7, 3)$. Equation (2.11) becomes

$$\begin{aligned} c_1(1, 2, 1) + c_2(2, 3, 1) + c_3(4, 7, 3) \\ &= (c_1 + 2c_2 + 4c_3, \quad 2c_1 + 3c_2 + 7c_3, \quad c_1 + c_2 + 3c_3) \\ &= (0, 0, 0) \end{aligned}$$

Each component of this vector equation is a scalar-valued linear algebraic equation. We write the three equations in the matrix form:

$$\begin{pmatrix} 1 & 2 & 4 \\ 2 & 3 & 7 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

We solve this equation by elimination (Section 1.5) to find $c_1 = -2c_3$ and $c_2 = -c_3$. Any choice for c_3 will yield a particular nonzero linear combination of the vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ which equals $\mathbf{0}$. The set is linearly dependent.

Definition. Let $\mathfrak{S} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be a set of vectors from a linear space \mathfrak{V} . The set of all linear combinations of vectors from \mathfrak{S} is called the subspace of \mathfrak{V} **spanned** (or **generated**) by \mathfrak{S} .^{*} We often refer to this subspace as $\text{span}(\mathfrak{S})$ or $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.

Bases and Coordinates

We have introduced the vector space concept in order to provide a common mathematical framework for different types of systems. We can make the similarities between systems more apparent by converting their vector space representations to a standard form. We perform this standardization by introducing coordinate systems. In the example of Figure 2.5, the vectors $\{\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ span a plane; yet any two of them will span the same plane. Two of them are redundant as far as generation of the plane is concerned.

Definition. A **basis** (or coordinate system) for a linear space \mathfrak{V} is a linearly independent set of vectors from \mathfrak{V} which spans \mathfrak{V} .

Example 5. *The Standard Bases for \mathfrak{R}^n , $\mathfrak{N}^n \times 1$, and \mathfrak{P}^n .* It is evident that any three linearly independent vectors in \mathfrak{R}^3 form a basis for \mathfrak{R}^3 . The n -tuples

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \dots, 0) \\ \mathbf{e}_2 &= (0, 1, 0, \dots, 0) \\ &\vdots \\ \mathbf{e}_n &= (0, \dots, 0, 1) \end{aligned} \tag{2.12}$$

form a basis for \mathfrak{R}^n . The set $\mathfrak{E} \triangleq \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is called the **standard basis for \mathfrak{R}^n** .

We use the same notation to represent the standard basis for $\mathfrak{N}^n \times 1$: $\mathfrak{E} \triangleq \{\mathbf{e}_i\}$, where \mathbf{e}_i is a column vector of zeros except for a 1 in the i th place. The set $\mathfrak{N} \triangleq \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ defined by $\mathbf{f}_k(t) = t^{k-1}$ forms a basis for \mathfrak{P}^n ; it is analogous to the standard bases for \mathfrak{R}^n and $\mathfrak{N}^n \times 1$.

^{*}The definition of the space spanned by an infinite set of vectors depends on limiting concepts. We delay the definition until Section 5.3.

Example 6. The Zero Vector Space. The set $\{\mathbf{0}\}$ together with the obvious definitions of addition and scalar multiplication forms a vector space which we denote \mathcal{O} . However, the vector $\mathbf{0}$, by itself, is a dependent set. Therefore \mathcal{O} has no basis.

If $\mathcal{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is a basis for the space \mathcal{V} , any vector \mathbf{x} in \mathcal{V} can be written uniquely as some linear combination

$$\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n \quad (2.13)$$

of vectors in \mathcal{X} . The multipliers c_i are called the **coordinates of \mathbf{x} relative to the ordered basis \mathcal{X}** . It is easy to show that the coordinates relative to a particular ordered basis are unique: just expand \mathbf{x} as in (2.13) for a second set $\{d_i\}$ of coordinates; then independence of the basis vectors implies $d_i = c_i$.

It is common to write the coordinates of a vector relative to a particular basis as a column matrix. We will denote by $[\mathbf{x}]_{\mathcal{X}}$ the **coordinate matrix** of the vector \mathbf{x} relative to the (ordered) basis \mathcal{X} ; thus corresponding to (2.13) we have

$$[\mathbf{x}]_{\mathcal{X}} \triangleq \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} \quad (2.14)$$

Some bases are more natural or convenient than others. We use the term **natural basis** to mean a basis relative to which we can find coordinates by inspection. The bases of Example 5 are natural bases for \mathcal{R}^n , $\mathcal{N}^{n \times 1}$, and \mathcal{P}^n . Thus if $\mathbf{f}(t) = \xi_1 + \xi_2 t + \cdots + \xi_n t^{n-1}$, then $[\mathbf{f}]_{\mathcal{X}} = (\xi_1 \ \xi_2 \ \cdots \ \xi_n)^T$.

Example 7. Coordinates for Vectors in \mathcal{R}^3 . Let $\mathcal{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ be an ordered basis for \mathcal{R}^3 , where $\mathbf{x}_1 = (1, 2, 3)$, $\mathbf{x}_2 = (2, 3, 2)$, and $\mathbf{x}_3 = (2, 5, 5)$. Let $\mathbf{x} = (1, 1, 1)$. To find $[\mathbf{x}]_{\mathcal{X}}$, we must solve (2.13):

$$\begin{aligned} (1, 1, 1) &= c_1(1, 2, 3) + c_2(2, 3, 2) + c_3(2, 5, 5). \\ &= (c_1 + 2c_2 + 2c_3, \quad 2c_1 + 3c_2 + 5c_3, \quad 3c_1 + 2c_2 + 5c_3) \end{aligned}$$

We rewrite the vector (3-tuple) equation in the matrix notation:

$$\begin{pmatrix} 1 & 2 & 2 \\ 2 & 3 & 5 \\ 3 & 2 & 5 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad (2.15)$$

We solved this equation in Example 1 of Section 1.5. The result is

$$[\mathbf{x}]_{\mathcal{X}} \stackrel{\Delta}{=} [(1, 1, 1)]_{\mathcal{X}} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} \\ \frac{3}{5} \\ -\frac{2}{5} \end{pmatrix}$$

The coordinate matrix of Example 7 is merely a simple way of stating that $\mathbf{x} = \frac{2}{5}\mathbf{x}_1 + \frac{2}{5}\mathbf{x}_2 - \frac{2}{5}\mathbf{x}_3$. We choose to write the coordinates of a vector \mathbf{x} as a column matrix because it allows us to carry out in a standard matrix format all manipulations involving the coordinates of \mathbf{x} .

In Example 4 of Section 1.5 we solved (2.15) with a general right-hand side; that is, for $\mathbf{x} = (\eta_1, \eta_2, \eta_3)$. That solution allows us to determine quickly the coordinate matrix, relative to the basis \mathcal{X} of Example 7, for *any* vector \mathbf{x} in \mathcal{R}^3 , including the case $\mathbf{x} = (0, 0, 0)$. In general, (2.13) includes (2.11); inherent in the process of finding coordinates for an arbitrary vector \mathbf{x} is the process of determining whether \mathcal{X} is a basis. If \mathcal{X} is not independent, there will exist nonzero coordinates for $\mathbf{x} = \mathbf{0}$. If \mathcal{X} does not span the space, there will be some vector \mathbf{x} for which no coordinates exist (P&C 2.7).

Example 8. Coordinates for Vectors in \mathcal{P}^3 . Let $\mathcal{F} \stackrel{\Delta}{=} \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ be an ordered basis for \mathcal{P}^3 , where $\mathbf{f}_1(t) = 1 + 2t + 3t^2$, $\mathbf{f}_2(t) = 2 + 3t + 2t^2$, and $\mathbf{f}_3(t) = 2 + 5t + 5t^2$. Let \mathbf{f} be defined by $\mathbf{f}(t) = 1 + t + t^2$. To find $[\mathbf{f}]_{\mathcal{F}}$, we solve (2.13), $\mathbf{f} = c_1\mathbf{f}_1 + c_2\mathbf{f}_2 + c_3\mathbf{f}_3$. To solve this equation, we evaluate both sides at t :

$$\begin{aligned} \mathbf{f}(t) &= (c_1\mathbf{f}_1 + c_2\mathbf{f}_2 + c_3\mathbf{f}_3)(t) \\ &= c_1\mathbf{f}_1(t) + c_2\mathbf{f}_2(t) + c_3\mathbf{f}_3(t) \end{aligned} \quad (2.16)$$

or

$$\begin{aligned} 1 + t + t^2 &= c_1(1 + 2t + 3t^2) + c_2(2 + 3t + 2t^2) + c_3(2 + 5t + 5t^2) \\ &= (c_1 + 2c_2 + 2c_3) + (2c_1 + 3c_2 + 5c_3)t + (3c_1 + 2c_2 + 5c_3)t^2 \end{aligned}$$

Equating coefficients on like powers of t we again obtain (2.15). The coordinate matrix of \mathbf{f} is

$$[\mathbf{f}]_{\mathcal{F}} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} \\ \frac{3}{5} \\ -\frac{2}{5} \end{pmatrix}$$

In order to solve the vector (function) equation (2.16) we converted it to a set of scalar equations expressed in matrix form. A second method for

converting (2.16) to a matrix equation in the unknowns $\{c_i\}$ is to evaluate the equation at three different values of t . Each such evaluation yields an algebraic equation in $\{c_i\}$. The resulting matrix equation is different from (2.15), but the solution is the same. We now describe a general method, built around a natural basis, for converting (2.13) to a matrix equation. The coordinate matrix of a vector \mathbf{x} relative to the basis $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is $[\mathbf{x}]_{\mathcal{X}} = (c_1 \cdots c_n)^T$, where the coordinates c_i are obtained by solving the vector equation

$$\mathbf{x} = c_1 \mathbf{x}_1 + \cdots + c_n \mathbf{x}_n$$

A general method for obtaining an equivalent matrix equation consists in taking coordinates of the vector equation relative to a natural basis \mathcal{N} —a basis relative to which coordinates can be obtained by inspection. The vector equation becomes

$$\begin{aligned} [\mathbf{x}]_{\mathcal{N}} &= \left[\sum_{i=1}^n c_i \mathbf{x}_i \right]_{\mathcal{N}} \\ &= \sum_{i=1}^n c_i [\mathbf{x}_i]_{\mathcal{N}} \\ &= ([\mathbf{x}_1]_{\mathcal{N}} \vdots \cdots \vdots [\mathbf{x}_n]_{\mathcal{N}}) [\mathbf{x}]_{\mathcal{X}} \end{aligned} \quad (2.17)$$

We determine $[\mathbf{x}]_{\mathcal{N}}, [\mathbf{x}_1]_{\mathcal{N}}, \dots, [\mathbf{x}_n]_{\mathcal{N}}$ by inspection. Then we solve (2.17) routinely for $[\mathbf{x}]_{\mathcal{X}}$.

Example 9. Finding Coordinates via a Natural Basis. Let the set $\mathcal{F} \stackrel{\Delta}{=} \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ be a basis for \mathcal{P}^3 , where $\mathbf{f}_1(t) = 1 + 2t + 3t^2$, $\mathbf{f}_2(t) = 2 + 3t + 2t^2$, and $\mathbf{f}_3(t) = 2 + 5t + 5t^2$. We seek $[\mathbf{f}]_{\mathcal{F}}$ for the vector $\mathbf{f}(t) = 1 + t + t^2$. To convert the defining equation for coordinates into a matrix equation, we use the natural basis $\mathcal{N} \stackrel{\Delta}{=} \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, where $\mathbf{g}_k(t) = t^{k-1}$. For this problem, (2.17) becomes

$$[\mathbf{f}]_{\mathcal{N}} = ([\mathbf{f}_1]_{\mathcal{N}} \vdots [\mathbf{f}_2]_{\mathcal{N}} \vdots [\mathbf{f}_3]_{\mathcal{N}}) [\mathbf{f}]_{\mathcal{F}}$$

or

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 3 & 5 \\ 3 & 2 & 5 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

The solution to this equation is $[\mathbf{f}]_{\mathcal{F}} = (\frac{2}{5} \frac{3}{5} - \frac{2}{5})^T$. (Compare with Example 8.)

Typically, the solution of (2.17) requires the elimination procedure

$$([\mathbf{x}_1]_{\mathcal{U}} \vdots \cdots \vdots [\mathbf{x}_n]_{\mathcal{U}} \vdots [\mathbf{x}]_{\mathcal{U}}) \rightarrow (\mathbf{I} \vdots [\mathbf{x}]_{\mathcal{U}}) \quad (2.18)$$

If we wish to solve for the coordinates of more than one vector, we still perform the elimination indicated in (2.18), but augment the matrix with all the vectors whose coordinates we desire. Thus if we wish the coordinates for \mathbf{z}_1 , \mathbf{z}_2 , and \mathbf{z}_3 , we perform elimination on

$$([\mathbf{x}_1]_{\mathcal{U}} \vdots \cdots \vdots [\mathbf{x}_n]_{\mathcal{U}} \vdots [\mathbf{z}_1]_{\mathcal{U}} \vdots [\mathbf{z}_2]_{\mathcal{U}} \vdots [\mathbf{z}_3]_{\mathcal{U}})$$

This elimination requires less computation than does the process which goes through inversion of the matrix $([\mathbf{x}_1]_{\mathcal{U}} \vdots \cdots \vdots [\mathbf{x}_n]_{\mathcal{U}})$, regardless of the number of vectors whose coordinates we desire (P&C 1.3).

Example 10. A Basis and Coordinates for a Subspace. Let \mathcal{W} be the **subspace** of \mathcal{P}^3 consisting in all functions \mathbf{f} defined by the rule $\mathbf{f}(t) = \xi_1 + \xi_2 t + (\xi_1 + \xi_2)t^2$ for some ξ_1 and ξ_2 . Note that the standard basis functions for \mathcal{P}^3 are not contained in \mathcal{W} . The functions defined by $\mathbf{g}_1(t) = 1 + t^2$ and $\mathbf{g}_2(t) = t + t^2$ are clearly independent vectors in \mathcal{W} . Because there are two “degrees of freedom” in \mathcal{W} (i.e., two parameters ξ_1 and ξ_2 must be given to specify a particular function in \mathcal{W}) we expect the set $\mathcal{G} \stackrel{\Delta}{=} \{\mathbf{g}_1, \mathbf{g}_2\}$ to span \mathcal{W} and thus be a basis. We seek the coordinate matrix $[\mathbf{f}]_{\mathcal{G}}$ of an arbitrary vector \mathbf{f} in \mathcal{W} . That is, we seek \mathbf{c}_1 and \mathbf{c}_2 such that

$$\mathbf{f}(t) = c_1 \mathbf{g}_1(t) + c_2 \mathbf{g}_2(t)$$

The matrix equation (2.17) can be written by inspection using the natural basis \mathcal{U} of Example 9:

$$[\mathbf{f}]_{\mathcal{U}} = ([\mathbf{g}_1]_{\mathcal{U}} \vdots [\mathbf{g}_2]_{\mathcal{U}}) [\mathbf{f}]_{\mathcal{G}}$$

or

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_1 + \xi_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Then $c_i = \xi_i$ and

$$[\mathbf{f}]_{\mathcal{G}} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

Because we were able to solve uniquely for the coordinates, we know that \mathcal{G} is indeed a basis for \mathcal{W} . The subspace \mathcal{W} is equivalent to the subspace of Example 1. Note that the elimination procedure does not agree precisely

with (2.18) because there are only two degrees of freedom among the three coefficients of the arbitrary vector \mathbf{f} in \mathcal{W} .

Dimension

The equivalence between the three vector spaces \mathcal{R}^3 , \mathcal{P}^3 , and $\mathcal{N}^{3 \times 1}$ is apparent from Examples 7 and 8; The subspace \mathcal{W} of Example 10, however, is equivalent to $\mathcal{N}^{2 \times 1}$ rather than $\mathcal{N}^{3 \times 1}$, even though the elements of \mathcal{W} are polynomials in \mathcal{P}^3 . The key to the equivalence lies not in the nature of the elements, but rather in the number of “degrees of freedom” in each space (the number of scalars which must be specified in order to specify a vector); more to the point, the key lies in the number of vectors in a basis for each space.

Definition. A vector space is **finite dimensional** if it is spanned by a finite number of vectors. It is intuitively clear that all bases for a finite-dimensional space contain the same number of vectors. The number of vectors in a basis for a finite-dimensional space \mathcal{V} is called the **dimension** of \mathcal{V} and is denoted by $\text{dim}(\mathcal{V})$.

Thus \mathcal{R}^3 and \mathcal{P}^3 are both three-dimensional spaces. The subspace \mathcal{W} of Example 10 has dimension 2. Knowledge of the dimension of a space (or a subspace) is obtained in the course of determining a basis for the space (subspace). Since the space $\mathcal{O} \triangleq \{\mathbf{0}\}$ has no basis, we assign it dimension zero.

Example 11. A Basis for a Space of Random Variables. A vector space \mathcal{V} of random variables, defined on the possible outcomes of a single die-throwing experiment, is described in Example 11 of Section 2.1. A natural basis for \mathcal{V} is the set of random variables $\mathcal{X} \triangleq \{\mathbf{x}_i, i=1, \dots, 6\}$, where

$$\begin{aligned} \mathbf{x}_i(\sigma) &\triangleq 1 \text{ for } \sigma = i \text{ (the die equals } i) \\ &\triangleq 0 \text{ for } \sigma \neq i \text{ (the die does not equal } i) \end{aligned}$$

That \mathcal{X} is a basis for \mathcal{V} can be seen from an attempt to determine the coordinates with respect to \mathcal{X} of an arbitrary random variable \mathbf{z} defined on the experiment. If

$$\begin{aligned} \mathbf{z}(\sigma) &\triangleq \begin{matrix} c_1 & \text{for } \sigma = 1 \\ \vdots \\ c_6 & \text{for } \sigma = 6 \end{matrix} \end{aligned}$$

then $[\mathbf{z}]_{\mathcal{X}} = (c_1 \cdots c_6)^T$; a unique representation exists.

The random variables $\{\mathbf{x}_1, \dots, \mathbf{x}_6\}$ are linearly independent. However, they are not *statistically* independent. **Statistical independence** of two random variables \mathbf{x} and \mathbf{y} means that knowledge of the *value* of one variable, say, \mathbf{x} , does not tell us anything about the outcome of the experiment which determines the value of the other variable \mathbf{y} , and therefore it tells us nothing about the value of \mathbf{y} . The random variables $\{\mathbf{x}_i\}$ are related by the underlying die-throwing experiment. If we know $\mathbf{x}_1 = 0$, for instance, then we know $\sigma \neq 1$ (the die is not equal to 1); the probability mass functions for $\mathbf{x}_2, \dots, \mathbf{x}_6$ and for all other vectors in \mathcal{V} are modified by the information concerning the value of \mathbf{x}_1 . The new probability mass functions for \mathbf{x} and \mathbf{y} of Example 11, Section 2.1, given that $\mathbf{x}_1 = 0$, are

$$\begin{aligned} \omega_{\mathbf{x}}(x; \mathbf{x}_1 = 0) &= \frac{3}{5} \quad \text{for } x = 0 & \omega_{\mathbf{y}}(y; \mathbf{x}_1 = 0) &= 1 \quad \text{for } y = 0 \\ &= \frac{2}{5} \quad \text{for } x = 1 & &= 0 \quad \text{for } y = 1 \end{aligned}$$

The space l_2 of square-summable sequences described in Example 4 of Section 2.1 is obviously *infinite dimensional*. A direct extension of the standard basis for \mathcal{R}^n seems likely to be a basis for l_2 . It is common knowledge that functions \mathbf{f} in $\mathcal{C}(0, 2\pi)$, the space of functions continuous on $[0, 2\pi]$, can be expanded uniquely in a Fourier series of the form $\mathbf{f}(t) = b_0 + \sum_{k=1}^{\infty} (a_k \sin kt + b_k \cos kt)$. This fact leads us to suspect that the set of functions

$$\mathcal{F} \triangleq \{ \mathbf{1}, \sin t, \cos t, \sin 2t, \cos 2t, \dots \} \quad (2.19)$$

forms a basis for $\mathcal{C}(0, 2\pi)$, and that the coordinates of \mathbf{f} relative to this basis are

$$(b_0, a_1, b_1, a_2, b_2, \dots)$$

This suspicion is correct. The coordinates (or Fourier coefficients) actually constitute a vector in l_2 . We show in Example 11 of Section 5.3 that l_2 serves as a convenient standard space of coordinate vectors for infinite-dimensional spaces; in that sense, it plays the same role that $\mathcal{N}^n \times 1$ does for n -dimensional spaces. Unfortunately, the concepts of independence, spanning sets, and bases do not extend easily to infinite-dimensional vector spaces. The concept of linear combination applies only to the combination of a finite number of vectors. We cannot add an infinite number of vectors without the concept of a limit; this concept is introduced in Chapter 5. Hence detailed examination of infinite-dimensional function spaces is left for that chapter.

Summary

There is no inherent basis in any space—one basis is as good as another. Yet a space may have one basis which appears more convenient than others. The standard basis for \mathcal{R}^n is an example. By picking units of measurement in a physical system (e.g., volts, feet, degrees centigrade) we tie together the system and the model; our choice of units may automatically determine convenient or standard basis vectors for the vector space of the model (based on, say, 1 V, 1 ft, or 1 °C).

By choosing a basis for a space, we remove the most distinguishing feature of that space, the nature of its elements, and thus tie each vector in the space to a unique coordinate matrix. Because of this unique connection which a basis establishes between the elements of a particular vector space and the elements of the corresponding space of coordinate matrices, we are able to carry out most vector manipulations in terms of coordinate matrices which represent the vectors. We have selected $\mathcal{N}^{n \times 1}$, rather than \mathcal{R}^n , as our standard n -dimensional space because matrix operations are closely tied to computer algorithms for solving linear algebraic equations (Section 1.5). Most vector space manipulations lead eventually to such equations.

Because coordinate matrices are themselves vectors in a vector space ($\mathcal{N}^{n \times 1}$), we must be careful to distinguish vectors from their coordinates. The confusion is typified by the problem of finding the coordinate matrix of a vector \mathbf{x} from $\mathcal{N}^{n \times 1}$ relative to the standard basis for $\mathcal{N}^{n \times 1}$. In this instance $[\mathbf{x}]_{\mathcal{E}} = \mathbf{x}$; the difference between the vector and its coordinate matrix is only conceptual. A vector is simply one of a set of elements, although we may use it to represent the physical condition of some system. The coordinate matrix of the vector, on the other hand, is the unique set of multipliers which specifies the vector as a linear combination of arbitrarily chosen basis vectors.

2.3 System Models

The concept of a vector as a model for the condition or change in condition of a system is explored in Sections 2.1 and 2.2. We usually separate the variables which pertain to the condition of the system into two broad sets: the independent (or input) variables, the values of which are determined outside of the system, and the dependent (or output) variables, whose values are determined by the system together with the independent variables. A model for the system itself consists in expressions of relations among the variables. In this section we identify properties of system models.

Example 1. An Economic System Let \mathbf{x} represent a set of inputs to the U. S. national economy (tax rates, interest rates, reinvestment policies, etc.); let \mathbf{y} represent a set of economic indicators (cost of living, unemployment rate, growth rate, etc.). The system model \mathbf{T} must describe the economic laws which relate \mathbf{y} to \mathbf{x} .

Example 2. A Baking Process. Suppose \mathbf{x} is the weight of a sample of clay before a baking process and \mathbf{y} is the weight after baking. Then the system model \mathbf{T} must describe the chemical and thermodynamic laws insofar as they relate \mathbf{x} and \mathbf{y} .

Example 3. A Positioning System. Suppose the system of interest is an armature-controlled motor which is used to position a piece of equipment. Let \mathbf{x} represent the armature voltage, a function of time; let \mathbf{y} be the shaft position, another function of time. The system model \mathbf{T} should describe the manner in which the dynamic system relates the function \mathbf{y} to the function \mathbf{x} .

The variables in the economic system of Example 1 clearly separate into input (or independent) variables and output (or system condition) variables. In Example 2, both the independent and dependent variables describe the condition of the system. Yet we can view the condition before baking as the input to the system and view the condition after baking as the output. The dynamic system of Example 3 is reciprocal; \mathbf{x} and \mathbf{y} are mutually related by \mathbf{T} . Since the system is used as a motor, we view the armature voltage \mathbf{x} as the input to the system and the shaft position \mathbf{y} as the output. We could, as well, use the machine as a dc generator; then we would view the shaft position as the input and the armature voltage as the output.

The notation $\mathbf{T}\mathbf{x} = \mathbf{y}$ that we introduced in (1.1) implies that the model \mathbf{T} does something to the vector \mathbf{x} to yield the vector \mathbf{y} . As a result, we may feel inclined to call \mathbf{x} the input and \mathbf{y} the output. Yet in Section 1.3 we note that equations are sometimes expressed in an inverse form. The positions of the variables in an equation do not determine whether they are independent or dependent variables. Furthermore, we can see from Example 3 that the input and output of a system in some instances may be determined arbitrarily. In general, we treat one of the vectors in the equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ as the input and the other as the output. However, unless we are exploring a problem for which the input is clearly defined, we use the terms input and output loosely in reference to the known and unknown variables, respectively.

Transformations on Vector Spaces

Our present purpose is to make more precise the vaguely defined model \mathbf{T} introduced in (1.1) and illustrated above.

Definition. A **transformation** or **function** $\mathbf{T}: \mathfrak{S}_1 \rightarrow \mathfrak{S}_2$ is a rule that

associates with each element of the set \mathcal{S}_1 a unique element from the set \mathcal{S}_2^* . The set \mathcal{S}_1 is called the **domain** of \mathbf{T} ; \mathcal{S}_2 is the **range of definition** of \mathbf{T} .

Our attention is directed primarily toward transformations where \mathcal{S}_1 and \mathcal{S}_2 are linear spaces. We speak of $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ as a transformation from the vector space \mathcal{V} into the vector space \mathcal{W} . An **operator** is another term for a transformation between vector spaces. We use this term primarily when the domain and range of definition are identical; we speak of $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{V}$ as an operator on \mathcal{V} . If $\mathcal{S}_{\mathcal{V}}$ is a subset of \mathcal{V} , we denote by $\mathbf{T}(\mathcal{S}_{\mathcal{V}})$ the set of all vectors $\mathbf{T}\mathbf{x}$ in \mathcal{W} for which \mathbf{x} is in $\mathcal{S}_{\mathcal{V}}$; we refer to $\mathbf{T}(\mathcal{S}_{\mathcal{V}})$ as the **image of $\mathcal{S}_{\mathcal{V}}$ under \mathbf{T}** . The **range of \mathbf{T}** is $\mathbf{T}(\mathcal{V})$, the image of \mathcal{V} under \mathbf{T} . The **nullspace of \mathbf{T}** is the set of all vectors \mathbf{x} in \mathcal{V} such that $\mathbf{T}\mathbf{x} = \theta_{\mathcal{W}}$ ($\theta_{\mathcal{W}}$ is the zero vector in the space \mathcal{W}). If $\mathcal{S}_{\mathcal{W}}$ is a subset of \mathcal{W} , we call the set of vectors \mathbf{x} in \mathcal{V} for which $\mathbf{T}\mathbf{x}$ is in $\mathcal{S}_{\mathcal{W}}$ the **inverse image of $\mathcal{S}_{\mathcal{W}}$** . Thus the nullspace of \mathbf{T} is the inverse image of the set $\{\theta_{\mathcal{W}}\}$. See Figure 2.6.

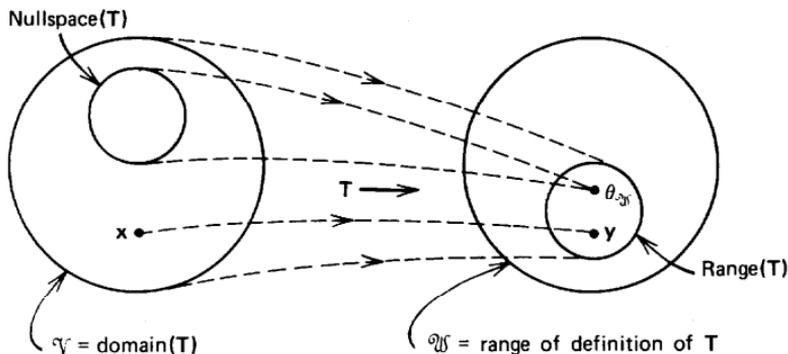


Figure 2.6. Abstract illustration of a transformation \mathbf{T} .

Example 4. A Transformation Define $\mathbf{T}: \mathcal{R}^2 \rightarrow \mathcal{R}^1$ by

$$\mathbf{T}(\xi_1, \xi_2) \triangleq \begin{cases} \sqrt{\xi_1^2 + \xi_2^2} - 1 & \text{for } \xi_1^2 + \xi_2^2 \geq 1 \\ 0 & \text{for } \xi_1^2 + \xi_2^2 < 1 \end{cases} \quad (2.20)$$

Physically, the vector $\mathbf{T}\mathbf{x}$ can be interpreted as the distance between \mathbf{x} and the unit circle in the two-dimensional arrow space. The variables ξ_1 and ξ_2 are “dummy” variables; they merely assist us in cataloguing the “values” of \mathbf{T} in the defining

*In the modeling process we use the function concept twice: once as a vector—a model for the condition of a system—and once as a relation between input and output vectors—a model for the system itself. In order to avoid confusion, we use the term function in referring to vectors in a vector space, but the term transformation in referring to the relation between vectors.

equation; we can use any other symbols in their place without changing the definition of \mathbf{T} . The range of \mathbf{T} is the set of positive numbers in \mathcal{R}^1 . The nullspace of \mathbf{T} is the set consisting of all vectors in the domain \mathcal{R}^2 which satisfy $\xi_1^2 + \xi_2^2 \leq 1$.

Suppose we wish to solve the equation $\mathbf{T}\mathbf{x} = 1$ for the transformation of Example 4. In effect, we ask which points in the arrow space are a unit distance from the unit circle—all points on the circle of radius 2. The solution is not unique because \mathbf{T} assigns to the single number 1 in \mathcal{R}^1 more than one vector in \mathcal{R}^2 . The equation $\mathbf{T}\mathbf{x} = -1$, on the other hand, has no solution because \mathbf{T} does not assign the number -1 in \mathcal{R}^1 to any vector in \mathcal{R}^2 . We now proceed to specify the properties of a transformation which are necessary in order that the transformation be uniquely reversible.

Definition. Let $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$. Then \mathbf{T} is **one-to-one** if

$$\mathbf{x}_1 \neq \mathbf{x}_2 \Rightarrow \mathbf{T}\mathbf{x}_1 \neq \mathbf{T}\mathbf{x}_2 \quad (2.21)$$

for all \mathbf{x}_1 and \mathbf{x}_2 in \mathcal{V} ; that is, if \mathbf{T} does not assign more than one \mathbf{x} in \mathcal{V} to a single \mathbf{y} in \mathcal{W} .

If \mathbf{T} is one-to-one, any solution to $\mathbf{T}\mathbf{x} = \mathbf{y}$ is unique. It might appear that the effect of \mathbf{T} is reversible if \mathbf{T} is one-to-one. The nonreversibility of \mathbf{T} in Example 4, however, arises only in part because \mathbf{T} is not one-to-one. In general, there may be vectors in the range of definition \mathcal{W} which are not associated in any way with vectors in \mathcal{V} . In point of fact, $\text{range}(\mathbf{T})$ consists precisely of those vectors \mathbf{y} in \mathcal{W} for which the equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ is solvable. Unless we know which vectors are in $\text{range}(\mathbf{T})$, we cannot reverse the transformation.

Definition. Let $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$. Then \mathbf{T} is **onto** if

$$\text{range}(\mathbf{T}) = \mathcal{W} \quad (2.22)$$

That is, \mathbf{T} is onto if every vector \mathbf{y} in \mathcal{W} is associated with at least one vector \mathbf{x} in \mathcal{V} .

Definition. If a transformation is one-to-one and onto, then it is **invertible**—it can be reversed uniquely. If $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ is invertible, we define the **inverse of \mathbf{T}** to be the transformation $\mathbf{T}^{-1}: \mathcal{W} \rightarrow \mathcal{V}$ which associates with each \mathbf{y} in \mathcal{W} the unique vector \mathbf{x} in \mathcal{V} for which $\mathbf{T}\mathbf{x} = \mathbf{y}$. See (2.29) for another characterization of \mathbf{T}^{-1} .

Example 5. The Identity Operator, \mathbf{I} . Let \mathcal{V} be a vector space. Define the operator \mathbf{I} on \mathcal{V} by

$$\mathbf{I}\mathbf{x} \stackrel{\Delta}{=} \mathbf{x} \quad (2.23)$$

for all \mathbf{x} in \mathcal{V} . The nullspace of \mathbf{I} is $\theta_{\mathcal{V}}$. Range (\mathbf{I}) = \mathcal{V} ; thus \mathbf{I} is onto. Furthermore, \mathbf{I} is one-to-one. Therefore, the identity operator is invertible.

Example 6. The Zero Transformation, Θ . Let \mathcal{V} and \mathcal{W} be vector spaces. Define $\Theta: \mathcal{V} \rightarrow \mathcal{W}$ by

$$\Theta \mathbf{x} \triangleq \theta_{\mathcal{W}} \tag{2.24}$$

for all \mathbf{x} in \mathcal{V} . The nullspace of Θ is \mathcal{V} . The range of Θ is $\theta_{\mathcal{W}}$. The zero transformation is neither one-to-one nor onto. It is clearly not invertible.

Example 7. A Transformation on a Function Space. Define $\mathbf{T}: \mathcal{C}(a, b) \rightarrow \mathcal{R}^1$ by

$$\mathbf{T}\mathbf{f} \triangleq \int_a^b \mathbf{f}^2(t) dt \tag{2.25}$$

for all \mathbf{f} in $\mathcal{C}(a, b)$. This transformation specifies an integral-square measure of the size of the function \mathbf{f} ; this measure is used often in judging the performance of a control system. The function \mathbf{f} is a dummy variable used to define \mathbf{T} ; the scalar t is a dummy variable used to define \mathbf{f} . In order to avoid confusion, we must carefully distinguish between the concept of the function \mathbf{f} in the vector space $\mathcal{C}(a, b)$ and the concept of the transformation \mathbf{T} which relates each function \mathbf{f} in $\mathcal{C}(a, b)$ to a vector in \mathcal{R}^1 . The transformation acts on the whole function \mathbf{f} —we must use all values of \mathbf{f} to find $\mathbf{T}\mathbf{f}$. The range of \mathbf{T} is the set of positive numbers in \mathcal{R}^1 ; thus \mathbf{T} is not onto the range of definition \mathcal{R}^1 . The nullspace of \mathbf{T} is the single vector $\theta_{\mathcal{V}}$. If we define \mathbf{f}_1 and \mathbf{f}_2 by $\mathbf{f}_1(t) = 1$ and $\mathbf{f}_2(t) = -1$, then $\mathbf{T}\mathbf{f}_1 = \mathbf{T}\mathbf{f}_2$; therefore \mathbf{T} is not one-to-one.

The transformations of Examples 4 and 7 are scalar valued; that is, the range of definition in each case is the space of scalars. We call a scalar-valued transformation a **functional**. Most functionals are not one-to-one.

Example 8. A Transformation for a Dynamic System. Let $\mathcal{C}^2(a, b)$ be the space of functions which have continuous second derivatives on $[a, b]$. Define $\mathbf{L}: \mathcal{C}^2(a, b) \rightarrow \mathcal{C}(a, b)$ by

$$(\mathbf{L}\mathbf{f})(t) \triangleq \mathbf{f}''(t) + \alpha(\mathbf{f}(t) + 0.01\mathbf{f}^3(t)) \tag{2.26}$$

for all \mathbf{f} in $\mathcal{C}^2(a, b)$ and all t in $[a, b]$. This transformation is a model for a particular mass-spring system in which the spring is nonlinear. The comments under Example 7 concerning the dummy variables \mathbf{f} and t apply here as well. As usual, the definition is given in terms of scalars, functions evaluated at t . Again, \mathbf{L} acts on the whole function \mathbf{f} . Even in this example we cannot determine any value of the function $\mathbf{L}\mathbf{f}$ without using an “interval” of values of \mathbf{f} , because the derivative

function \mathbf{f}' is defined in terms of a limit of values of \mathbf{f} in the neighborhood of t :

$$\mathbf{f}'(t) \triangleq \lim_{\Delta t \rightarrow 0} \frac{\mathbf{f}(t + \Delta t) - \mathbf{f}(t)}{\Delta t}$$

The nullspace of \mathbf{L} consists in all solutions of the nonlinear differential equation, $\mathbf{L}\mathbf{f} = \theta_{\text{off}}$; restated in terms of the values of $\mathbf{L}\mathbf{f}$, this equation is

$$\mathbf{f}''(t) + \alpha(\mathbf{f}(t) + 0.01\mathbf{f}^3(t)) = 0 \quad a \leq t \leq b$$

To determine these solutions is not a simple task. By selecting $\mathcal{C}(a, b)$ as the range of definition, we ask that the function $\mathbf{L}\mathbf{f}$ be continuous; since $\mathbf{L}\mathbf{f}$ represents a force in the mass-spring system described by (2.26) continuity seems a practical assumption. By choosing $\mathcal{C}^2(a, b)$ as the domain, we guarantee that $\mathbf{L}\mathbf{f}$ is continuous. Yet the range of \mathbf{L} is not clear. It is in the range of definition, but is it equal to the range of definition? In other words, can we solve the nonlinear differential equation $\mathbf{L}\mathbf{f} = \mathbf{u}$ for *any* continuous \mathbf{u} ? The function \mathbf{f} represents the displacement versus time in the physical mass-spring system. The function \mathbf{u} represents the force applied to the system as a function of time. Physical intuition leads us to believe that for given initial conditions there is a unique displacement pattern \mathbf{f} associated with each continuous forcing pattern \mathbf{u} . Therefore, \mathbf{L} should be onto. On the other hand, since no initial conditions are specified, we expect two degrees of freedom in the solution to $\mathbf{L}\mathbf{f} = \mathbf{u}$ for each continuous \mathbf{u} . Thus the dimension of nullspace (\mathbf{L}) is two, and \mathbf{L} is not one-to-one.

Combining Transformations

The transformation introduced in Example 8 is actually a composite of several simpler transformations. In developing a model for a system, we usually start with simple models for portions of the system, and then combine the parts into the total system model. Suppose \mathbf{T} and \mathbf{U} are both transformations from \mathcal{V} into \mathcal{W} . We define the transformation $a\mathbf{T} + b\mathbf{U}$: $\mathcal{V} \rightarrow \mathcal{W}$ by

$$(a\mathbf{T} + b\mathbf{U})\mathbf{x} \triangleq a\mathbf{T}\mathbf{x} + b\mathbf{U}\mathbf{x} \quad (2.27)$$

for all \mathbf{x} in \mathcal{V} . If $\mathbf{G}: \mathcal{W} \rightarrow \mathcal{U}$, we define the transformation \mathbf{GT} : $\mathcal{V} \rightarrow \mathcal{U}$ by

$$(\mathbf{GT})\mathbf{x} \triangleq \mathbf{G}(\mathbf{T}\mathbf{x}) \quad (2.28)$$

for all \mathbf{x} in \mathcal{V} . Equations (2.27) and (2.28) define **linear combination** and **composition** of transformations, respectively.

Example 9. Composition of Matrix Multiplications. Define $\mathbf{G}: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ by

$$\mathbf{G} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} \triangleq \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 3 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}$$

and $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ by

$$\mathbf{T} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \triangleq \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

Then $\mathbf{GT}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is described by

$$\begin{aligned} \mathbf{GT} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} &= \mathbf{G} \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 3 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \\ &= \begin{pmatrix} 5 & 3 \\ 14 & 9 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \end{aligned}$$

Exercise 1. Let $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$. Show that \mathbf{T} is invertible if and only if $\mathcal{V} = \mathcal{W}$ and there is a transformation $\mathbf{T}^{-1}: \mathcal{W} \rightarrow \mathcal{V}$ such that

$$\mathbf{T}^{-1}\mathbf{T} = \mathbf{TT}^{-1} = \mathbf{I} \quad (2.29)$$

Exercise 2. Suppose \mathbf{G} and \mathbf{T} of (2.26) are invertible. Show that

$$(\mathbf{GT})^{-1} = \mathbf{T}^{-1}\mathbf{G}^{-1} \quad (2.30)$$

The composition (or product) of two transformations has two nasty characteristics. First, unlike scalars, transformations usually **do not commute**; that is, $\mathbf{GT} \neq \mathbf{TG}$. As illustrated in Example 9, \mathbf{G} and \mathbf{T} generally do not even act on the same vector space, and \mathbf{TG} has no meaning. Even if \mathbf{G} and \mathbf{T} both act on the same space, we must not expect commutability, as demonstrated by the following matrix multiplications:

$$\begin{aligned} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} &= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Commutable operators do exist. In fact, since any operator commutes with itself, we can write \mathbf{G}^2 , as we do in Example 10 below, without being ambiguous. Operators which commute act much like scalars in their behavior toward each other (see P&C 4.29).

If two scalars satisfy $ab = 0$, then either $a = 0$, $b = 0$, or both. The second matrix multiplication above demonstrates that this property does not extend even to simple transformations. This second difficulty with the composition of transformations is sometimes called the existence of **divisors of zero**. If $\mathbf{GT} = \mathbf{\Theta}$ and $\mathbf{G} \neq \mathbf{\Theta}$, we cannot conclude that $\mathbf{T} = \mathbf{\Theta}$; the cancellation laws of algebra do not apply to transformations. The difficulty lies in the fact that for transformations there is a “gray” region between being invertible and being zero. The range of \mathbf{T} can lie in the nullspace of \mathbf{G} .

Example 10. Linear Combination and Composition of Transformations. The space $\mathcal{C}^n(a, b)$ consists in all functions with continuous n th derivatives on $[a, b]$. Define $\mathbf{G}: \mathcal{C}^n(a, b) \rightarrow \mathcal{C}^{n-1}(a, b)$ by $\mathbf{Gf} \triangleq \mathbf{f}'$ for all \mathbf{f} in $\mathcal{C}^n(a, b)$. Then $\mathbf{G}^2: \mathcal{C}^2(a, b) \rightarrow \mathcal{C}(a, b)$ is well defined. Let $\mathbf{U}: \mathcal{C}^2(a, b) \rightarrow \mathcal{C}(a, b)$ be defined by $(\mathbf{Uf})(t) \triangleq \mathbf{f}(t) + 0.01\mathbf{f}^3(t)$ for all \mathbf{f} in $\mathcal{C}^2(a, b)$ and all t in $[a, b]$. The transformation \mathbf{L} of Example 8 can be described by $\mathbf{L} \triangleq \mathbf{G}^2 + \alpha\mathbf{U}$.

As demonstrated by the above examples, the domain and range of definition are essential parts of the definition of a transformation. This importance is emphasized by the notation $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$. The spaces \mathcal{V} and \mathcal{W} are selected to fit the structure of the situation we wish to model. If we pick a domain that is too large, the operator will not be one-to-one. If we pick a range of definition that is too large, the operator will not be onto. Thus both \mathcal{V} and \mathcal{W} affect the invertibility of \mathbf{T} . We apply loosely the term *finite (infinite) dimensional transformation* to those transformations that act on a finite (infinite) dimensional domain.

2.4 Linear Transformations

One of the most common and useful transformations is the matrix multiplication introduced in Chapter 1. It is well suited for automatic computation using a digital computer. Let \mathbf{A} be an $m \times n$ matrix. We define $\mathbf{T}: \mathcal{N}^{n \times 1} \rightarrow \mathcal{N}^{m \times 1}$ by

$$\mathbf{T}\mathbf{x} \triangleq \mathbf{A}\mathbf{x} \quad (2.3\ 1)$$

for all \mathbf{x} in $\mathcal{N}^{n \times 1}$. We distinguish carefully between \mathbf{T} and \mathbf{A} . \mathbf{T} is not \mathbf{A} , but rather *multiplication* by \mathbf{A} . The nullspace of \mathbf{T} is the set of solutions to

the matrix equation $\mathbf{Ax} = \mathbf{0}$. Even though \mathbf{T} and \mathbf{A} are conceptually different, we sometimes refer to the nullspace of \mathbf{T} as the nullspace of \mathbf{A} . Similarly, we define $\text{range}(\mathbf{A}) \triangleq \text{range}(\mathbf{T})$.

Suppose \mathbf{A} is square ($m = n$) and invertible; then the equation $\mathbf{T}\mathbf{x} = \mathbf{Ax} = \mathbf{y}$ has a unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ for each \mathbf{y} in $\mathfrak{N}^{n \times 1}$. But \mathbf{T}^{-1} is defined as precisely that transformation which associates with each \mathbf{y} in $\mathfrak{N}^{n \times 1}$ the unique solution to the equation $\mathbf{T}\mathbf{x} = \mathbf{y}$. Therefore, \mathbf{T} is invertible, and $\mathbf{T}^{-1}: \mathfrak{N}^{m \times 1} \rightarrow \mathfrak{N}^{n \times 1}$ is given by $\mathbf{T}^{-1}\mathbf{y} \triangleq \mathbf{A}^{-1}\mathbf{y}$.

The properties of matrix multiplication (Appendix 1) are such that $\mathbf{A}(a\mathbf{x}_1 + b\mathbf{x}_2) = a\mathbf{Ax}_1 + b\mathbf{Ax}_2$. That is, matrix multiplication preserves linear combinations. This property of matrix multiplication allows **superposition** of solutions to a matrix equation: if \mathbf{x}_1 solves $\mathbf{Ax} = \mathbf{y}_1$ and \mathbf{x}_2 solves $\mathbf{Ax} = \mathbf{y}_2$, then the solution to $\mathbf{Ax} = \mathbf{y}_1 + \mathbf{y}_2$ is $\mathbf{x}_1 + \mathbf{x}_2$. From one or two input-output relationships we can infer others. Many other familiar transformations preserve linear combinations and allow superposition of solutions.

Definition. The transformation $\mathbf{T}: \mathfrak{V} \rightarrow \mathfrak{W}$ is **linear** if

$$\mathbf{T}(a\mathbf{x}_1 + b\mathbf{x}_2) = a\mathbf{T}\mathbf{x}_1 + b\mathbf{T}\mathbf{x}_2 \tag{2.32}$$

for all vectors \mathbf{x}_1 and \mathbf{x}_2 in \mathfrak{V} and all scalars a and b .

Example 1. Integration. Define $\mathbf{T}: \mathcal{C}(0, 1) \rightarrow \mathcal{C}(0, 1)$ by

$$(\mathbf{T}\mathbf{f})(t) \triangleq \int_0^t \mathbf{f}(s) ds \tag{2.33}$$

for all \mathbf{f} in $\mathcal{C}(0, 1)$ and all t in $[0, 1]$. The linearity of this indefinite integration operation is a fundamental fact of integral calculus; that is,

$$\int_0^t [a\mathbf{f}_1(s) + b\mathbf{f}_2(s)] ds = a \int_0^t \mathbf{f}_1(s) ds + b \int_0^t \mathbf{f}_2(s) ds$$

The operator (2.33) is a special case of the linear integral operator $\mathbf{T}: \mathcal{C}(a, b) \rightarrow \mathcal{C}(c, d)$ defined by

$$(\mathbf{T}\mathbf{f})(t) \triangleq \int_a^b k(t, s)\mathbf{f}(s) ds \tag{2.34}$$

for all \mathbf{f} in $\mathcal{C}(a, b)$ and all t in $[c, d]$. We can substitute for the domain $\mathcal{C}(a, b)$ any other space of functions for which the integral exists. We can use any range of definition which includes the integrals (2.34) of all functions in the domain. The function k is called the **kernel** of the integral transformation. Another special case of (2.34) is $\mathbf{T}: \mathcal{L}_2(-\infty, \infty) \rightarrow \mathcal{L}_2(-\infty, \infty)$ defined by

$$(\mathbf{T}\mathbf{f})(t) \triangleq \int_{-\infty}^{\infty} \mathbf{g}(t-s)\mathbf{f}(s) ds$$

for some \mathbf{g} in $\mathcal{L}_2(-\infty, \infty)$, all \mathbf{f} in $\mathcal{L}_2(-\infty, \infty)$, and all t in $(-\infty, \infty)$. This \mathbf{T} is known as the convolution of \mathbf{f} with the function \mathbf{g} . It arises in connection with the solution of linear constant-coefficient differential equations (Appendix 2).

The integral transformation (2.34) is the analogue for function spaces of the matrix multiplication (2.31). That matrix transformation can be expressed

$$(\mathbf{T}\mathbf{x})_i \triangleq \sum_{j=1}^n \mathbf{A}_{ij}\xi_j \quad i=1, \dots, m \quad (2.35)$$

for all vectors \mathbf{x} in $\mathfrak{N}^{n \times 1}$. The symbol ξ_j represents the j th element of \mathbf{x} ; the symbol $(\mathbf{T}\mathbf{x})_i$ means the i th element of $\mathbf{T}\mathbf{x}$. In (2.35) the matrix is treated as a function of two discrete variables, the row variable i and the column variable j . In analogy with the integral transformation, we call the matrix multiplication [as viewed in the form of (2.35)] a **summation transformation**; we refer to the function \mathbf{A} (with values \mathbf{A}_{ij}) as the **kernel** of the summation transformation.

Example 2. Differentiation Define $\mathbf{D}: \mathcal{C}^1(a, b) \rightarrow \mathcal{C}(a, b)$ by

$$(\mathbf{D}\mathbf{f})(t) \triangleq \mathbf{f}'(t) \triangleq \lim_{\Delta t \rightarrow 0} \frac{\mathbf{f}(t + \Delta t) - \mathbf{f}(t)}{\Delta t} \quad (2.36)$$

for all \mathbf{f} in $\mathcal{C}^1(a, b)$ and all t in $[a, b]$; $\mathbf{f}'(t)$ is the slope of the graph of \mathbf{f} at t ; \mathbf{f}' (or $\mathbf{D}\mathbf{f}$) is the whole "slope" function. We also use the symbols $\dot{\mathbf{f}}$ and $\mathbf{f}^{(1)}$ in place of $\mathbf{D}\mathbf{f}$. We can substitute for the above domain and range of definition any pair of function spaces for which the derivatives of all functions in the domain lie in the range of definition. Thus we could define \mathbf{D} on $\mathcal{C}(a, b)$ if we picked a range of definition which contains the appropriate discontinuous functions. The nullspace of \mathbf{D} is $\text{span}\{\mathbf{1}\}$, where $\mathbf{1}$ is the function defined by $\mathbf{1}(t) = 1$ for all t in $[a, b]$. It is well known that differentiation is linear; $\mathbf{D}(c_1\mathbf{f}_1 + c_2\mathbf{f}_2) = c_1\mathbf{D}\mathbf{f}_1 + c_2\mathbf{D}\mathbf{f}_2$.

We can define more general differential operators in terms of (2.36). The general linear constant-coefficient differential operator $\mathbf{L}: \mathcal{C}^n(a, b) \rightarrow \mathcal{C}(a, b)$ is defined, for real scalars $\{a_i\}$, by

$$\mathbf{L} \triangleq \mathbf{D}^n + a_1\mathbf{D}^{n-1} + \dots + a_n\mathbf{I} \quad (2.37)$$

where we have used (2.27) and (2.28) to combine transformations. A **variable-coefficient** (or "time-varying") extension of (2.37) is the operator $\mathbf{L}: \mathcal{C}^n(a, b) \rightarrow \mathcal{C}(a, b)$ defined by*

$$(\mathbf{L}\mathbf{f})(t) \triangleq g_0(t)\mathbf{f}^{(n)}(t) + g_1(t)\mathbf{f}^{(n-1)}(t) + \dots + g_n(t)\mathbf{f}(t) \quad (2.37)$$

*Note that we use boldface print for some of the functions in (2.38) but not for others. As indicated in the Preface, we use boldface print only to emphasize the vector or transformation interpretation of an object. We sometimes describe the same function both ways, \mathbf{f} and f .

for all \mathbf{f} in $\mathcal{C}^n(a, b)$ and all t in $[a, b]$. (We have denoted the k th derivative $\mathbf{D}^k \mathbf{f}$ by $\mathbf{f}^{(k)}$.) If the interval $[a, b]$ is finite, if the functions g_i are continuous, and if $g_0(t) \neq 0$ on $[a, b]$, we refer to (2.38) as a regular n th-order differential operator. [With $g_0(t) \neq 0$, we would lose no generality by letting $g_0(t) = 1$ in (2.38).] We can apply the differential operators (2.37) and (2.38) to other function spaces than $\mathcal{C}^n(a, b)$.

Example 3. Evaluation of a Function. Define $\mathbf{T}: \mathcal{C}(a, b) \rightarrow \mathfrak{R}^1$ by

$$\mathbf{T}\mathbf{f} \triangleq \mathbf{f}(t_1) \quad (2.39)$$

for all \mathbf{f} in the function space $\mathcal{C}(a, b)$. In this example, \mathbf{f} is a dummy variable, but t_1 is not. The transformation is a *linear functional* called "evaluation at t_1 ." The range of \mathbf{T} is \mathfrak{R}^1 ; \mathbf{T} is onto. The nullspace of \mathbf{T} is the set of continuous functions which pass through zero at t_1 . Because many functions have the same value at t_1 , \mathbf{T} is not one-to-one. This functional can also be defined using some other function space for its domain.

Example 4. A One-Sided Laplace Transform, \mathcal{L} . Suppose \mathfrak{W} is the space of complex-valued functions defined on the positive-real half of the complex plane. (See Example 10, Section 2.1.) Let \mathfrak{V} be the space of functions which are defined and continuous on $[0, \infty]$ and for which $e^{-ct}|f(t)|$ is bounded for some constant c and all values of t greater than some finite number. We define the one-sided Laplace transform $\mathcal{L}: \mathfrak{V} \rightarrow \mathfrak{W}$ by

$$(\mathcal{L}\mathbf{f})(s) \triangleq \int_0^{\infty} e^{-st} \mathbf{f}(t) dt \quad (2.40)$$

for all complex s with $\text{real}(s) > 0$. The functions in \mathfrak{V} are such that (2.40) converges for $\text{real}(s) > 0$. We sometimes denote the transformed function $\mathcal{L}\mathbf{f}$ by \mathbf{F} . This integral transform, like that of (2.34), is linear. The Laplace transform is used to convert linear constant-coefficient differential equations into linear algebraic equations. •

Exercise 1. Suppose the transformations \mathbf{T} , \mathbf{U} , and \mathbf{G} of (2.27) and (2.28) are linear and \mathbf{T} is invertible. Show that the transformations $a\mathbf{T} + b\mathbf{U}$, $\mathbf{G}\mathbf{T}$, and \mathbf{T}^{-1} are also linear.

Exercise 2. Let \mathfrak{V} be an n -dimensional linear space with basis \mathfrak{X} . Define $\mathbf{T}: \mathfrak{V} \rightarrow \mathfrak{R}^{n \times 1}$ by

$$\mathbf{T}\mathbf{x} \triangleq [\mathbf{x}]_{\mathfrak{X}} \quad (2.41)$$

Show that \mathbf{T} , the process of taking coordinates, is a linear, invertible transformation.

*It can be shown that $[\mathcal{L}(\mathbf{D}\mathbf{f})](s) = s(\mathcal{L}\mathbf{f})(s) - \mathbf{f}(0^+)$, where $\mathbf{f}(0^+)$ is the limit of $\mathbf{f}(t)$ as $t \rightarrow 0$ from the positive side of 0.

The vector space \mathcal{V} of Exercise 2 is equivalent to $\mathfrak{N}^{n \times 1}$ in every sense we might wish. The linear, invertible transformation is the key. We say two vector spaces \mathcal{V} and \mathcal{W} are **isomorphic** (or equivalent) if there exists an invertible linear transformation from \mathcal{V} into \mathcal{W} . Each real n -dimensional vector space is isomorphic to each other real n -dimensional space and, in particular, to the real space $\mathfrak{N}^{n \times 1}$. A similar statement can be made using complex scalars for each space. Infinite-dimensional spaces also exhibit isomorphism. In Section 5.3 we show that all well behaved infinite-dimensional spaces are isomorphic to l_2 .

Nullspace and Range—Keys to Invertibility

Even linear transformations may have troublesome properties. In point of fact, the example in which we demonstrate *noncommutability* and *noncancellation* of products of transformations uses linear transformations (matrix multiplications). Most difficulties with a linear transformation can be understood through investigation of the range and nullspace of the transformation.*

Let $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ be linear. Suppose \mathbf{x}_h is a vector in the nullspace of \mathbf{T} (any solution to $\mathbf{T}\mathbf{x} = \mathbf{0}$); we call \mathbf{x}_h a homogeneous solution for the transformation \mathbf{T} . Denote by \mathbf{x}_p a particular solution to the equation $\mathbf{T}\mathbf{x} = \mathbf{y}$. (An \mathbf{x}_p exists if and only if \mathbf{y} is in **range**(\mathbf{T} .) Then $\mathbf{x}_p + \alpha\mathbf{x}_h$ is also a solution to $\mathbf{T}\mathbf{x} = \mathbf{y}$ for any scalar α . One of the most familiar uses of the principle of superposition is in obtaining the general solution to a linear differential equation by combining particular and homogeneous solutions. The general solution to any linear operator equation can be obtained in this manner.

Example 5. The General Solution to a Matrix Equation. Define the linear operator $\mathbf{T}: \mathfrak{N}^{2 \times 1} \rightarrow \mathfrak{N}^{2 \times 1}$ by

$$\mathbf{T} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \triangleq \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

Then the equation

$$\mathbf{T}\mathbf{x} = \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \triangleq \mathbf{y} \quad (2.42)$$

has as its general solution $\mathbf{x} = (\xi_1 \ 2 - 2\xi_1)^T$. A particular solution is $\mathbf{x}_p = (1 \ 0)^T$. The nullspace of \mathbf{T} consists in the vector $\mathbf{x}_h = (-1 \ 2)^T$ and all its multiples. The general solution can be expressed as $\mathbf{x} = \mathbf{x}_p + \alpha\mathbf{x}_h$ where α is arbitrary. Figure 2.7 shows an

*See Sections 4.4 and 4.6 for further insight into noncancellation and noncommutability of linear operators.

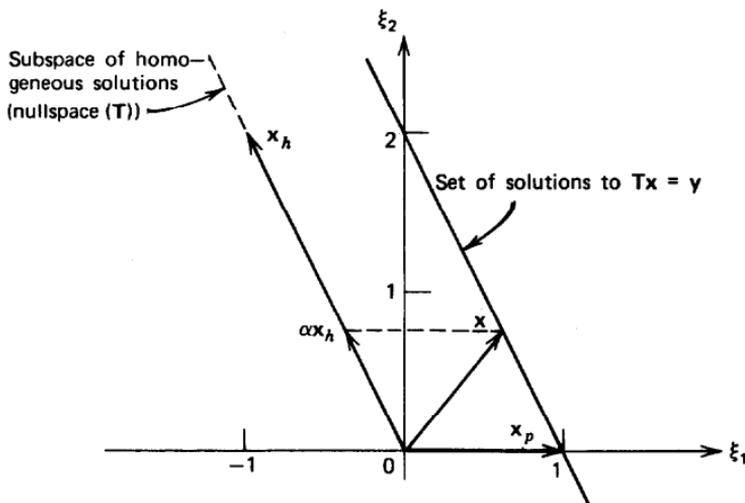


Figure 2.7. Solutions to the linear equation of Example 5.

arrow-space equivalent of these vectors. The nullspace of \mathbf{T} is a subspace of $\mathfrak{N}^{2 \times 1}$. The general solution (the set of all solutions to $\mathbf{T}\mathbf{x} = \mathbf{y}$) consists of a line in $\mathfrak{N}^{2 \times 1}$; specifically, it is the nullspace of \mathbf{T} shifted by the addition of any particular solution.

The nullspace of a linear transformation is always a subspace of the domain \mathfrak{V} . The freedom in the general solution to $\mathbf{T}\mathbf{x} = \mathbf{y}$ lies only in $\text{nullspace}(\mathbf{T})$, the subspace of homogeneous solutions. For if $\hat{\mathbf{x}}_p$ is another particular solution to $\mathbf{T}\mathbf{x} = \mathbf{y}$, then

$$\mathbf{T}(\mathbf{x}_p - \hat{\mathbf{x}}_p) = \mathbf{T}\mathbf{x}_p - \mathbf{T}\hat{\mathbf{x}}_p = \mathbf{y} - \mathbf{y} = \mathbf{0}$$

The difference between \mathbf{x}_p and $\hat{\mathbf{x}}_p$ is a vector in $\text{nullspace}(\mathbf{T})$. If $\text{nullspace}(\mathbf{T}) = \mathbf{0}$, there is no freedom in the solution to $\mathbf{T}\mathbf{x} = \mathbf{y}$; it is unique.

Definition. A transformation $\mathbf{G}: \mathfrak{V} \rightarrow \mathfrak{W}$ is nonsingular if $\text{nullspace}(\mathbf{G}) = \mathbf{0}$.

Exercise 3. Show that a *linear* transformation is one-to-one if and only if it is nonsingular.

Because a linear transformation $\mathbf{T}: \mathfrak{V} \rightarrow \mathfrak{W}$ preserves linear combinations, it necessarily transforms $\mathbf{0}_{\mathfrak{V}}$ into $\mathbf{0}_{\mathfrak{W}}$. Furthermore, \mathbf{T} acts on the vectors in \mathfrak{V} by subspaces—whatever \mathbf{T} does to \mathbf{x} it does also to $c\mathbf{x}$, where c is any scalar. The set of vectors in \mathfrak{V} which are taken to zero, for example, is the subspace which we call $\text{nullspace}(\mathbf{T})$. Other subspaces of \mathfrak{V} are “rotated” or “stretched” by \mathbf{T} . This fact becomes more clear during our discussion of spectral decomposition in Chapter 4.

Example 6. The Action of a Linear Transformation on Subspaces. Define $\mathbf{T}: \mathcal{R}^3 \rightarrow \mathcal{R}^2$ by $\mathbf{T}(\xi_1, \xi_2, \xi_3) \stackrel{\Delta}{=} (\xi_3, 0)$. The set $\{\mathbf{x}_1 = (1, 0, 0), \mathbf{x}_2 = (0, 1, 0)\}$ forms a basis for $\text{nullspace}(\mathbf{T})$. By adding a third independent vector, say, $\mathbf{x}_3 = (1, 1, 1)$, we obtain a basis for the domain \mathcal{R}^3 . The subspace spanned by $\{\mathbf{x}_1, \mathbf{x}_2\}$ is annihilated by \mathbf{T} . The subspace spanned by $\{\mathbf{x}_3\}$ is transformed by \mathbf{T} into a subspace of \mathcal{R}^2 —the range of \mathbf{T} . The vector \mathbf{x}_3 itself is transformed into a basis for $\text{range}(\mathbf{T})$. Because \mathbf{T} acts on the vectors in \mathcal{R}^3 by subspaces, the dimension of $\text{nullspace}(\mathbf{T})$ is a measure of the degree to which \mathbf{T} acts like zero; the dimension of $\text{range}(\mathbf{T})$ indicates the degree to which \mathbf{T} acts invertible. Specifically, of the three dimensions in \mathcal{R}^3 , \mathbf{T} takes two to zero. The third dimension of \mathcal{R}^3 is taken into the one-dimensional $\text{range}(\mathbf{T})$.

The characteristics exhibited by Example 6 extend to any linear transformation on a finite-dimensional space. Let $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ be linear with $\dim(\mathcal{V}) = n$. We call the dimension of $\text{nullspace}(\mathbf{T})$ the **nullity of \mathbf{T}** . The **rank of \mathbf{T}** is the dimension of $\text{range}(\mathbf{T})$. Let $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ be a basis for $\text{nullspace}(\mathbf{T})$. Pick vectors $\{\mathbf{x}_{k+1}, \dots, \mathbf{x}_n\}$ which extend the basis for $\text{nullspace}(\mathbf{T})$ to a basis for \mathcal{V} (P&C 2.9). We show that \mathbf{T} takes $\{\mathbf{x}_{k+1}, \dots, \mathbf{x}_n\}$ into a basis for $\text{range}(\mathbf{T})$. Suppose $\mathbf{x} = c_1\mathbf{x}_1 + \dots + c_n\mathbf{x}_n$ is an arbitrary vector in \mathcal{V} . The linear transformation \mathbf{T} annihilates the first k components of \mathbf{x} . Only the remaining $n - k$ components are taken into $\text{range}(\mathbf{T})$. Thus the vectors $\{\mathbf{T}\mathbf{x}_{k+1}, \dots, \mathbf{T}\mathbf{x}_n\}$ must span $\text{range}(\mathbf{T})$. To show that these vectors are independent, we use the test (2.11):

$$\xi_{k+1}(\mathbf{T}\mathbf{x}_{k+1}) + \dots + \xi_n(\mathbf{T}\mathbf{x}_n) = \theta_{\mathcal{W}}$$

Since \mathbf{T} is linear,

$$\mathbf{T}(\xi_{k+1}\mathbf{x}_{k+1} + \dots + \xi_n\mathbf{x}_n) = \theta_{\mathcal{W}}$$

Then $\xi_{k+1}\mathbf{x}_{k+1} + \dots + \xi_n\mathbf{x}_n$ is in $\text{nullspace}(\mathbf{T})$, and

$$\xi_{k+1}\mathbf{x}_{k+1} + \dots + \xi_n\mathbf{x}_n = d_1\mathbf{x}_1 + \dots + d_k\mathbf{x}_k$$

for some $\{d_i\}$. The independence of $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ implies $d_1 = \dots = d_k = \xi_{k+1} = \dots = \xi_n = 0$; thus $\{\mathbf{T}\mathbf{x}_{k+1}, \dots, \mathbf{T}\mathbf{x}_n\}$ is an independent set and is a basis for $\text{range}(\mathbf{T})$.

We have shown that a linear transformation \mathbf{T} acting on a finite-dimensional space \mathcal{V} obeys a “conservation of dimension” law:

$$\dim(\mathcal{V}) = \text{rank}(\mathbf{T}) + \text{nullity}(\mathbf{T}) \quad (2.43)$$

Nullity(\mathbf{T}) is the “dimension” annihilated by \mathbf{T} . **Rank(\mathbf{T})** is the “dimension” \mathbf{T} retains. If $\text{nullspace}(\mathbf{T}) = \{\theta\}$, then $\text{nullity}(\mathbf{T}) = 0$ and $\text{rank}(\mathbf{T}) = \dim(\mathcal{V})$. If, in addition, $\dim(\mathcal{W}) = \dim(\mathcal{V})$, then $\text{rank}(\mathbf{T}) = \dim(\mathcal{W})$ (\mathbf{T} is

onto), and \mathbf{T} is invertible. A linear $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ cannot be invertible unless $\dim(\mathcal{W}) = \dim(\mathcal{V})$.

We sometimes refer to the vectors $\mathbf{x}_{k+1}, \dots, \mathbf{x}_n$ as **progenitors of the range of \mathbf{T}** . Although the nullspace and range of \mathbf{T} are unique, the space spanned by the progenitors is not; we can add any vector in nullspace to any progenitor without changing the basis for the range (see Example 6).

The Near Nullspace

In contrast to mathematical analysis, mathematical *computation* is not clear-cut. For example, a set of equations which is mathematically invertible can be so “nearly singular” that the inverse cannot be computed to an acceptable degree of precision. On the other hand, because of the finite number of significant digits used in the computer, a mathematically singular system will be indistinguishable from a “nearly singular” system. The phenomenon merits serious consideration.

The matrix operator of Example 5 is singular. Suppose we modify the matrix slightly to obtain the nonsingular, but “nearly singular” matrix equation

$$\begin{pmatrix} 2 & 1 \\ 2 & 1 + \epsilon \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \quad (2.44)$$

where ϵ is small. Then the arrow space diagram of Figure 2.7 must also be modified to show a pair of almost parallel lines. (Figure 1.7 of Section 1.5 is the arrow space diagram of essentially this pair of equations.) Although the solution (the intersection of the nearly parallel lines) is unique, it is difficult to compute accurately; the nearly singular equations are very ill conditioned. Slight errors in the data and roundoff during computing lead to significant uncertainty in the computed solution, even if the computation is handled carefully (Section 1.5). The uncertain component of the solution lies essentially in the nullspace of the operator; that is, it is almost parallel to the nearly parallel lines in the arrow-space diagram. The above pair of nearly singular algebraic equations might represent a nearly singular system. On the other hand, the underlying system might be precisely singular; the equations in the model of a singular system may be only *nearly singular* because of inaccuracies in the data. Regardless of which of these interpretations is correct, determining the “near nullspace” of the matrix is an important part of the analysis of the system. If the underlying system is singular, a description of the near nullspace is a description of the *freedom* in the solutions for the system. If the underlying system is just nearly singular, a description of the near nullspace is a description of the *uncertainty* in the solution.

Definition. Suppose \mathbf{T} is a *nearly singular* linear operator on a vector space \mathcal{V} . We use the term **near nullspace of \mathbf{T}** to mean those vectors that are taken *nearly* to zero by \mathbf{T} ; that is, those vectors which \mathbf{T} drastically reduces in “size.”*

In the two-dimensional example described above, the near nullspace consists in vectors which are *nearly* parallel to the vector $\mathbf{x} = (-1 \ 2)^T$. The near nullspace of \mathbf{T} is *not a subspace* of \mathcal{V} . Rather, it consists in a set of vectors which are *nearly* in a subspace of \mathcal{V} . We can think of the near nullspace as a “fuzzy” subspace of \mathcal{V} .

We now present a method, referred to as **inverse iteration**, for describing the near nullspace of a nearly singular operator \mathbf{T} acting on a vector space \mathcal{V} . Let \mathbf{x}_0 be an arbitrary vector in \mathcal{V} . Assume \mathbf{x}_0 contains a component which is in the near nullspace of \mathbf{T} . (If it does not, such a component will be introduced by roundoff during the ensuing computation.) Since \mathbf{T} reduces such components drastically, compared to its effect on the other components of \mathbf{x}_0 , \mathbf{T}^{-1} must drastically emphasize such components. Therefore, if we solve $\mathbf{T}\mathbf{x}_1 = \mathbf{x}_0$ (in effect determining $\mathbf{x}_1 = \mathbf{T}^{-1}\mathbf{x}_0$), the computed solution \mathbf{x}_1 contains a *significant* component in the near nullspace of \mathbf{T} . (This component is the error vector which appears during the solution of the nearly singular equation.) The inverse iteration method consists in iteratively solving $\mathbf{T}\mathbf{x}_{k+1} = \mathbf{x}_k$. After a few iterations, \mathbf{x}_k is dominated by its near-nullspace component; we use \mathbf{x}_k as a partial basis for the near nullspace of \mathbf{T} . (The number of iterations required is at the discretion of the analyst. We are not looking for a precisely defined subspace, but rather, a subspace that is fuzzy.) By repeating the above process for several different starting vectors \mathbf{x}_0 , we usually obtain a set of vectors which spans the near nullspace of \mathbf{T} .

Example 7. Describing a Near Nullspace. Define a linear operator \mathbf{T} on $\mathcal{R}^{2 \times 1}$ by means of the nearly singular matrix multiplication described above:

$$\mathbf{T}\mathbf{x} \triangleq \begin{pmatrix} 2 & 1 \\ 2 & 1+\epsilon \end{pmatrix} \mathbf{x}$$

For this simple example we can invert \mathbf{T} explicitly

$$\mathbf{T}^{-1}\mathbf{x} = \frac{1}{2\epsilon} \begin{pmatrix} 1+\epsilon & -1 \\ -2 & 2 \end{pmatrix} \mathbf{x}$$

We apply the inverse iteration method to the vector $\mathbf{x}_0 = (1 \ 1)^T$; of course, we have no roundoff in our computations:

$$\mathbf{x}_1 = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = \frac{1}{2\epsilon} \begin{pmatrix} (1+\epsilon)/2 \\ -1 \end{pmatrix}, \quad \mathbf{x}_3 = \frac{1}{(2\epsilon)^2} \begin{pmatrix} (\epsilon^2 + 2\epsilon + 3)/2 \\ -(\epsilon + 3) \end{pmatrix}, \dots$$

*In Section 4.2 we describe the near nullspace more precisely as the eigenspace for the smallest eigenvalue of \mathbf{T} .

If ϵ is small, say $\epsilon = 0.01$, then

$$\mathbf{x}_2 = 50 \begin{pmatrix} 0.505 \\ -1 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_3 = (50)^2 \begin{pmatrix} 1.51 \\ -3.01 \end{pmatrix}$$

After only three iterations, the sequence \mathbf{x}_k has settled; the vector \mathbf{x}_3 provides a good description of the near nullspace of \mathbf{T} . If $\epsilon = 0$, \mathbf{T} is singular; \mathbf{x}_3 lies almost in the nullspace of this singular operator (Figure 2.7). Were we to try other starting vectors \mathbf{x}_0 , we would obtain other vectors \mathbf{x}_k nearly parallel to $(-1 \ 2)^T$. This near nullspace of \mathbf{T} should be considered one-dimensional.

We note from Example 7 that the vector \mathbf{x}_k in the inverse iteration grows drastically in size. Practical computer implementations of inverse iteration include normalization of \mathbf{x}_k at each step in order to avoid numbers too large for the computer. A description for a two-dimensional near nullspace is sought in P&C 2.26. In Section 4.2 we analyze the inverse iteration more precisely in terms of eigenvalues and eigenvectors. Forsythe [2.3] gives some interesting examples of the treatment of nearly singular operators.

The Role of Linear Transformations

The purpose of modeling a system is to develop insight concerning the system, to develop an intuitive feel for the input-output relationship. In order to decide whether or not a particular model, linear or nonlinear, is a good model, we must compare the input-output relationship of the model with the corresponding, but measurable, input-output relationship of the system being modeled. If the model and the system are sufficiently in agreement for our purposes, we need not distinguish between the system and the model.

Almost all physical systems are to some degree nonlinear. Yet most systems act in a nearly linear manner if the range of variation of the variables is restricted. For example, the current through a resistor is essentially proportional to the applied voltage if the current is not large enough to heat the resistor significantly. We are able to develop adequate models for a wide variety of static and dynamic physical systems using only linear transformations. For linear models there is available a vast array of mathematical results; most mathematical analysis is linear analysis. Furthermore, the analysis or optimization of a *nonlinear* system is usually based on linearization (Chapters 7 and 8). Even in solving a nonlinear equation for a given input, we typically must resort to repetitive linearization.

The examples and exercises of this section have demonstrated the variety of familiar transformations which are linear: matrix multiplication, differentiation, integration, etc. We introduce other linear transformations

as we need them. The next few chapters pertain only to linear transformations. In Chapter 3 we focus on the peculiarities of linear differential systems. In Chapter 4 we develop the concepts of spectral decomposition of linear systems. The discussion of infinite-dimensional systems in Chapter 5 is also directed toward linear systems. Because we use the symbols \mathbf{T} and \mathbf{U} so much in reference to linear transformations, hereinafter we employ the symbols \mathbf{F} and \mathbf{G} to emphasize concepts which apply as well to nonlinear transformations. We begin to examine nonlinear concepts in Chapter 6. We do not return fully to the subject of nonlinear systems, however, until we introduce the concepts of linearization and repetitive linearization in Chapters 7 and 8.

2.5 Matrices of Linear Transformations

By the process of picking an ordered basis for an n -dimensional vector space \mathcal{V} , we associate with each vector in \mathcal{V} a unique $n \times 1$ column matrix. In effect, we convert the vectors in \mathcal{V} into an equivalent set of vectors which are suitable for matrix manipulation and, therefore, automatic computation by computer. By taking coordinates, we can also convert a linear equation, $\mathbf{T}\mathbf{x} = \mathbf{y}$, into a matrix equation. Suppose $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ is a linear transformation, $\dim(\mathcal{V}) = n$, and $\dim(\mathcal{W}) = m$. Pick as bases for \mathcal{V} and \mathcal{W} the sets $\mathcal{X} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and $\mathcal{Y} \triangleq \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$, respectively. The vectors \mathbf{x} in \mathcal{V} and $\mathbf{T}\mathbf{x}$ in \mathcal{W} can be represented by their coordinate matrices $[\mathbf{x}]_{\mathcal{X}}$ and $[\mathbf{T}\mathbf{x}]_{\mathcal{Y}}$. The vectors \mathbf{x} and $\mathbf{T}\mathbf{x}$ are linearly related (by the linear transformation \mathbf{T}). By (2.41), we know that a vector and its coordinates are also linearly related. Therefore, we expect $[\mathbf{x}]_{\mathcal{X}}$ and $[\mathbf{T}\mathbf{x}]_{\mathcal{Y}}$ to be linearly related as well. Furthermore, we intuitively expect the linear relation between the $n \times 1$ matrix $[\mathbf{x}]_{\mathcal{X}}$ and the $m \times 1$ matrix $[\mathbf{T}\mathbf{x}]_{\mathcal{Y}}$ to be multiplication by an $m \times n$ matrix. We denote this matrix by $[\mathbf{T}]_{\mathcal{Y}\mathcal{X}}$ and refer to it as the **matrix of \mathbf{T} relative to the ordered bases \mathcal{X} and \mathcal{Y}** ; it must satisfy

$$[\mathbf{T}]_{\mathcal{Y}\mathcal{X}} [\mathbf{x}]_{\mathcal{X}} \stackrel{\Delta}{=} [\mathbf{T}\mathbf{x}]_{\mathcal{Y}} \quad (2.45)$$

for all \mathbf{x} in \mathcal{V} . Assume we can find such a matrix. Then by taking coordinates (with respect to \mathcal{Y}) of each side of the linear equation $\mathbf{T}\mathbf{x} = \mathbf{y}$, we convert the equation to the equivalent matrix equation.

$$[\mathbf{T}]_{\mathcal{Y}\mathcal{X}} [\mathbf{x}]_{\mathcal{X}} = [\mathbf{y}]_{\mathcal{Y}} \quad (2.46)$$

We will show that we can represent any linear transformation of \mathcal{V} into \mathcal{W} by a matrix multiplication by selecting bases for \mathcal{V} and \mathcal{W} —we can

convert any linear equation involving finite-dimensional vector spaces into a matrix equation. We first show how to determine the matrix of \mathbf{T} , then we show that it satisfies the defining equation (2.45) for all vectors \mathbf{x} in \mathcal{V} .

Example 1. Determining the Matrix of a Linear Transformation Let $\mathbf{x} = (\xi_1, \xi_2, \xi_3)$, an arbitrary vector in \mathcal{R}^3 . Define $\mathbf{T}: \mathcal{R}^3 \rightarrow \mathcal{R}^2$ by

$$\mathbf{T}(\xi_1, \xi_2, \xi_3) \triangleq (2\xi_2 - \xi_1, \xi_1 + \xi_2 + \xi_3)$$

We now find $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$, where \mathcal{E}_3 and \mathcal{E}_2 are the standard bases for \mathcal{R}^3 and \mathcal{R}^2 , respectively. By (2.45), we have

$$[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}[(\xi_1, \xi_2, \xi_3)]_{\mathcal{E}_3} = [(2\xi_2 - \xi_1, \xi_1 + \xi_2 + \xi_3)]_{\mathcal{E}_2}$$

for all vectors (ξ_1, ξ_2, ξ_3) , or

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 2\xi_2 - \xi_1 \\ \xi_1 + \xi_2 + \xi_3 \end{pmatrix} \quad (2.47)$$

where we have used $\{a_{ij}\}$ to represent the elements of $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$. By making three independent choices of the scalars ξ_1 , ξ_2 , and ξ_3 , we could convert this matrix equation into six equations in the six unknowns $\{a_{ij}\}$. However, by using a little ingenuity, we reduce this effort. Think of the matrix multiplication in terms of the columns of $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$. The i th element of $[\mathbf{x}]_{\mathcal{E}_3}$ multiplies the i th column of $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$. If we choose $\mathbf{x} = (1, 0, 0)$, then $[(1, 0, 0)]_{\mathcal{E}_3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, and (2.47) becomes

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

We have found the first column of $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$ directly. We obtain the other two columns of $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$ from (2.47) by successive substitution of $\mathbf{x} = (0, 1, 0)$ and $\mathbf{x} = (0, 0, 1)$. The result is

$$[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2} = \begin{pmatrix} -1 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

In Example 1 we avoided the need for simultaneous equations by substituting the basis vectors \mathbf{e}_1 , \mathbf{e}_2 , and \mathbf{e}_3 into (2.47) to pick out the columns of $[\mathbf{T}]_{\mathcal{E}_3, \mathcal{E}_2}$. This same technique can be used to find the matrix of any linear transformation acting on a finite-dimensional space. We refer again to $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$, with $\dim(\mathcal{V}) = n$, $\dim(\mathcal{W}) = m$, \mathcal{X} a basis for \mathcal{V} , and \mathcal{Y} a basis for \mathcal{W} . If we substitute into (2.45) the vector \mathbf{x}_i , the i th vector of

the basis \mathcal{X} , we pick out the i th column of $[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}$:

$$[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}[\mathbf{x}_i]_{\mathcal{X}} = [\mathbf{T}]_{\mathcal{X}\mathcal{Y}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} = i\text{th column of } [\mathbf{T}]_{\mathcal{X}\mathcal{Y}} = [\mathbf{T}\mathbf{x}_i]_{\mathcal{Y}}$$

We can find each column of $[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}$ independently. The only computational effort is that in determining the coordinate matrices $[\mathbf{T}\mathbf{x}_i]_{\mathcal{Y}}$. Therefore,

$$[\mathbf{T}]_{\mathcal{X}\mathcal{Y}} = ([\mathbf{T}\mathbf{x}_1]_{\mathcal{Y}} \vdots [\mathbf{T}\mathbf{x}_2]_{\mathcal{Y}} \vdots \cdots \vdots [\mathbf{T}\mathbf{x}_n]_{\mathcal{Y}}) \quad (2.48)$$

Example 2. The Matrix of a Linear Operator. Define the differential operator $\mathbf{D}: \mathcal{P}^3 \rightarrow \mathcal{P}^3$ as in (2.36). The set $\mathcal{X} \triangleq \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$, where $\mathbf{f}_1(t) = 1$, $\mathbf{f}_2(t) = t$, $\mathbf{f}_3(t) = t^2$, is a natural basis for \mathcal{P}^3 . We use (2.48) to find

$$\begin{aligned} [\mathbf{D}]_{\mathcal{X}\mathcal{X}} &= ([\mathbf{D}\mathbf{f}_1]_{\mathcal{X}} \vdots [\mathbf{D}\mathbf{f}_2]_{\mathcal{X}} \vdots [\mathbf{D}\mathbf{f}_3]_{\mathcal{X}}) \\ &= ([\mathbf{0}]_{\mathcal{X}} \vdots [\mathbf{f}_1]_{\mathcal{X}} \vdots [2\mathbf{f}_2]_{\mathcal{X}}) \\ &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

From the method used to determine $[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}$ in (2.48), we know that this matrix correctly represents the action of \mathbf{T} on the basis vectors $\{\mathbf{x}_i\}$. We now show that the matrix (2.48) also represents correctly the action of \mathbf{T} on all other vectors in \mathcal{V} . An arbitrary vector \mathbf{x} in \mathcal{V} may be written in terms of the basis vectors for \mathcal{V} :

$$\mathbf{x} = \sum_{i=1}^n \xi_i \mathbf{x}_i$$

Since the transformation \mathbf{T} is linear,

$$\mathbf{T}\mathbf{x} = \sum_{i=1}^n \xi_i \mathbf{T}\mathbf{x}_i$$

Because the process of taking coordinates is linear [see (2.41)],

$$\begin{aligned} [\mathbf{T}\mathbf{x}]_{\mathcal{Y}} &= \sum_{i=1}^n \xi_i [\mathbf{T}\mathbf{x}_i]_{\mathcal{Y}} \\ &= ([\mathbf{T}\mathbf{x}_1]_{\mathcal{Y}} \vdots \cdots \vdots [\mathbf{T}\mathbf{x}_n]_{\mathcal{Y}}) \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \\ &= [\mathbf{T}]_{\mathcal{Y}\mathcal{X}} [\mathbf{x}]_{\mathcal{X}} \end{aligned}$$

Thus, continuing Example 2 above, if \mathbf{f} is the arbitrary vector defined by $\mathbf{f}(t) \triangleq \xi_1 + \xi_2 t + \xi_3 t^2$, then

$$(\mathbf{D}\mathbf{f})(t) = \xi_2 + 2\xi_3 t, [\mathbf{f}]_{\mathcal{X}} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}, [\mathbf{D}\mathbf{f}]_{\mathcal{X}} = \begin{pmatrix} \xi_2 \\ 2\xi_3 \\ 0 \end{pmatrix}, \text{ and } [\mathbf{D}]_{\mathcal{X}\mathcal{X}} [\mathbf{f}]_{\mathcal{X}} = [\mathbf{D}\mathbf{f}]_{\mathcal{X}}$$

When the domain and range space of \mathbf{T} are identical, and the same basis is used for both spaces (as it is in Example 2), we sometimes refer to the matrix $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$ as the **matrix of the operator \mathbf{T} relative to the basis \mathcal{X}** .

We expect the matrix of a linear transformation to possess the basic characteristics of that transformation. The only basic characteristics of a linear transformation that we have discussed thus far are its rank and nullity. The picking of coordinate systems \mathcal{X} and \mathcal{Y} converts the transformation equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ to a precisely equivalent matrix equation, $[\mathbf{T}\mathbf{x}]_{\mathcal{Y}} = [\mathbf{T}]_{\mathcal{Y}\mathcal{X}} [\mathbf{x}]_{\mathcal{X}} = [\mathbf{y}]_{\mathcal{Y}}$; for every \mathbf{x} and \mathbf{y} in the one equation, there is a unique $[\mathbf{x}]_{\mathcal{X}}$ and $[\mathbf{y}]_{\mathcal{Y}}$ in the other. The dimensions of the nullspace and range of the transformation “multiplication by $[\mathbf{T}]_{\mathcal{Y}\mathcal{X}}$ ” must be the same, therefore, as the dimensions of the nullspace and range of \mathbf{T} . We speak loosely of the rank and nullity of $[\mathbf{T}]_{\mathcal{Y}\mathcal{X}}$ when we actually mean the rank and nullity of the transformation “multiplication by $[\mathbf{T}]_{\mathcal{Y}\mathcal{X}}$.” We refer to the nullity and rank of a matrix as if it were the matrix of a linear transformation. The nullspace and range of matrix multiplications are explored in P&C 2.19; the problem demonstrates that for an $m \times n$ matrix \mathbf{A} ,

$$\begin{aligned} \text{rank}(\mathbf{A}) &= \text{the number of independent columns of } \mathbf{A} \\ &= \text{the number of independent rows of } \mathbf{A} \end{aligned}$$

$$\text{nullity}(\mathbf{A}) = n - \text{rank}(\mathbf{A})$$

$$\text{nullity}(\mathbf{A}^T) = m - \text{rank}(\mathbf{A})$$

Once again referring to Example 2, we see that the nullity of \mathbf{D} is 1 [the vector \mathbf{f}_1 is a basis for $\text{nullspace}(\mathbf{D})$]. The nullity of $[\mathbf{D}]_{\mathcal{X}\mathcal{X}}$ is also 1 ($[\mathbf{D}]_{\mathcal{X}\mathcal{X}}$ contains one dependent column). The matrix $[\mathbf{D}]_{\mathcal{X}\mathcal{X}}$ does possess the same nullity and rank as the operator \mathbf{D} .

It is apparent that determination of the matrix of a transformation reduces to the determination of coordinate matrices for the set of vectors $\{\mathbf{T}\mathbf{x}_i\}$ of (2.48). We found in Section 2.2 that determination of the coordinate matrix of a vector \mathbf{x} with respect to a basis $\mathcal{X} = \{\mathbf{x}_i\}$ can be reduced to performing elimination on the matrix equation (2.17):

$$[\mathbf{x}]_{\mathcal{X}} = ([\mathbf{x}_1]_{\mathcal{X}} \cdots [\mathbf{x}_n]_{\mathcal{X}})[\mathbf{x}]_{\mathcal{X}}$$

where \mathcal{X} is a natural basis for the space \mathcal{V} of which \mathbf{x} is a member (i.e., a basis with respect to which coordinates can be determined by inspection).

Exercise 1. Show that $[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}$ of (2.48) can be obtained by the row reduction

$$\left([y_1]_{\mathcal{X}} \cdots [y_n]_{\mathcal{X}} \mid [\mathbf{T}\mathbf{x}_1]_{\mathcal{X}} \cdots [\mathbf{T}\mathbf{x}_n]_{\mathcal{X}} \right) \rightarrow \left(\mathbf{I} \mid [\mathbf{T}]_{\mathcal{X}\mathcal{Y}} \right) \quad (2.49)$$

where \mathcal{X} is a natural basis for the range of definition \mathcal{W} . (Hint: if the elements of $[\mathbf{T}\mathbf{x}_i]_{\mathcal{Y}}$ are denoted by $[\mathbf{T}\mathbf{x}_i]_{\mathcal{Y}} = (c_{i1} \cdots c_{in})^T$, then $\mathbf{T}\mathbf{x}_i = \sum_j c_{ji} \mathbf{y}_j$, and $[\mathbf{T}\mathbf{x}_i]_{\mathcal{X}} = \sum_j c_{ji} [y_j]_{\mathcal{X}}$.) Use this approach to find $[\mathbf{T}]_{\mathcal{E}_n \mathcal{E}_2}$ of Example 1.

Example 3. The Matrix of a Matrix Transformation. Let $\mathbf{T}: \mathcal{N}^{n \times 1} \rightarrow \mathcal{N}^{m \times 1}$ be defined by $\mathbf{T}\mathbf{x} \triangleq \mathbf{A}\mathbf{x}$, where \mathbf{A} is an $m \times n$ matrix. Denoting the standard bases for $\mathcal{N}^{n \times 1}$ and $\mathcal{N}^{m \times 1}$ by \mathcal{E}_n and \mathcal{E}_m , respectively, we find $[\mathbf{T}]_{\mathcal{E}_n \mathcal{E}_m} = \mathbf{A}$. Although $[\mathbf{x}]_{\mathcal{X}}$ and \mathbf{x} are identical in this example, we should distinguish between them, for it is certainly incorrect to equate the matrix $[\mathbf{T}]_{\mathcal{E}_n \mathcal{E}_m}$ to the transformation \mathbf{T} .

Suppose $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ is invertible and linear; \mathcal{V} and \mathcal{W} are finite-dimensional with bases \mathcal{X} and \mathcal{Y} , respectively. It follows from (2.45) that

$$[\mathbf{T}^{-1}]_{\mathcal{Y}\mathcal{X}}[\mathbf{y}]_{\mathcal{Y}} = [\mathbf{T}^{-1}\mathbf{y}]_{\mathcal{X}} \quad (2.50)$$

for all \mathbf{y} in \mathcal{W} . Then, for each \mathbf{x} in \mathcal{V} ,

$$[\mathbf{x}]_{\mathcal{X}} = [\mathbf{T}^{-1}\mathbf{T}\mathbf{x}]_{\mathcal{X}} = [\mathbf{T}^{-1}]_{\mathcal{Y}\mathcal{X}}[\mathbf{T}\mathbf{x}]_{\mathcal{Y}} = [\mathbf{T}^{-1}]_{\mathcal{Y}\mathcal{X}}[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}[\mathbf{x}]_{\mathcal{X}}$$

A similar relationship can be established with \mathbf{T} and \mathbf{T}^{-1} reversed. Then as a consequence of (2.29),

$$[\mathbf{T}^{-1}]_{\mathcal{Y}\mathcal{X}} = [\mathbf{T}]_{\mathcal{X}\mathcal{Y}}^{-1} \quad (2.51)$$

Exercise 2. Suppose \mathcal{V} , \mathcal{W} , and \mathcal{U} are finite-dimensional vector spaces with bases \mathcal{X} , \mathcal{Y} , and \mathcal{Z} , respectively. Show that

a. If $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ and $\mathbf{U}: \mathcal{V} \rightarrow \mathcal{W}$ are linear, then

$$[a\mathbf{T} + b\mathbf{U}]_{\mathcal{X}\mathcal{Y}} = a[\mathbf{T}]_{\mathcal{X}\mathcal{Y}} + b[\mathbf{U}]_{\mathcal{X}\mathcal{Y}} \quad (2.52)$$

b. If $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ and $\mathbf{U}: \mathcal{W} \rightarrow \mathcal{U}$ are linear, then

$$[\mathbf{UT}]_{\mathcal{X}\mathcal{Z}} = [\mathbf{U}]_{\mathcal{Y}\mathcal{Z}}[\mathbf{T}]_{\mathcal{X}\mathcal{Y}} \quad (2.53)$$

Changes in Coordinate System

In Chapter 4 we discuss coordinate systems which are particularly suitable for analysis of a given linear transformation—coordinate systems for which the matrix of the transformation is diagonal. In preparation for that discussion we now explore the effect of a change of coordinate system on a coordinate matrix $[\mathbf{x}]$ and on the matrix of a transformation $[\mathbf{T}]$.

Suppose \mathcal{X} and \mathcal{Z} are two different bases for an n -dimensional vector space \mathcal{V} . We know by (2.41) that the transformations

$$\mathbf{x} \rightarrow [\mathbf{x}]_{\mathcal{X}} \quad \text{and} \quad \mathbf{x} \rightarrow [\mathbf{x}]_{\mathcal{Z}}$$

are linear and invertible. Thus we expect $[\mathbf{x}]_{\mathcal{X}}$ and $[\mathbf{x}]_{\mathcal{Z}}$ to be related by

$$\mathbf{S}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{x}]_{\mathcal{Z}} \quad (2.54)$$

where \mathbf{S} is an $n \times n$ invertible matrix. In fact, multiplication of $[\mathbf{x}]_{\mathcal{X}}$ by any invertible matrix represents a change from the coordinate system \mathcal{X} to some new coordinate system. We sometimes denote the matrix \mathbf{S} of (2.54) by the symbol $\mathbf{S}_{\mathcal{X}\mathcal{Z}}$, thereby making explicit the fact that \mathbf{S} converts coordinates relative to \mathcal{X} into coordinates relative to \mathcal{Z} . Then $(\mathbf{S}_{\mathcal{X}\mathcal{Z}})^{-1} = \mathbf{S}_{\mathcal{Z}\mathcal{X}}$.

Determination of the specific change-of-coordinates matrix \mathbf{S} defined in (2.54) follows the same line of thought as that used to determine $[\mathbf{T}]$ in (2.48). By successively substituting into (2.54) the vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ from the basis \mathcal{X} , we isolate the columns of \mathbf{S} : the i th column of \mathbf{S} is $[\mathbf{x}_i]_{\mathcal{Z}}$. Thus the unique invertible matrix \mathbf{S} which transforms coordinate matrices relative to \mathcal{X} into coordinate matrices relative to \mathcal{Z} is

$$\mathbf{S} = \mathbf{S}_{\mathcal{X}\mathcal{Z}} = ([\mathbf{x}_1]_{\mathcal{Z}} \ : \ \cdots \ : \ [\mathbf{x}_n]_{\mathcal{Z}}) \quad (2.55)$$

where the \mathbf{x}_i are the vectors in the basis \mathcal{X} .

Since a change-of-coordinates matrix is always invertible, we determine

from (2.54) that

$$\mathbf{S}^{-1}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{x}]_{\mathcal{X}}$$

and

$$\mathbf{S}^{-1} = \mathbf{S}_{\mathcal{X}\mathcal{X}}^{-1} = \mathbf{S}_{\mathcal{X}\mathcal{X}} = ([\mathbf{z}_1]_{\mathcal{X}} \ \cdots \ [\mathbf{z}_n]_{\mathcal{X}}) \quad (2.56)$$

where the \mathbf{z}_i are the vectors in the basis \mathcal{Z} . If \mathcal{Z} is a natural basis for the space, then \mathbf{S} can be found by inspection. On the other hand, if \mathcal{X} is a natural basis, we find \mathbf{S}^{-1} by inspection. It is appropriate to use either (2.55) or (2.56) in determining \mathbf{S} . We need both \mathbf{S} and \mathbf{S}^{-1} to allow conversion back and forth between the two coordinate systems. Besides, the placing of \mathbf{S} on the left side of (2.54) was arbitrary.

Example 4. A Change-of-Coordinates Matrix. Let \mathcal{E} be the standard basis for \mathcal{R}^3 . Another basis for \mathcal{R}^3 is $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$, where $\mathbf{z}_1 = (1, 1, 1)$, $\mathbf{z}_2 = (1, 1, 0)$, and $\mathbf{z}_3 = (1, 0, 0)$. Since \mathcal{E} is a natural basis for \mathcal{R}^3 , we use (2.56) to find

$$\mathbf{S}^{-1} = ([\mathbf{z}_1]_{\mathcal{E}} \ [\mathbf{z}_2]_{\mathcal{E}} \ [\mathbf{z}_3]_{\mathcal{E}}) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad (2.57)$$

A straightforward elimination (Section 1.5) yields

$$\mathbf{S} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix} \quad (2.58)$$

We note that for an arbitrary vector $\mathbf{x} = (\xi_1, \xi_2, \xi_3)$ in \mathcal{R}^3 , $[\mathbf{x}]_{\mathcal{E}} = (\xi_1 \ \xi_2 \ \xi_3)^T$. By (2.54),

$$[\mathbf{x}]_{\mathcal{Z}} = \mathbf{S}[\mathbf{x}]_{\mathcal{E}} = (\xi_3 \ \xi_2 - \xi_3 \ \xi_1 - \xi_2)^T \quad (2.59)$$

But then,

$$\begin{aligned} \mathbf{x} &= (\xi_3)\mathbf{z}_1 + (\xi_2 - \xi_3)\mathbf{z}_2 + (\xi_1 - \xi_2)\mathbf{z}_3 \\ &= (\xi_3)(1, 1, 1) + (\xi_2 - \xi_3)(1, 1, 0) + (\xi_1 - \xi_2)(1, 0, 0) \\ &= (\xi_1, \xi_2, \xi_3) \end{aligned} \quad (2.60)$$

and the validity of the change of coordinates matrix \mathbf{S} is verified.

If neither \mathcal{X} nor \mathcal{Z} is a natural basis, the determination of \mathbf{S} can still be systematized by the introduction of an intermediate step which does involve a natural basis.

Exercise 3. Suppose we need the change-of-coordinates matrix \mathbf{S} such that $\mathbf{S}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{x}]_{\mathcal{Z}}$, where neither \mathcal{X} nor \mathcal{Z} is a natural basis for \mathcal{V} . Suppose \mathcal{U} is a natural basis. Show, by introducing an intermediate change to the coordinates $[\mathbf{x}]_{\mathcal{U}}$, that

$$\mathbf{S} = ([\mathbf{z}_1]_{\mathcal{U}} \cdots [\mathbf{z}_n]_{\mathcal{U}})^{-1} ([\mathbf{x}_1]_{\mathcal{U}} \cdots [\mathbf{x}_n]_{\mathcal{U}}) \quad (2.61)$$

Example 5. Change of Coordinates via an Intermediate Natural Basis. Two bases for \mathcal{P}^3 are $\mathcal{F} \triangleq \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ and $\mathcal{G} \triangleq \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, where

$$\mathbf{f}_1(t) = 1, \quad \mathbf{f}_2(t) = 1 + t, \quad \mathbf{f}_3(t) = 1 + t^2$$

$$\mathbf{g}_1(t) = 1 + t, \quad \mathbf{g}_2(t) = t, \quad \mathbf{g}_3(t) = t + t^2$$

To find \mathbf{S} such that $\mathbf{S}[\mathbf{f}]_{\mathcal{F}} = [\mathbf{f}]_{\mathcal{G}}$, we introduce the natural basis $\mathcal{U} \triangleq \{\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3\}$, where $\mathbf{h}_i(t) = t^{i-1}$. Then, by (2.61),

$$\begin{aligned} \mathbf{S} &= ([\mathbf{g}_1]_{\mathcal{U}} \cdots [\mathbf{g}_3]_{\mathcal{U}})^{-1} ([\mathbf{f}_1]_{\mathcal{U}} \cdots [\mathbf{f}_3]_{\mathcal{U}}) \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & -2 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

Similarity and Equivalence Transformations

Now that we have a process for changing coordinate systems, we explore the effect of such a change on the matrix of a transformation. Suppose \mathbf{T} is a linear operator on \mathcal{V} , and that \mathcal{X} and \mathcal{Z} are two different bases for \mathcal{V} . Then $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$ is defined by

$$[\mathbf{T}]_{\mathcal{X}\mathcal{X}}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{T}\mathbf{x}]_{\mathcal{X}}$$

The change from the \mathcal{X} to the \mathcal{Z} coordinate system is described by

$$\mathbf{S}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{x}]_{\mathcal{Z}}$$

The change-of-coordinates matrix \mathbf{S} also applies to the vector $\mathbf{T}\mathbf{x}$ in \mathcal{V} :

$$\mathbf{S}[\mathbf{T}\mathbf{x}]_{\mathcal{X}} = [\mathbf{T}\mathbf{x}]_{\mathcal{Z}}$$

By substituting $[\mathbf{x}]_{\mathcal{X}}$ and $[\mathbf{T}\mathbf{x}]_{\mathcal{X}}$ from these last two equations into the defining equation for $[\mathbf{T}\mathbf{x}]_{\mathcal{Z}}$, we find

$$[\mathbf{T}]_{\mathcal{X}\mathcal{X}} \mathbf{S}^{-1}[\mathbf{x}]_{\mathcal{Z}} = \mathbf{S}^{-1}[\mathbf{T}\mathbf{x}]_{\mathcal{Z}}$$

or

$$(\mathbf{S}[\mathbf{T}]_{\mathcal{X}\mathcal{X}} \mathbf{S}^{-1})[\mathbf{x}]_{\mathcal{Z}} = [\mathbf{T}\mathbf{x}]_{\mathcal{Z}}$$

But this is the defining equation for $[\mathbf{T}]_{\mathcal{Z}\mathcal{Z}}$. It is apparent that

$$[\mathbf{T}]_{\mathcal{Z}\mathcal{Z}} = \mathbf{S}[\mathbf{T}]_{\mathcal{X}\mathcal{X}} \mathbf{S}^{-1} \quad (2.62)$$

where \mathbf{S} converts from the \mathcal{X} coordinate system to the \mathcal{Z} coordinate system. Equation (2.62) describes an invertible linear transformation on $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$ known as a **similarity transformation**. In Section 4.2, we find that a similarity transformation preserves the basic spectral properties of the matrix. It is comforting to know that any two matrix representations of a linear system have the same properties—these properties are inherent in the model, \mathbf{T} , and should not be affected by the coordinate system we select.

Example 6. A Similarity Transformation. In Example 2 we found the matrix of the differential operator on \mathcal{P}^3 relative to the natural basis for \mathcal{P}^3 :

$$[\mathbf{D}]_{\mathcal{X}\mathcal{X}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

Another basis for \mathcal{P}^3 is $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, where $\mathbf{g}_1(t) = 1 + t$, $\mathbf{g}_2(t) = t$, and $\mathbf{g}_3(t) = t + t^2$. The change-of-coordinates matrix which relates the two bases \mathcal{X} and \mathcal{G} is defined by $\mathbf{S}[\mathbf{f}]_{\mathcal{X}} = [\mathbf{f}]_{\mathcal{G}}$; we find it using (2.56):

$$\begin{aligned} \mathbf{S}^{-1} &= ([\mathbf{g}_1]_{\mathcal{X}} \ : \ [\mathbf{g}_2]_{\mathcal{X}} \ : \ [\mathbf{g}_3]_{\mathcal{X}}) \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

The inverse matrix is

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

Then, by (2.62),

$$\begin{aligned} [\mathbf{D}]_{\mathcal{G}\mathcal{G}} &= \mathbf{S}[\mathbf{D}]_{\mathcal{X}\mathcal{X}}\mathbf{S}^{-1} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

Exercise 4. Let $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ be a linear transformation. Assume \mathcal{V} and \mathcal{W} are finite dimensional. Let the invertible matrix $\mathbf{S}_{\mathcal{X}\mathcal{F}}$ convert from the basis \mathcal{X} to the basis \mathcal{F} in \mathcal{V} . Let the invertible matrix $\mathbf{S}_{\mathcal{Y}\mathcal{G}}$ convert from the basis \mathcal{Y} to the basis \mathcal{G} in \mathcal{W} . Show that

$$[\mathbf{T}]_{\mathcal{F}\mathcal{G}} = \mathbf{S}_{\mathcal{Y}\mathcal{G}}[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}\mathbf{S}_{\mathcal{X}\mathcal{F}}^{-1} \quad (2.63)$$

This transformation of the matrix $[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}$ is called an **equivalence transformation**. The similarity transformation (2.42) is a special case. The term “equivalence” is motivated by the fact that $[\mathbf{T}]_{\mathcal{X}\mathcal{Y}}$ and $[\mathbf{T}]_{\mathcal{F}\mathcal{G}}$ are equivalent models of the system. The system equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ is equally well represented by the matrix equations which result from the introduction of any coordinate systems for \mathcal{V} and \mathcal{W} .

The discussion of matrices of transformations has been limited to transformations on finite-dimensional vector spaces. The primary reason for avoiding the infinite-dimensional counterparts is our inability to speak meaningfully about bases for infinite-dimensional spaces before discussing convergence of an infinite sequence of vectors (Section 5.3). However, matrices of infinite dimension are more difficult to work with (to invert, etc.) than are finite-dimensional matrices.

2.6 Problems and Comments

*2.1 Let \mathcal{S}_1 and \mathcal{S}_2 be subsets of a vector space \mathcal{V} . Let \mathcal{W}_1 and \mathcal{W}_2 be subspaces of \mathcal{V} .

- (a) The **intersection** $\mathcal{S}_1 \cap \mathcal{S}_2$ of the sets \mathcal{S}_1 and \mathcal{S}_2 is the set of vectors which belong to both \mathcal{S}_1 and \mathcal{S}_2 ; if $\mathcal{S}_1 \cap \mathcal{S}_2$ is empty or if $\mathcal{S}_1 \cap \mathcal{S}_2 = \mathbf{0}$, we say \mathcal{S}_1 and \mathcal{S}_2 are **disjoint**.
- (b) The **union** $\mathcal{S}_1 \cup \mathcal{S}_2$ of the sets \mathcal{S}_1 and \mathcal{S}_2 is the set of vectors which belong either to \mathcal{S}_1 or to \mathcal{S}_2 or to both.

- (c) The **sum** $\mathfrak{S}_1 + \mathfrak{S}_2$ of the sets \mathfrak{S}_1 and \mathfrak{S}_2 is the set of vectors of the form $\mathbf{x}_1 + \mathbf{x}_2$, where \mathbf{x}_1 is in \mathfrak{S}_1 and \mathbf{x}_2 is in \mathfrak{S}_2 .
- (d) $\mathfrak{W}_1 \cap \mathfrak{W}_2$ is a subspace.
- (e) $\mathfrak{W}_1 \cup \mathfrak{W}_2$ is usually not a subspace.
- (f) $\mathfrak{W}_1 + \mathfrak{W}_2$ is the subspace spanned by $\mathfrak{W}_1 \cup \mathfrak{W}_2$.
- (g) $\dim(\mathfrak{W}_1) + \dim(\mathfrak{W}_2) = \dim(\mathfrak{W}_1 + \mathfrak{W}_2) + \dim(\mathfrak{W}_1 \cap \mathfrak{W}_2)$.

2.2 Prove that the real 3-tuple space \mathfrak{R}^3 introduced in Equation (2.2) is a vector space.

2.3 Determine whether or not the following sets of vectors are linearly independent:

- (a) The column vectors $(2 \ 1 \ 0 \ 1)^T$, $(1 \ 2 \ -1 \ 1)^T$, and $(3 \ 0 \ 1 \ 1)^T$ in $\mathfrak{N}^{4 \times 1}$
- (b) The functions $\mathbf{f}_1(t) = 1 + 2t - t^2$, $\mathbf{f}_2(t) = 2 + 2t + t^2$, and $\mathbf{f}_3(t) = -1 + 3t + t^2$ in \mathfrak{P}^3 .
- (c) The functions $\mathbf{g}_1(t) = 1 + 2t + t^2 - t^3$, $\mathbf{g}_2(t) = 1 + t - t^2 + t^3$, and $\mathbf{g}_3(t) = 1 + 3t + 3t^2 - 3t^3$ in \mathfrak{P}^4 .

*2.4 *Modulo-2 scalars:* data transmitted by radio or telephone usually consist in strings of binary numbers (ones and zeros). A character or number to be transmitted is represented by a binary code word of length n . It is a sequence of these code words which makes up the transmitted string. We can think of the set of all possible code words of length n as vectors in a vector space. We call the space a binary linear code (see [2.8]). The scalars used in vector space manipulations can be restricted to binary numbers if ordinary addition of scalars is replaced by **modulo-2 addition**:

$$\begin{array}{ll} 0+0=0 & 0+1=1 \\ 1+0=1 & 1+1=0 \end{array}$$

The rules for multiplication of scalars need not be changed. One way to check for errors in data transmission is to let the n th element of each code word equal the sum (mod-2) of the other elements in the word. If a single error appears in the transmitted word, the n th element will fail to give the proper sum.

- (a) Let \mathfrak{V} be the set of 5×1 matrices with the mod-2 scalars as elements. Show that \mathfrak{V} is a vector space. (Assume that addition and scalar multiplication of the matrices is based on the mod-2 scalars.)
- (b) Let \mathfrak{W} be the subset of \mathfrak{V} consisting in vectors for which the fifth element equals the sum of the other four elements. Show that \mathfrak{W} is a subspace of \mathfrak{V} .
- (c) Find a basis \mathfrak{X} for \mathfrak{W} . Determine $[\mathbf{x}]_{\mathfrak{X}}$, where $\mathbf{x} = (1 \ 1 \ 0 \ 1 \ 1)^T$.

(d) The subspace \mathcal{W} is a binary linear code. A code can also be described by a “parity check” matrix \mathbf{P} for which the code is the nullspace. Find the parity check matrix for the code \mathcal{W} .

2.5 The set of all real $m \times n$ matrices, together with the usual definitions of addition and scalar multiplication of matrices, forms a vector space which we denote by $\mathcal{M}^{m \times n}$. Determine the dimension of this linear space by exhibiting a basis for the space.

*2.6 Let \mathcal{V} and \mathcal{W} be vector spaces. With the definition of linear combination of transformations given in (2.27),

(a) The set of all transformations from \mathcal{V} into \mathcal{W} forms a vector space.

(b) The set $\mathcal{L}(\mathcal{V}, \mathcal{W})$ of all linear transformations from \mathcal{V} into \mathcal{W} forms a subspace of the vector space in (a).

(c) The set of all linear transformations which take a particular subspace of \mathcal{V} into $\theta_{\mathcal{W}}$ constitutes a subspace of $\mathcal{L}(\mathcal{V}, \mathcal{W})$.

(d) If $\dim(\mathcal{V}) = n$ and $\dim(\mathcal{W}) = m$, then $\dim(\mathcal{L}(\mathcal{V}, \mathcal{W})) = mn$.

*2.7 Exploring linear combinations by row reduction. Let $\mathcal{Y} \triangleq \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ be a set of $m \times 1$ column vectors. The linear combination $\mathbf{y} = c_1\mathbf{y}_1 + \dots + c_n\mathbf{y}_n$ can be expressed as $\mathbf{y} = \mathbf{A}\mathbf{x}$ by defining $\mathbf{A} \triangleq (\mathbf{y}_1 \vdots \mathbf{y}_2 \vdots \dots \vdots \mathbf{y}_n)$ and $\mathbf{x} \triangleq (c_1 \dots c_n)^T$. Row reduction of the matrix $(\mathbf{A} \vdots \mathbf{y})$ for an unspecified vector $\mathbf{y} \triangleq (\eta_1 \dots \eta_m)^T$, or the equivalent row reduction of $(\mathbf{A} \vdots \mathbf{I})$ for an $m \times m$ matrix \mathbf{I} , determines the form of the vectors in $\text{span}(\mathcal{Y})$ and pinpoints any linear dependency in the set \mathcal{Y} . If \mathcal{Y} is linearly independent, the row reduction also determines the coordinates with respect to \mathcal{Y} of each vector \mathbf{y} in $\text{span}(\mathcal{Y})$. Let

$$\mathbf{y} = \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}, \mathbf{y}_1 = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}, \mathbf{y}_4 = \begin{pmatrix} 3 \\ 3 \\ 9 \end{pmatrix}$$

(a) Row reduce $(\mathbf{A} \vdots \mathbf{I})$.

(b) Determine the space spanned by \mathcal{Y} ; that is, determine the relationships that must exist among the elements $\{\eta_i\}$ of \mathbf{y} in order that \mathbf{y} be some linear combination of the vectors in \mathcal{Y} . Determine a basis for $\text{span}(\mathcal{Y})$.

(c) Determine which linear combinations of the vectors in \mathcal{Y} equal the specific vector \mathbf{y} given above.

(d) The form of $\text{span}(\mathcal{Y})$ can also be determined by row reduction of \mathbf{A}^T . The nonzero rows of the row-reduced matrix constitute a basis for $\text{span}(\mathcal{Y})$. Any zero rows which appear indicate the linear dependence of the set \mathcal{Y} .

- 2.8 For the following sets of vectors, determine if \mathbf{y} is in $\text{span}\{\mathbf{y}_i\}$. If so, express \mathbf{y} as a linear combination of the vectors $\{\mathbf{y}_i\}$.

$$(a) \quad \mathbf{y} = \begin{pmatrix} 9 \\ 3 \\ 7 \end{pmatrix}, \quad \mathbf{y}_1 = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{y}_3 = \begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix}$$

$$(b) \quad \mathbf{y} = \begin{pmatrix} 9 \\ 12 \\ 10 \\ 10 \end{pmatrix}, \quad \mathbf{y}_1 = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 3 \\ 4 \\ 3 \\ 5 \end{pmatrix}, \quad \mathbf{y}_3 = \begin{pmatrix} 4 \\ 5 \\ 3 \\ 6 \end{pmatrix}$$

$$(c) \quad \mathbf{y} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{pmatrix}, \quad \mathbf{y}_1 = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{y}_3 = \begin{pmatrix} -2 \\ 1 \\ -4 \\ 3 \end{pmatrix}$$

- 2.9 Find a basis for the subspace of \mathcal{P}^4 spanned by the functions $\mathbf{f}_1(t) = 1 + t + 2t^2$, $\mathbf{f}_2(t) = 2t + t^2 + t^3$, and $\mathbf{f}_3(t) = 2 + 3t^2 - t^3$. Extend the basis for the subspace to a basis for \mathcal{P}^4 by adding appropriate vectors to the basis.
- 2.10 Find the coordinate matrix of the vector $\mathbf{x} = (1, 1, 1)$ in \mathcal{R}^3 :
- (a) Relative to the basis $\mathcal{X} = \{(1, 0, 0), (1, -1, 0), (0, 1, -1)\}$.
- (b) Relative to the basis $\mathcal{Y} = \{(1, 1, -1), (1, -1, 1), (-1, 1, 1)\}$.
- 2.11 Find the coordinate matrix of the function f relative to the basis $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, where $\mathbf{f}(t) \triangleq t$, $\mathbf{g}_1(t) \triangleq 1 + t$, $\mathbf{g}_2(t) \triangleq 1 + t^2$, and $\mathbf{g}_3(t) \triangleq 1 - t^2$.
- 2.12 Find the coordinate matrix of the function \mathbf{g} relative to the basis $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$, where $\mathbf{f}_1(t) \triangleq 1 - t$, $\mathbf{f}_2(t) \triangleq 1 - t^2$, $\mathbf{f}_3(t) \triangleq 1 + t - t^2$, and $\mathbf{g}(t) \triangleq \xi_1 + \xi_2 t + \xi_3 t^2$.
- 2.13 Find the coordinates of the vector \mathbf{x} in $\mathcal{P}^{2 \times 2}$ relative to the basis $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$, where

$$\mathbf{x}_1 \triangleq \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{x}_2 \triangleq \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix},$$

$$\mathbf{x}_3 \triangleq \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{x}_4 \triangleq \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix}, \quad \text{and} \quad \mathbf{x} \triangleq \begin{pmatrix} 3 & -1 \\ -3 & 1 \end{pmatrix}$$

- 2.14 Let $\mathcal{P}^{2 \times 2}$ denote the space of polynomial functions of the form $\mathbf{f}(s, t) = a_{11} + a_{12}s + a_{21}t + a_{22}st$. Find a basis for $\mathcal{P}^{2 \times 2}$ which includes the function $\mathbf{f}_1(s, t) = 2s - t - 1$. Find the coordinate matrix of the general vector \mathbf{f} in $\mathcal{P}^{2 \times 2}$ relative to that basis.

2.15 Let \mathcal{C} be the space of continuous functions. Define the forward difference operator $\Delta_\delta: \mathcal{C} \rightarrow \mathcal{C}$ by $(\Delta_\delta \mathbf{f})(t) \triangleq [\mathbf{f}(t + \delta) - \mathbf{f}(t)]/\delta$ for all \mathbf{f} in \mathcal{C} and for all t , where $\delta > 0$ is a fixed real number. Show that Δ_δ is linear.

2.16 *Financial planning:* the financial condition of a family unit at time t can be described by $\mathbf{f}(t) = \mathbf{f}(t - \delta) + a\mathbf{f}(t - \delta) + \mathbf{g}(t)$ where $\mathbf{f}(t)$ is the family savings at time t , $\mathbf{f}(t - \delta)$ is the savings at a previous time $t - \delta$, a is the interest rate per time interval δ , and $\mathbf{g}(t)$ is the deposit at time t . (No deposits occur between $t - \delta$ and t .)

(a) Let the time interval δ be 1 month. If we consider t only at monthly intervals, the above financial model can be expressed as the difference equation, $\mathbf{f}(k) = (1 + a)\mathbf{f}(k - 1) + \mathbf{g}(k)$. Given $\mathbf{f}(0) = \$100$, $a = 0.005$ (i.e., 6% compounded monthly), and $\mathbf{g}(k) = \$10$ for $k = 1, 2, \dots$, determine the savings versus time over 1 year by computing $\mathbf{f}(1)$ from $\mathbf{f}(0)$, $\mathbf{f}(2)$ from $\mathbf{f}(1)$, etc. (This computation is known as “marching.”)

(b) The above financial model can be rewritten as

$$\frac{\mathbf{f}(t) - \mathbf{f}(t - \delta)}{\delta} = \frac{a}{\delta} \mathbf{f}(t - \delta) + \frac{\mathbf{g}(t)}{\delta}$$

The quantity $b \triangleq a/\delta$ is the interest rate per unit time; $\mathbf{u}(t) \triangleq \mathbf{g}(t)/\delta$ is the deposit rate for the interval. If we let $\delta \rightarrow 0$, the model becomes a differential equation, $\dot{\mathbf{f}}(t) = b\mathbf{f}(t) + \mathbf{u}(t)$. Let $\mathbf{f}(0) = \$100$, $b = 0.005$ per month, and $\mathbf{u}(t) = \$10$ per month for $t > 0$; find the savings versus time over 1 year by solving the differential equation. Compare the result with (a).

(c) An arbitrary nonlinear time-varying differential equation with initial conditions can be approximated by a difference equation in order to obtain an approximate solution via the simple marching technique of (a). Approximate the differential equation of (b) by using the *forward-difference approximation* $\mathbf{f}(t) \approx (1/\epsilon)(\mathbf{f}(t + \epsilon) - \mathbf{f}(t))$, $\epsilon = 1$ month, and considering t only at monthly intervals. Solve the difference equation for a 1 year period using $\mathbf{f}(0)$, b , and $\mathbf{u}(t)$ as given in (b). Compare the result with (b). How can the difference approximation be improved?

2.17 The electrostatic potential distribution within a two-dimensional charge free region satisfies Laplace's equation:

$$(\nabla^2 \mathbf{f})(s, t) \triangleq \frac{\partial^2 \mathbf{f}(s, t)}{\partial s^2} + \frac{\partial^2 \mathbf{f}(s, t)}{\partial t^2} = 0$$

For the potential distribution between two parallel plates of spacing d , the model reduces to $\mathbf{f}''(s) = 0$ with $\mathbf{f}(0)$ and $\mathbf{f}(d)$ given.

- (a) Assume the differential operator \mathbf{D}^2 acts on $\mathcal{C}^2(0, d)$, a space of twice-differentiable functions. Find the nullspace of \mathbf{D}^2 , a subspace of $\mathcal{C}^2(0, d)$. The nullspace is the solution space for the above differential equation. Express the solution space in terms of the known boundary values $\mathbf{f}(0)$ and $\mathbf{f}(d)$. What is the dimension of the nullspace of \mathbf{D}^2 ?
- (b) Define the *central-difference operator* Δ on $\mathcal{C}^2(0, d)$ by

$$(\Delta \mathbf{f})(s) \triangleq \mathbf{f}\left(s + \frac{\delta}{2}\right) - \mathbf{f}\left(s - \frac{\delta}{2}\right)$$

The derivative of \mathbf{f} can be expressed as the limit of the central-difference approximation, $\mathbf{f}'(s) \approx (\Delta \mathbf{f})(s) / \delta$. Verify that \mathbf{D}^2 , as it acts on $\mathcal{C}^2(0, d)$, can be approximated arbitrarily closely by the second-central-difference approximation, $\mathbf{D}^2 \approx \Delta^2 / \delta^2$.

- (c) Suppose the plate spacing is $d = 5$. Let $\delta = 1$, and evaluate the finite-difference approximation $\Delta^2 \mathbf{f} = \boldsymbol{\theta}$ at $s = 1, 2, 3$, and 4 to obtain four algebraic equations in the variables $\mathbf{f}(0), \mathbf{f}(1), \dots, \mathbf{f}(5)$. Formulate these algebraic equations as a 4×6 matrix equation $\mathbf{Ax} = \boldsymbol{\theta}$. Compare this matrix equation with the differential equation $\mathbf{D}^2 \mathbf{f} = \boldsymbol{\theta}$; that is, compare the spaces on which the operators act; also compare the dimensions of their solution spaces. Solve the matrix equation in terms of the boundary values $\mathbf{f}(0)$ and $\mathbf{f}(5)$. Compare the discrete solution with the continuous solution found in (a).

This problem can also be carried out for the two-dimensional case, where $\mathbf{f}(s, t)$ is given on a closed boundary. The finite-difference approach in (b) and (c) is widely used in the solution of practical problems of this type. The equations, sometimes numbering as many as 100,000, are solved by iterative computer techniques. See Forsythe and Wasow [2.4].

- 2.18 According to the trapezoidal rule for approximate integration, if we subdivide the interval $[a, b]$ into n segments of length δ , and denote $\mathbf{g}(a + j\delta)$ by \mathbf{g}_j , $j = 0, 1, \dots, n$, then for a continuous \mathbf{g} ,

$$\begin{aligned} \int_a^b \mathbf{g}(s) ds &\approx \frac{\delta}{2} (\mathbf{g}_0 + \mathbf{g}_1) + \frac{\delta}{2} (\mathbf{g}_1 + \mathbf{g}_2) + \cdots + \frac{\delta}{2} (\mathbf{g}_{n-1} + \mathbf{g}_n) \\ &= \delta \left(\frac{\mathbf{g}_0}{2} + \mathbf{g}_1 + \cdots + \mathbf{g}_{n-1} + \frac{\mathbf{g}_n}{2} \right) \end{aligned}$$

We can view the trapezoidal rule as an approximation of a function space integral operation by a matrix multiplication \mathbf{Ax} , where \mathbf{A} is $1 \times n$ and $\mathbf{x} \triangleq (g_0 \cdots g_n)^T$.

- (a) Find the matrix \mathbf{A} which expresses the trapezoidal rule for $\delta = 1$ and $n = 5$. Apply the trapezoidal rule to accurately represent the integral of the discontinuous function $\mathbf{g}(s) \triangleq 1$ for $0 < s < 2$, $\mathbf{g}(s) \triangleq 0$ for $2 < s < 5$. Hint: at the discontinuity use the midpoint value, $(g_2^- + g_2^+)/2$.
- (b) We can also approximate a *general* integral operator by a matrix multiplication. Suppose $(\mathbf{Tf})(t) \triangleq \int_a^b k(t,s)\mathbf{f}(s) ds$ for t in $[a, b]$. We can treat the function $k(t,s)\mathbf{f}(s)$ as we did $\mathbf{g}(s)$ in (a). Subdivide both the s and t intervals into n segments of length δ , and use the same subscript notation for function values as above. Then if $k(t,s)\mathbf{f}(s)$ is continuous,

$$\left(\frac{k_{j,0}f_0}{2} + k_{j,1}f_1 + \cdots + k_{j,n-1}f_{n-1} + \frac{k_{j,n}f_n}{2} \right)$$

for $j=0, 1, \dots, n$. We can approximate the integral operation by a matrix multiplication, $\mathbf{y} = \mathbf{Ax}$, where $\mathbf{x} = (f_0 \cdots f_n)^T$ and $\mathbf{y} = ((\mathbf{Tf})_0 \cdots (\mathbf{Tf})_n)^T$. Find \mathbf{A} for $\delta = 1$, $n = 5$, $a = 0$, $b = 5$, and

$$\begin{aligned} k(t,s) &= 1 \quad \text{for } 0 \leq s < t \\ &= 0 \quad \text{for } t < s \leq 5 \end{aligned}$$

Hint: use midpoint values as in (a). Note that the operator is ordinary indefinite integration.

- (c) Apply the matrix multiplication found in (b) to obtain the approximate integral of $\mathbf{f}(s) = 3s^2$. Compare the approximation to the actual integral at the points $t = 0, 1, \dots, 5$.

*2.19 Exploring the nullspace and range by row reduction: Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 & 4 \\ 2 & 1 & 3 & 0 \\ 4 & 5 & 3 & 8 \end{pmatrix}$$

Multiplication by \mathbf{A} is a linear transformation from $\mathfrak{N}^{4 \times 1}$ into $\mathfrak{N}^{3 \times 1}$. Multiplication by \mathbf{A}^T is a linear transformation from $\mathfrak{N}^{3 \times 1}$ into $\mathfrak{N}^{4 \times 1}$. In Section 5.4 we find that if \mathbf{y} is in $\text{range}(\mathbf{A})$ and \mathbf{x} is in $\text{nullspace}(\mathbf{A}^T)$, then $\mathbf{x} \cdot \mathbf{y} = 0$ where $\mathbf{x} \cdot \mathbf{y}$ is the dot product of analytic geometry. Furthermore, if \mathbf{z} is in $\text{range}(\mathbf{A}^T)$ and \mathbf{w} is in

$\text{nullspace}(\mathbf{A})$, then $\mathbf{z} \cdot \mathbf{w} = 0$. By means of these dot product equations, we can use bases for $\text{nullspace}(\mathbf{A}^T)$ and $\text{range}(\mathbf{A}^T)$ to find bases for $\text{range}(\mathbf{A})$ and $\text{nullspace}(\mathbf{A})$, and vice versa. We can also show that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$. In this problem we obtain $\text{nullspace}(\mathbf{A})$ and $\text{range}(\mathbf{A})$ directly from \mathbf{A}^T .

(a) Row reduce $(\mathbf{A} \vdots \mathbf{I})$. Use the results of the row reduction to determine bases for $\text{nullspace}(\mathbf{A})$ and $\text{range}(\mathbf{A})$.

(b) Row reduce $(\mathbf{A}^T \vdots \mathbf{I})$. Show that the nonzero rows in the left-hand block of the row-reduced matrix constitute a basis for $\text{range}(\mathbf{A})$. Show that the rows of the right-hand block which correspond to zero rows of the left-hand block of the row-reduced matrix constitute a basis for $\text{nullspace}(\mathbf{A})$.

2.20 Define $\mathbf{T}: \mathcal{P}^3 \rightarrow \mathcal{C}(0, 1)$ by $(\mathbf{T}\mathbf{f})(t) \triangleq \int_0^1 k(t, s)\mathbf{f}(s) ds$ for all \mathbf{f} in \mathcal{P}^3 , where

$$\begin{aligned} k(t, s) &= t(1-s) \text{ for } t \leq s \\ &= (1-t)s \text{ for } t \geq s \end{aligned}$$

Find a basis for $\text{range}(\mathbf{T})$. Describe $\text{nullspace}(\mathbf{T})$.

2.21 Let \mathcal{W} be the space of polynomial functions \mathbf{f} of the form $\mathbf{f}(s, t) \triangleq c_1 + c_2s + c_3t + c_4st$ for all s and t . Define $\mathbf{T}: \mathcal{W} \rightarrow \mathcal{W}$ by $(\mathbf{T}\mathbf{f})(s, t) \triangleq (\partial/\partial s)\mathbf{f}(s, t)$ for all \mathbf{f} in \mathcal{W} .

(a) Find a basis for the range of \mathbf{T} .

(b) Determine the rank and nullity of \mathbf{T} .

2.22 Define $\mathbf{T}: \mathcal{N}^{2 \times 2} \rightarrow \mathcal{N}^{2 \times 2}$ by

$$\mathbf{T} \begin{pmatrix} c_1 & c_2 \\ c_3 & c_4 \end{pmatrix} \triangleq \begin{pmatrix} c_1 - c_2 & c_1 \\ c_2 & c_4 - c_3 \end{pmatrix}$$

for all choices of the scalars c_1, c_2, c_3 , and c_4 . Find $\text{nullspace}(\mathbf{T})$ and $\text{range}(\mathbf{T})$ by exhibiting a basis for each.

2.23 *Expected value:* the throws of a single die constitute an experiment. Let \mathcal{V} be the space of random variables defined on this experiment. We can think of the probability mass function $\omega(\sigma)$ as the relative frequency with which the outcome σ occurs: $\omega(\sigma) = \frac{1}{6}$ for $\sigma = 1, 2, \dots, 6$.

(a) A random variable \mathbf{x} in \mathcal{V} associates a value $\mathbf{x}(\sigma)$ with each possible outcome of the experiment. The value which \mathbf{x} associates with an actual trial of the experiment is called a sample value of \mathbf{x} . The probability mass function $\omega_{\mathbf{x}}(\mathbf{x})$ specifies the relative frequency with which the sample value \mathbf{x} occurs during repeated trials. Find $\omega_{\mathbf{y}}(\mathbf{y})$ for the random

variable \mathbf{y} defined by $\mathbf{y}(\sigma) \stackrel{\Delta}{=} 2$ for $\sigma = 1$ or 2 and by $\mathbf{y}(\sigma) \stackrel{\Delta}{=} 0$ for $\sigma = 3, 4, 5,$ or 6 .

- (b) The *expected value* of \mathbf{x} is the average, over many trials, of the sample values of \mathbf{x} . Thus

$$\mathbf{E}(\mathbf{x}) = \sum_x x \omega_x(x) = \sum_{\sigma} \mathbf{x}(\sigma) \omega(\sigma)$$

Find $\mathbf{E}(\mathbf{y})$ for the random variable \mathbf{y} given in (a).

- (c) Show that the functional $\mathbf{E}: \mathcal{V} \rightarrow \mathcal{R}$ is linear. Pick a basis \mathcal{X} for \mathcal{V} . Let $\mathcal{E} \stackrel{\Delta}{=} \{(1)\}$ be a basis for \mathcal{R} . Find $[\mathbf{y}]_{\mathcal{X}}$ and $[\mathbf{E}]_{\mathcal{X}\mathcal{E}}$, where \mathbf{y} is the random variable in (a).
- (d) If $\mathbf{f}: \mathcal{V} \rightarrow \mathcal{V}$ then $\mathbf{f}(\mathbf{x})$ is a random variable. Express $\mathbf{E}(\mathbf{f}(\mathbf{x}))$ in terms of $\omega(\sigma)$. Find $\mathbf{E}(\mathbf{y}^2)$ for the random variable \mathbf{y} given in (a). If $\mathbf{g}: \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$, can \mathbf{E} be applied to $\mathbf{g}(\mathbf{x}, \mathbf{y})$?

2.24 *Hadamard matrices*: let $\mathbf{f}(s)$ represent the light intensity versus position in one line of a television picture. Let the $n \times 1$ column vector \mathbf{x} be a discrete approximation to \mathbf{f} . Then \mathbf{x} can be viewed as a one-dimensional photograph. Suppose the data \mathbf{x} must be transmitted for remote viewing. One way to reduce the effect of transmission errors and to reduce the amount of data transmitted is to transmit, instead, a transformed version of \mathbf{x} . A computationally simple transformation is the Hadamard transform—multiplication by a Hadamard matrix. A symmetric Hadamard matrix \mathbf{H} consists in plus and minus ones, and satisfies $\mathbf{H}^{-1} = \mathbf{H}$ (see [2.9]). Denote the transformed vector by $\mathbf{X} = \mathbf{H}\mathbf{x}$. Let $n = 8$ and

$$\mathbf{H} = \frac{1}{8} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

The Hadamard transform spreads throughout the elements of \mathbf{X} the information which is concentrated in a single element of \mathbf{x} ; it concentrates information which is spread out.

- (a) Determine the effect of \mathbf{H} on the photographs $\mathbf{x} = (11111111)^T$ and $\mathbf{x} = \mathbf{e}_i$, where \mathbf{e}_i is the i th standard basis vector for $\mathcal{N}^{8 \times 1}$.

- (b) Find the transform of the photograph $\mathbf{x} = (2 \ 2 \ 2 \ 3 \ 2 \ 2 \ 2 \ 2)^T$. Assume that an error during transmission of \mathbf{X} reduces the third element of \mathbf{X} to zero. Determine the effect of the error on the reconstructed photograph.
- (c) The inverse transform, $\mathbf{x} = \mathbf{H}\mathbf{X}$, can be interpreted as an expansion of \mathbf{x} in terms of the columns of \mathbf{H} . The columns of \mathbf{H} are analogous to sinusoidal functions; the number of zero crossings corresponds to frequency. Let \mathbf{x} be the photograph in (b). Determine the effect on the reconstructed photograph of not transmitting the highest frequency component of \mathbf{X} (i.e., the effect of making the second element of \mathbf{X} zero). Determine the effect on the reconstructed photograph of eliminating the zero frequency component (i.e., the effect of making the first component of \mathbf{X} zero).

2.25 The space $\mathcal{C}^1(0, \infty)$ consists in the continuously differentiable functions on $[0, \infty]$. Define the Cartesian product space \mathcal{V} by $\mathcal{V} \triangleq \mathcal{C}^1(0, \infty) \times \cdots \times \mathcal{C}^1(0, \infty)$. Denote the vector-valued functions in \mathcal{V} by \mathbf{x} . We can treat the values of \mathbf{x} as vectors in $\mathfrak{R}^n \times 1$; that is, $\mathbf{x}(t) = (\mathbf{f}_1(t) \cdots \mathbf{f}_n(t))^T$, where \mathbf{f}_i is in $\mathcal{C}^1(0, \infty)$. Let \mathbf{A} be a real $n \times n$ matrix. Define the linear transformation $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{W}$ by

$$(\mathbf{T}\mathbf{x})(t) = \begin{pmatrix} \dot{\mathbf{f}}_1(t) \\ \vdots \\ \dot{\mathbf{f}}_n(t) \end{pmatrix} - \mathbf{A} \begin{pmatrix} \mathbf{f}_1(t) \\ \vdots \\ \mathbf{f}_n(t) \end{pmatrix} = \dot{\mathbf{x}}(t) - \mathbf{A}\mathbf{x}(t)$$

This transformation is central to the state-space analysis of dynamic systems.

- (a) Determine an appropriate range of definition \mathcal{W} for \mathbf{T} .
- (b) Find a basis for nullspace if $n = 2$ and

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

2.26 Assume $\epsilon < \delta \ll 1$. Then the following matrix is nearly singular:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \epsilon & 0 \\ 1 & 0 & \delta \end{pmatrix}$$

Use inverse iteration to find a basis for the near nullspace of \mathbf{A} .

2.27 Define $\mathbf{T}: \mathfrak{R}^2 \rightarrow \mathfrak{R}^2$ by $\mathbf{T}(\xi_1, \xi_2) \triangleq (\xi_1 + 2\xi_2, \xi_1 - 2\xi_2)$ for all (ξ_1, ξ_2) in \mathfrak{R}^2 . Let $\mathcal{X} = \{(1, 1), (1, -1)\}$. Find $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$.

2.28 Define $\mathbf{T}: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ by

$$\mathbf{T}(\xi_1, \xi_2, \xi_3) \stackrel{\Delta}{=} (\xi_1 + \xi_2, 2\xi_3 - \xi_1)$$

- (a) Determine $[\mathbf{T}]_{\mathfrak{B}_3, \mathfrak{B}_2}$, the matrix of \mathbf{T} relative to the standard bases for \mathbb{R}^3 and \mathbb{R}^2 .
 (b) Determine $[\mathbf{T}]_{\mathfrak{X}\mathfrak{Y}}$, where $\mathfrak{X} = \{(1, 0, -1), (1, 1, 1), (1, 0, 0)\}$ and $\mathfrak{Y} = \{(1, 0), (1, 1)\}$.

2.29 Define $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ by $\mathbf{T}(\xi_1, \xi_2) \stackrel{\Delta}{=} (\xi_1 + \xi_2, \xi_1 - \xi_2, 2\xi_2)$ for all (ξ_1, ξ_2) in \mathbb{R}^2 . Let $\mathfrak{X} = \{(1, 1), (1, -1)\}$ and $\mathfrak{Y} = \{(1, 1, -1), (1, -1, 1), (-1, 1, 1)\}$. Find $[\mathbf{T}]_{\mathfrak{X}\mathfrak{Y}}$.

2.30 Let $\mathcal{P}^{2 \times 2}$ denote the space of polynomial functions of the form $\mathbf{f}(s, t) = a_{11} + a_{12}s + a_{21}t + a_{22}st$. Define $\mathbf{T}: \mathcal{P}^{2 \times 2} \rightarrow \mathcal{W}$ by

$$(\mathbf{T}\mathbf{f})(s, t) = \int_0^s \mathbf{f}(\sigma, t) d\sigma$$

where $\mathcal{W} = \text{range}(\mathbf{T})$.

(a) Find bases, \mathfrak{F} for $\mathcal{P}^{2 \times 2}$ and \mathfrak{G} for \mathcal{W} .

(b) Find $[\mathbf{T}]_{\mathfrak{F}\mathfrak{G}}$.

(c) Determine \mathbf{T}^{-1} and $[\mathbf{T}^{-1}]_{\mathfrak{G}\mathfrak{F}}$. How else might $[\mathbf{T}^{-1}]_{\mathfrak{G}\mathfrak{F}}$ be obtained?

2.31 The sets $\mathfrak{X} = \{(1, -1, 0), (1, 0, 1), (1, 1, 1)\}$ and $\mathfrak{Y} = \{(1, 1, 0), (0, 1, 1), (1, -1, 1)\}$ are bases for \mathbb{R}^3 . Find the change of coordinates matrix $\mathbf{S}_{\mathfrak{X}\mathfrak{Y}}$ which converts coordinates relative to \mathfrak{X} into coordinates relative to \mathfrak{Y} .

2.32 Let $\mathbf{g}_1(t) = 1 - t$, $\mathbf{g}_2(t) = 1 - t^2$, and $\mathbf{g}_3(t) = 1 + t - t^2$. The set $\mathfrak{G} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$ is a basis for \mathcal{P}^3 . Another basis is $\mathfrak{F} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ where $\mathbf{f}_k(t) = t^{k-1}$.

(a) Find $[\mathbf{f}]_{\mathfrak{G}}$ for the arbitrary vector $\mathbf{f}(t) = \xi_1 + \xi_2 t + \xi_3 t^2$.

(b) Find the coordinate-transformation matrix \mathbf{S} such that $[\mathbf{f}]_{\mathfrak{G}} = \mathbf{S}[\mathbf{f}]_{\mathfrak{F}}$.

2.33 Define $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ by $\mathbf{T}(\xi_1, \xi_2) \stackrel{\Delta}{=} (\xi_2 - \xi_1, \xi_1, 2\xi_1 - \xi_2)$ for all (ξ_1, ξ_2) in \mathbb{R}^2 . The sets $\mathfrak{X} = \{(1, 1), (1, -1)\}$ and $\mathfrak{Z} = \{(1, 2), (2, 1)\}$ are bases for \mathbb{R}^2 . The sets $\mathfrak{Y} = \{(1, 1, -1), (1, -1, 1), (-1, 1, 1)\}$ and $\mathfrak{K} = \{(1, 1, 1), (0, 1, 1), (0, 0, 1)\}$ are bases for \mathbb{R}^3 .

(a) Find $[\mathbf{T}]_{\mathfrak{X}\mathfrak{Y}}$.

(b) Find the coordinate transformations $\mathbf{S}_{\mathfrak{X}\mathfrak{Z}}$ and $\mathbf{S}_{\mathfrak{Y}\mathfrak{K}}$.

(c) Use the answers to (a) and (b) to compute $[\mathbf{T}]_{\mathfrak{Z}\mathfrak{K}}$ by means of an equivalence transformation.

2.34 Define $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ by $\mathbf{T}(\xi_1, \xi_2) \stackrel{\Delta}{=} (\xi_1 + 2\xi_2, \xi_1 - 2\xi_2)$ for all (ξ_1, ξ_2) in \mathbb{R}^2 . The sets $\mathfrak{X} = \{(1, 2), (2, 1)\}$ and $\mathfrak{Y} = \{(1, 1), (1, -1)\}$ are bases for \mathbb{R}^2 .

- (a) Find $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$.
- (b) Find the coordinate transformation $\mathbf{S}_{\mathcal{X}\mathcal{Y}}$.
- (c) Use the answers to (a) and (b) to compute $[\mathbf{T}]_{\mathcal{Y}\mathcal{Y}}$ by means of a similarity transformation.

2.35 Multiplication by an invertible matrix can be interpreted either as a linear transformation or as a change of coordinates. Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2\}$ be a basis for a two-dimensional space \mathcal{V} and \mathbf{x} a vector in \mathcal{V} . Then $[\mathbf{x}_1]_{\mathcal{X}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $[\mathbf{x}_2]_{\mathcal{X}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Let

$$[\mathbf{x}]_{\mathcal{X}} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$$

- (a) Alias interpretation: assume $\mathbf{A}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{x}]_{\mathcal{Y}}$, where $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2\}$ is a second basis for \mathcal{V} . Find $[\mathbf{y}_1]_{\mathcal{X}}$ and $[\mathbf{y}_2]_{\mathcal{X}}$. Sketch $[\mathbf{x}_1]_{\mathcal{X}}$, $[\mathbf{x}_2]_{\mathcal{X}}$, $[\mathbf{x}]_{\mathcal{X}}$, $[\mathbf{y}_1]_{\mathcal{X}}$, and $[\mathbf{y}_2]_{\mathcal{X}}$ as arrows in a plane. What is the relationship between $[\mathbf{x}]_{\mathcal{X}}$ and the basis $\{[\mathbf{y}_1]_{\mathcal{X}}, [\mathbf{y}_2]_{\mathcal{X}}\}$; that is, what is meant by the notation $[\mathbf{x}]_{\mathcal{Y}}$?
- (b) Alibi interpretation: assume $\mathbf{A}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{T}\mathbf{x}]_{\mathcal{X}}$. Sketch $[\mathbf{x}_1]_{\mathcal{X}}$, $[\mathbf{x}_2]_{\mathcal{X}}$, $[\mathbf{x}]_{\mathcal{X}}$, and $[\mathbf{T}\mathbf{x}]_{\mathcal{X}}$ as arrows in a plane. What is the relationship between $[\mathbf{T}\mathbf{x}]_{\mathcal{X}}$ and the basis $\{[\mathbf{x}_1]_{\mathcal{X}}, [\mathbf{x}_2]_{\mathcal{X}}\}$; that is, what is meant by the notation $[\mathbf{T}\mathbf{x}]_{\mathcal{X}}$?

2.7 References

- [2.1] Churchill, R. V., *Fourier Series and Boundary Value Problems*, McGraw-Hill, New York, 1941.
- [2.2] Cramer, Harald and M. R. Leadbetter, *Stationary and Related Stochastic Processes*, Wiley, New York, 1967.
- [2.3] Forsythe, George E., "Singularity and Near Singularity in Numerical Analysis," *Am. Math. Mon.*, **65** (1958), 229-40.
- [2.4] Forsythe, George E. and Wolfgang R. Wasow, *Finite Difference Methods for Partial Differential Equations*, Wiley, New York, 1960.
- *[2.5] Halmos, P. R., *Finite-Dimensional Vector Spaces*, Van Nostrand, Princeton, N. J., 1958.
- *[2.6] Hoffman, Kenneth and Ray Kunze, *Linear Algebra*, Prentice-Hall, Englewood Cliffs, N. J., 1961.
- [2.7] Papoulis, Athanasios, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1965.
- [2.8] Peterson, W. Wesley, *Error-Correcting Codes*, M.I.T. Press and Wiley, New York, 1961.
- [2.9] Pratt, William K., Julius Kane, and Harry C. Andrews, "Hadamard Transform Image Coding," *Proc. IEEE*, **57**, 1 (January 1969), 58-68.
- [2.10] Royden, H. L., *Real Analysis*, 2nd ed., Macmillan, New York, 1968.
- [2.11] Wylie, C. R., Jr., *Advanced Engineering Mathematics*, 3rd ed., McGraw-Hill, New York, 1966.

Linear Differential Operators

Differential equations seem to be well suited as models for systems. Thus an understanding of differential equations is at least as important as an understanding of matrix equations. In Section 1.5 we inverted matrices and solved matrix equations. In this chapter we explore the analogous inversion and solution process for linear differential equations.

Because of the presence of boundary conditions, the process of inverting a differential operator is somewhat more complex than the analogous matrix inversion. The notation ordinarily used for the study of differential equations is designed for easy handling of boundary conditions rather than for understanding of differential operators. As a consequence, the concept of the inverse of a differential operator is not widely understood among engineers. The approach we use in this chapter is one that draws a strong analogy between linear differential equations and matrix equations, thereby placing both these types of models in the same conceptual framework. The key concept is the Green's function. It plays the same role for a linear differential equation as does the inverse matrix for a matrix equation.

There are both practical and theoretical reasons for examining the process of inverting differential operators. The inverse (or integral form) of a differential equation displays explicitly the input-output relationship of the system. Furthermore, integral operators are computationally and theoretically less troublesome than differential operators; for example, differentiation emphasizes data errors, whereas integration averages them. Consequently, the theoretical justification for applying many of the computational procedures of later chapters to differential systems is based on the inverse (or integral) description of the system. Finally, the application of the optimization techniques of Chapters 6-8 to differential systems often depends upon the prior determination of the integral forms of the systems.

One of the reasons that matrix equations are widely used is that we have a practical, automatable scheme, Gaussian elimination, for inverting a matrix or solving a matrix equation. It is also possible to invert certain types of differential equations by computer automation. The greatest progress in understanding and automation has been made for linear,

constant-coefficient differential equations with initial conditions. These equations are good models for many dynamic systems (systems which evolve with time). In Section 3.4 we examine these linear constant-coefficient models in state-space form and also in the form of n th-order differential equations. The inversion concept can be extended to partial differential equations.

3.1 A Differential Operator and Its Inverse

Within the process of inverting a differential operator there is an analogue of the elimination technique for matrix inversion. However, the analogy between the matrix equation and the differential equation is clouded by the presence of the boundary conditions. As an example of a linear differential equation and its associated boundary conditions, we use

$$-\mathbf{f}'' = \mathbf{u} \quad \text{with } \mathbf{f}(0) = \alpha_1 \text{ and } \mathbf{f}(b) = \alpha_2 \quad (3.1)$$

Equation (3.1) can be viewed as a description of the relationship between the steady-state temperature distribution and the sources of heat in an insulated bar of length b . The temperature distribution \mathbf{f} varies only as a function of position t along the bar. The temperature distribution is controlled partly by \mathbf{u} , the heat generated (say, by induction heating) throughout the bar, and partly by constant temperature baths (of temperatures α_1 and α_2 , respectively) at the two ends $t = 0$ and $t = b$. Thus both the distributed input \mathbf{u} and the boundary inputs $\{\alpha_i\}$ have practical significance. The concepts of *distributed and boundary inputs* extend to other ordinary and partial differential equations.

A Discrete Approximation of the Differential System

In order to obtain a more transparent analogy to matrix equations and thereby clarify the role of the boundary conditions, we temporarily approximate the differential equation by a set of difference equations.* Let $b = 4$, substitute into (3.1) the finite-difference approximation

$$\begin{aligned} -\frac{d^2\mathbf{f}(t)}{dt^2} &\approx -\frac{[\mathbf{f}(t+1) - \mathbf{f}(t)]/1 - [\mathbf{f}(t) - \mathbf{f}(t-1)]/1}{1} \\ &= -\mathbf{f}(t-1) + 2\mathbf{f}(t) - \mathbf{f}(t+1) \end{aligned}$$

*The approximation of derivatives by finite differences is a practical numerical approach to the solution of ordinary and partial differential equations. The error owing to the finite-difference approximation can be made as small as desired by using a sufficiently fine approximation to the derivatives. (See Forsythe and Wasow [3.3].) Special techniques are usually used to solve the resulting algebraic equations. See P&C 3.3 and Varga [3.12].

and evaluate the equation at $t = 1, 2,$ and 3 :

$$\begin{aligned}
 -\mathbf{f}(0) + 2\mathbf{f}(1) - \mathbf{f}(2) &= \mathbf{u}(1) \\
 -\mathbf{f}(1) + 2\mathbf{f}(2) - \mathbf{f}(3) &= \mathbf{u}(2) \\
 -\mathbf{f}(2) + 2\mathbf{f}(3) - \mathbf{f}(4) &= \mathbf{u}(3) \\
 \mathbf{f}(0) &= \alpha_1 \\
 \mathbf{f}(4) &= \alpha_2
 \end{aligned} \tag{3.4}$$

It is obvious that this set of algebraic equations would not be invertible without the boundary conditions. We can view the boundary conditions either as an increase in the number of equations or as a decrease in the number of unknowns. The left side of (3.2), including the boundary conditions, is a matrix multiplication of the general vector $(\mathbf{f}(0) \mathbf{f}(1) \cdots \mathbf{f}(4))^T$ in the space $\mathfrak{N}^{5 \times 1}$. The corresponding right-hand side of (3.2) is $(\mathbf{u}(1) \mathbf{u}(2) \mathbf{u}(3) \alpha_1 \alpha_2)^T$; the boundary values increase the dimension of the range of definition by two. On the other hand, if we use the boundary conditions to eliminate two variables, we reduce the dimension of the domain of the matrix operator by two, $\mathbf{f}(0)$ and $\mathbf{f}(4)$ become part of the right-hand side, and the reduced matrix operates on the general vector $(\mathbf{f}(1) \mathbf{f}(2) \mathbf{f}(3))^T$ in $\mathfrak{N}^{3 \times 1}$. By either the “expanded” or the “reduced” view, the transformation with its boundary conditions is invertible. In the next section we explore the differential equation and its boundary conditions along the same lines as we have used for this discrete approximation.

The Role of the Boundary Conditions

A differential operator without boundary conditions is like a matrix with fewer rows than columns: it leads to an underdetermined differential equation. In the same manner as in the discrete approximation (3.2), appropriate boundary conditions make a linear differential operator invertible. In order that we be able to denote the inverse of (3.1) in a simple manner as we do for matrix equations, we must combine the differential operator $-\mathbf{D}^2$ and the two boundary conditions into a single operator on a vector space. We can do so using the “increased equations” view of the boundary conditions. Let \mathbf{f} be a function in the space $\mathcal{C}^2(\mathbf{0}, b)$ of twice continuously differentiable functions; then $-\mathbf{f}''$ will be in $\mathcal{C}(\mathbf{0}, b)$, the space of continuous functions. Define the differential system operator $\mathbf{T}: \mathcal{C}^2(\mathbf{0}, b) \rightarrow \mathcal{C}(\mathbf{0}, b) \times \mathfrak{R}^2$ by

$$\mathbf{T}\mathbf{f} \triangleq (-\mathbf{f}'', \mathbf{f}(0), \mathbf{f}(b)) \tag{3.3}$$

The system equations become

$$\mathbf{T}\mathbf{f} = (\mathbf{u}, \alpha_1, \alpha_2) \tag{3.4}$$

We are seeking an explicit expression of \mathbf{T}^{-1} such that $\mathbf{f} = \mathbf{T}^{-1}(\mathbf{u}, \alpha_1, \alpha_2)$. Because of the abstractness of \mathbf{T} , an operation which produces a mixture of a distributed quantity \mathbf{u} and discrete quantities $\{\alpha_i\}$, it is not clear how to proceed to determine \mathbf{T}^{-1} .

Standard techniques for solution of differential equations are more consistent with the "decreased unknowns" interpretation of the boundary conditions. Ordinarily, we solve the differential equation, $-\mathbf{f}'' = \mathbf{u}$, ignoring the boundary conditions. Then we apply the boundary conditions to eliminate the arbitrary constants in the solution. If we think of the operator $-\mathbf{D}^2$ as being restricted through the whole solution process to act only on functions which satisfy the boundary conditions, then the "arbitrary" constants in the solution to the differential equation are not arbitrary; rather, they are specific (but unknown) functions of the boundary values, $\{\alpha_i\}$. We develop this interpretation of the inversion process into an explicit expression for the inverse of the operator \mathbf{T} of (3.3).

How do we express the "restriction" of $-\mathbf{D}^2$ in terms of an operator on a vector space? The set of functions which satisfy the boundary conditions is not a subspace of $\mathcal{C}^2(0, b)$; it does not include the zero function (unless $\alpha_1 = \alpha_2 = 0$). The analogue of this set of functions in the three-dimensional arrow space is a plane which does not pass through the origin. We frame the problem in terms of vector space concepts by separating the effects of the distributed and boundary inputs. In point of fact, it is the difference in the nature of these two types of inputs that has prevented the differential equation and the boundary conditions from being expressed as a single equation.* Decompose the differential system (3.1) into two parts, one involving only the distributed input, the other only the boundary inputs:

$$-\mathbf{f}_d'' = \mathbf{u} \quad \text{with} \quad \mathbf{f}_d(0) = \mathbf{f}_d(b) = 0 \quad (3.5)$$

$$-\mathbf{f}_b'' = \mathbf{0} \quad \text{with} \quad \mathbf{f}_b(0) = \alpha_1, \quad \mathbf{f}_b(b) = \alpha_2 \quad (3.6)$$

Equations (3.5) and (3.6) possess unique solutions. By superposition, these solutions combine to yield the unique solution \mathbf{f} to (3.1); that is, $\mathbf{f} = \mathbf{f}_d + \mathbf{f}_b$. Each of these differential systems can be expressed as a single operator on a vector space. We invert the two systems separately.

We work first with (3.5). The operator $-\mathbf{D}^2$ is onto $\mathcal{C}(0, b)$; that is, we can obtain any continuous function by twice differentiating some function in $\mathcal{C}^2(0, b)$. However, $-\mathbf{D}^2$ is singular; the general vector in nullspace ($-\mathbf{D}^2$) is of the form $\mathbf{f}(t) = \mathbf{c}_1 + \mathbf{c}_2 t$. We modify the definition of the operator $-\mathbf{D}^2$ by reducing its domain. Let \mathcal{V} be the subspace of functions in $\mathcal{C}^2(0, b)$ which satisfy the homogeneous boundary conditions of (3.5),

*Friedman [3.4] does include the boundary conditions in the differential equation by treating the boundary conditions as delta functions superimposed on the distributed input.

$\mathbf{f}(0) = \mathbf{f}(b) = 0$. Define the modified differential operator $\mathbf{T}_d: \mathcal{V} \rightarrow \mathcal{C}(0, b)$ by $\mathbf{T}_d \mathbf{f} \triangleq -\mathbf{D}^2 \mathbf{f}$ for all \mathbf{f} in \mathcal{V} . The “distributed input” differential system (3.5) becomes

$$\mathbf{T}_d \mathbf{f}_d = \mathbf{u} \tag{3.7}$$

The boundary conditions are now included in the definition of the operator; in effect, we have “reduced” the operator $-\mathbf{D}^2$ to the operator \mathbf{T}_d by using the two boundary conditions to eliminate two “variables” or two degrees of freedom from the domain of the operator $-\mathbf{D}^2$. The operator \mathbf{T}_d is nonsingular; the equation $-\mathbf{f}''(t) = 0$ has no nonzero solutions in \mathcal{V} . Furthermore, \mathbf{T}_d is onto; eliminating from the domain of $-\mathbf{D}^2$ those functions which do not satisfy the zero boundary conditions of (3.5) does not eliminate any functions from the range of $-\mathbf{D}^2$. Suppose \mathbf{g} is in $\mathcal{C}^2(0, b)$, and that $\mathbf{g}(0)$ and $\mathbf{g}(b)$ are not zero. Define the related function \mathbf{f} in \mathcal{V} by $\mathbf{f}(t) \triangleq \mathbf{g}(t) - [\mathbf{g}(0) + t(\mathbf{g}(b) - \mathbf{g}(0))/b]$. We have simply subtracted a “straight line” to remove the nonzero end points from \mathbf{g} ; as a result, $\mathbf{f}(0) = \mathbf{f}(b) = 0$. But $-\mathbf{D}^2 \mathbf{f} = -\mathbf{D}^2 \mathbf{g}$. Both \mathbf{f} and \mathbf{g} lead to the same function in $\mathcal{C}(0, b)$. Every vector in $\mathcal{C}(0, b)$ comes (via $-\mathbf{D}^2$) from some function in \mathcal{V} . Thus \mathbf{T}_d is onto and invertible.

The differential system (3.6) can also be expressed as a single invertible operator. The nonzero boundary conditions of (3.6) describe a transformation $\mathbf{U}: \mathcal{C}^2(0, b) \rightarrow \mathcal{R}^2$, where

$$\mathbf{U} \mathbf{f} \triangleq (\mathbf{f}(0), \mathbf{f}(b))$$

Since $\mathcal{C}^2(0, b)$ is infinite dimensional but \mathcal{R}^2 is not, \mathbf{U} must be singular. We modify the operator \mathbf{U} by reducing its domain. Let \mathcal{W} be the subspace of functions in $\mathcal{C}^2(0, b)$ which satisfy the homogeneous differential equation of (3.6), $-\mathbf{f}_b''(t) = 0$; \mathcal{W} is the two-dimensional space \mathcal{P}^2 consisting in functions of the form $\mathbf{f}(t) = \mathbf{c}_1 + \mathbf{c}_2 t$. We define the modified operator $\mathbf{T}_b: \mathcal{P}^2 \rightarrow \mathcal{R}^2$ by $\mathbf{T}_b \mathbf{f} \triangleq (\mathbf{f}(0), \mathbf{f}(b))$ for all \mathbf{f} in \mathcal{P}^2 . The “boundary input” differential system (3.6) can be expressed as the two-dimensional equation

$$\mathbf{T}_b \mathbf{f}_b = (\alpha_1, \alpha_2) \tag{3.8}$$

The differential equation and boundary conditions of (3.6) have been combined into the single operator, \mathbf{T}_b . It is apparent that \mathbf{T}_b is invertible—the operator equation is easily solved for its unique solution.

The Inverse Operator

We have rephrased (3.5) and (3.6) in terms of the invertible operators \mathbf{T}_d and \mathbf{T}_b , respectively. Because (3.5) and (3.6) constitute a restructuring of

(3.1), we can express the solution to (3.1) as

$$\begin{aligned}\mathbf{f} &= \mathbf{f}_a + \mathbf{f}_b \\ &= \mathbf{T}_a^{-1} \mathbf{u} + \mathbf{T}_b^{-1}(\alpha_1, \alpha_2) \\ &= \mathbf{T}^{-1}(\mathbf{u}, \alpha_1, \alpha_2)\end{aligned}\tag{3.9}$$

where \mathbf{T} is the operator of (3.3).

Since \mathbf{T}_a is a differential operator, we expect $\mathbf{T}_a^{-1} : \mathcal{C}(0, b) \rightarrow \mathcal{V}$ to be an integral operator. We express it explicitly in the general form (2.34):

$$\mathbf{f}_a(t) = (\mathbf{T}_a^{-1} \mathbf{u})(t) = \int_0^b k(t, s) \mathbf{u}(s) ds \tag{3.10}$$

The kernel function k is commonly referred to as the **Green's function** for the differential system (3.1). In order that (3.10) correctly express the inverse of \mathbf{T}_a , $\mathbf{f}_a(t)$ must satisfy the differential system (3.5) from which \mathbf{T}_a is derived. Substituting (3.10) into (3.5) yields

$$\begin{aligned}-\mathbf{f}_a''(t) &= -\frac{d^2}{dt^2} \int_0^b k(t, s) \mathbf{u}(s) ds \\ &= \int_0^b -\frac{d^2 k(t, s)}{dt^2} \mathbf{u}(s) ds = \mathbf{u}(t)\end{aligned}$$

with

$$\begin{aligned}\mathbf{f}_a(0) &= \int_0^b k(0, s) \mathbf{u}(s) ds = 0 \\ \mathbf{f}_a(b) &= \int_0^b k(b, s) \mathbf{u}(s) ds = 0\end{aligned}$$

for all \mathbf{u} in $\mathcal{C}(0, b)$. These equations are satisfied for all continuous \mathbf{u} if and only if

$$-\frac{d^2 k(t, s)}{dt^2} = \delta(t - s) \quad \text{with} \quad k(0, s) = k(b, s) = 0 \tag{3.11}$$

That is, the Green's function k , as a function of its first variable t , must satisfy the differential equation and boundary conditions (3.5) for $\mathbf{u}(t) = \delta(t - s)$, where $\delta(t - s)$ is a unit impulse (or Dirac delta function) applied at the point $t = s$.* We can use (3.11) to determine the Green's function.

*See Appendix 2 for a discussion of delta functions. We use some license in interchanging the order of differentiation and integration when delta functions are present. The interchange can be justified, however, through the theory of distributions (Schwartz [3.10]).

For practical purposes we can think of $\delta(t - s)$ as a narrow continuous pulse of unit area, centered at $t = s$. [In terms of the steady-state heat-flow problem (3.1), the function $\delta(t - s)$ in (3.11) represents the generation of a unit quantity of heat per unit time in the cross section of the bar at $t = s$.] However, $\delta(t - s)$ is not a function in the usual sense; its value is not defined at $t = s$. It is not in $\mathcal{C}^2(0, b)$. Therefore, the solution k to (3.11) cannot be in $\mathcal{C}^2(0, b)$. We simply note that the domain $\mathcal{C}^2(0, b)$ and range of definition $\mathcal{C}(0, b)$ of the operator $-\mathbf{D}^2$ were defined somewhat arbitrarily. We can allow a "few" discontinuities or delta functions in $-\mathbf{D}^2\mathbf{f}$ if we also add to $\mathcal{C}^2(0, b)$ those functions whose second derivatives contain a "few" discontinuities or delta functions.

The operator $\mathbf{T}_b^{-1}: \mathfrak{R}^2 \rightarrow \mathfrak{P}^2$ can also be expressed explicitly. Since \mathbf{T}_b^{-1} acts linearly on the vector (α_1, α_2) in \mathfrak{R}^2 to yield a polynomial in \mathfrak{P}^2 , we express \mathbf{T}_b^{-1} as

$$\mathbf{f}_b = \mathbf{T}_b^{-1}(\alpha_1, \alpha_2) = \alpha_1 \rho_1 + \alpha_2 \rho_2 \quad (3.12)$$

where ρ_1 and ρ_2 are functions in \mathfrak{P}^2 . We refer to the function $\rho_j(t)$ as the **boundary kernel** for the differential system (3.1). Just as the Green's function is a function of two variables, t and s , so the boundary kernel is a function of both the continuous variable t and the discrete variable j . Because of the simplicity of the differential operator of this example, the introduction of the boundary kernel seems unnecessary and artificial. For more complicated differential operators, however, the boundary kernel provides a straightforward approach to determination of the full inverse operator. In order that (3.12) correctly describe \mathbf{T}_b^{-1} , \mathbf{f}_b must satisfy the differential system (3.6):

$$\begin{aligned} -\mathbf{f}_b'' &= -\alpha_1 \rho_1'' - \alpha_2 \rho_2'' = 0 \\ \mathbf{f}_b(0) &= \alpha_1 \rho_1(0) + \alpha_2 \rho_2(0) = \alpha_1 \\ \mathbf{f}_b(b) &= \alpha_1 \rho_1(b) + \alpha_2 \rho_2(b) = \alpha_2 \end{aligned}$$

for all α_1 and α_2 . Thus the boundary kernel ρ must obey

$$\begin{aligned} -\rho_1''(t) &= 0 & \text{with } \rho_1(0) &= 1, \rho_1(b) = 0 \\ -\rho_2''(t) &= 0 & \text{with } \rho_2(0) &= 0, \rho_2(b) = 1 \end{aligned} \quad (3.13)$$

We can use (3.13) to determine the boundary kernel.

We have defined carefully the differential system operator \mathbf{T} , the "distributed input" system operator \mathbf{T}_d , and the "boundary input" system operator \mathbf{T}_b in order to be precise about the vector space concepts involved with inversion of differential equations. However, to continue use of this

precise notation would require an awkward transition back and forth between the vector space notation and the notation standard to the field of differential equations. We rely primarily on the standard notation. We use the term **differential system** to refer to the differential operator with its boundary conditions (denoted $\{-\mathbf{D}^2, \mathbf{f}(0), \mathbf{f}(b)\}$ in this example) and also to the differential equation with its boundary conditions [denoted as in (3.1)]. We refer to both the inverse of the operator and the inverse of the equation as the *inverse of the differential system*. Where we refer to the purely differential part of the system separately, we usually denote it explicitly, for example, as $-\mathbf{D}^2$ or as $-\mathbf{f}'' = u$.

A Green's Function and Boundary Kernel

We solve for the Green's function k of the system (3.1) by direct integration of (3.11). The successive integration steps are depicted graphically in Figure 3.1. It is clear from the figure that the integral of $-d^2k/dt^2$ is constant for $t < s$ and $t > s$, and contains a jump of size 1 at $t = s$. We permit the value of the constant c to depend upon the point s at which the unit impulse is applied.

$$\begin{aligned} -\frac{dk(t,s)}{dt} &= c(s), & t < s \\ &= c(s) + 1, & t > s \end{aligned}$$

Integration of $-dk/dt$ yields continuity of $-k$ at s :

$$\begin{aligned} -k(t,s) &= c(s)t + d(s), & t \leq s \\ &= c(s)s + d(s) + (c(s) + 1)(t - s), & t \geq s \end{aligned}$$

Applying the boundary conditions we find

$$\begin{aligned} -k(0,s) &= c(s)(0) + d(s) = 0 & \Rightarrow d(s) = 0 \\ -k(b,s) &= c(s)s + (c(s) + 1)(b - s) = 0 & \Rightarrow c(s) = \frac{s - b}{b} \end{aligned}$$

Thus

$$\begin{aligned} k(t,s) &= \frac{(b-s)t}{b}, & t \leq s \\ &= \frac{(b-t)s}{b}, & t \geq s \end{aligned} \tag{3.14}$$

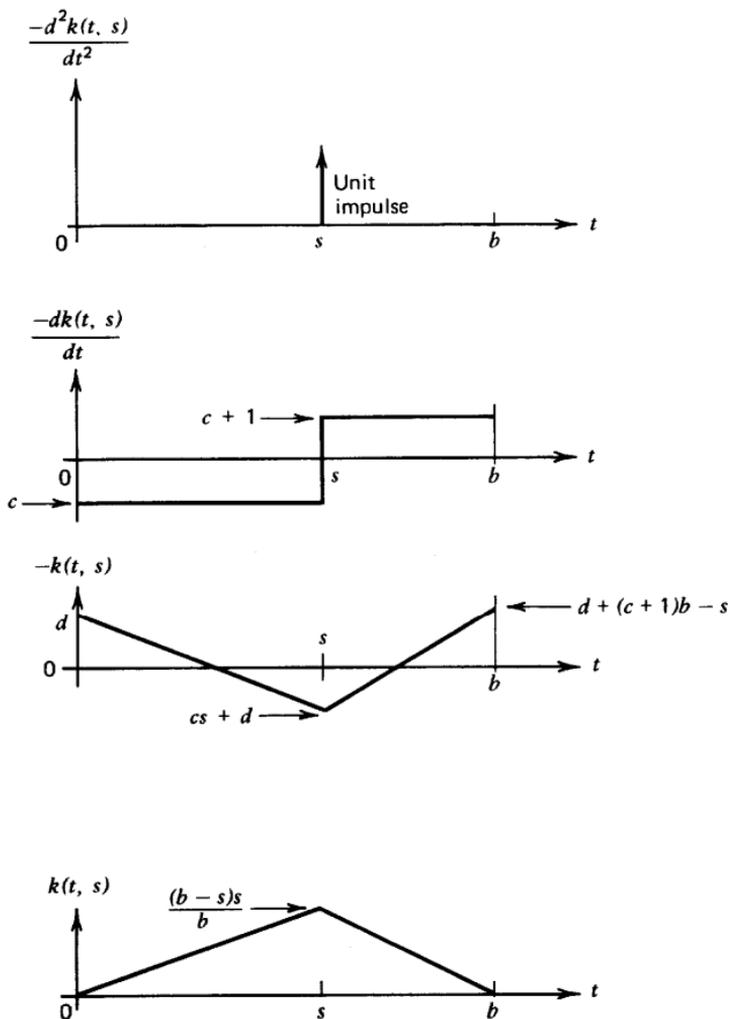


Figure 3.1. Graphical integration of (3.11).

where both t and s lie in the interval $[0, b]$.

By integration of (3.13) we determine the boundary kernel ρ associated with (3.1). The general solution to the j th differential equation is $\rho_j(t) = c_{j1} + c_{j2}t$. Using the boundary conditions we find

$$\begin{aligned} \rho_1(t) &= \frac{b-t}{b} \\ \rho_2(t) &= \frac{t}{b} \end{aligned} \tag{3.15}$$

Having found k and ρ , we insert them into (3.10) and (3.12) to obtain

\mathbf{T}_a^{-1} and \mathbf{T}_b^{-1} . Combining the two inverses as in (3.9) produces

$$\begin{aligned} \mathbf{f}(t) &= \int_0^b k(t,s)\mathbf{u}(s)ds + \alpha_1\rho_1(t) + \alpha_2\rho_2(t) \\ &= \int_0^t \frac{(b-t)s}{b}\mathbf{u}(s)ds + \int_t^b \frac{(b-s)t}{b}\mathbf{u}(s)ds + \alpha_1\frac{b-t}{b} + \alpha_2\frac{t}{b} \quad (3.16) \end{aligned}$$

Equation (3.16) is an explicit description of the inverse of the linear differential system (3.1).

A Matrix Analogy

A differential equation with an appropriate set of boundary conditions is analogous to a square matrix equation. We explore this analogy in order to remove some of the abstractness and mystery from differential operators and their inverses. An example of a matrix equation and its corresponding inverse is

$$\begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

Any such pair of equations can be expressed as $\mathbf{Ax} = \mathbf{y}$ and $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, respectively, for some square matrix \mathbf{A} . The inverse matrix equation is more clearly analogous to an inverse differential equation (or integral equation) if we express the matrix multiplication in the form of a summation. Denote the elements of \mathbf{x} and \mathbf{y} by ξ_i and η_j , respectively. Then the equation $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ becomes*

$$\xi_i = \sum_{j=1}^n (\mathbf{A}^{-1})_{ij}\eta_j \quad i=1, \dots, n \quad (3.17)$$

The symbol $(\mathbf{A}^{-1})_{ij}$ represents the element in row i and column j of the $n \times n$ matrix \mathbf{A}^{-1} . Thus the inverse matrix, a function of the two integer variables i and j , is the kernel of a summation operator. In the form (3.17), the inverse matrix equation $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ is obviously a discrete analogue of the integral equation $\mathbf{f}_a(t) = \int_0^b k(t,s)\mathbf{u}(s)ds$ of (3.10). The Green's function $k(t,s)$ is the analogue of the inverse matrix \mathbf{A}^{-1} . If we compare the inverse matrix equation (3.17) to (3.16), the full inverse of the differential system (3.1), the analogy is clouded somewhat by the presence of the boundary terms. The true analogue of \mathbf{A}^{-1} is the pair of kernel functions, k and ρ .

*See (2.35).

Because $k(t,s)$ and $\rho_j(t)$ appear as "weights" in an integral or summation, the inverse form of the differential system is somewhat more useful to the intuition than is the differential system itself.

We can also draw an analogy between the process of inverting the matrix \mathbf{A} and the process of solving for k and ρ . The solution to the equation $\mathbf{Ax} = \mathbf{e}_i$, where \mathbf{e}_i is the i th standard basis vector for $\mathcal{N}^{n \times 1}$, is the i th column of \mathbf{A}^{-1} . The solution process is analogous to solving (3.11) for $k(t,s)$ with s fixed; it is also analogous to solving (3.13) for ρ_j with j fixed. The row reduction $(\mathbf{A} : \mathbf{I}) \rightarrow (\mathbf{I} : \mathbf{A}^{-1})$ produces all columns of the inverse matrix simultaneously. Thus the inversion of \mathbf{A} by row reduction is analogous to the determination of k and ρ by solving (3.11) and (3.13), respectively. In general, the process of computing k and ρ requires more effort than does the direct solution of (3.1) for specific inputs \mathbf{u} and $\{\alpha_i\}$. However, the resulting inverse equation (3.16) contains information about the solution for any set of inputs.

3.2 Properties of n th-Order Systems and Green's Functions

In Section 3.1 we introduced the concepts of a differential operator and its inverse by means of a simple second-order example, (3.1). We now explore these concepts in detail for more general linear differential systems. Included in this section is an examination of noninvertible differential systems and a development of conditions for invertibility. Techniques for explicit determination of the Green's function and boundary kernel are treated in Section 3.3.

We define a **regular** n th-order linear differential operator $\mathbf{L} : \mathcal{C}^n(a,b) \rightarrow \mathcal{C}(a,b)$ by

$$(\mathbf{L}\mathbf{f})(t) \triangleq g_0(t)\mathbf{f}^{(n)}(t) + g_1(t)\mathbf{f}^{(n-1)}(t) + \dots + g_n(t)\mathbf{f}(t) \quad (3.18)$$

where the coefficients $\{g_i\}$ are continuous and $g_0(t) \neq 0$ on $[a,b]$.* The corresponding n th-order differential equation is $\mathbf{L}\mathbf{f} = \mathbf{u}$, where the distributed input function \mathbf{u} is continuous on $[a,b]$. It is well known that \mathbf{L} is onto $\mathcal{C}(a,b)$; the n th-order differential equation without boundary conditions always has solutions (Ince [3.6]). The **homogeneous differential equation** is defined as the equation $\mathbf{L}\mathbf{f} = 0$, without boundary conditions (the input \mathbf{u} is zero). The homogeneous differential equation for the operator

*If the interval $[a,b]$ were infinite, if g_0 were zero at some point, or if one of the coefficient functions were discontinuous, we would refer to (3.18) as a *singular* differential operator. In Section 5.5 we refer to the regular second-order linear differential operator as a regular Sturm-Liouville operator.

(3.18) always has n linearly independent solutions[†]; we call a set $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ of independent solutions a **fundamental set of solutions** for L . We sometimes express such a set as the **complementary function** for L :

$$\mathbf{f}_c \stackrel{\Delta}{=} c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n \quad (3.19)$$

where c_1, \dots, c_n are unspecified constants. Both the complementary function \mathbf{f}_c and the fundamental set of solutions $\{\mathbf{v}_i\}$ are, in reality, descriptions of the n -dimensional nullspace of L .

In order that L of (3.18) be invertible, we must add n appropriate boundary conditions to eliminate the n arbitrary constants in the complementary function. We denote the i th boundary condition for (3.18) by $\beta_i(\mathbf{f}) = \alpha_i$, where α_i is a scalar and β_i is a linear functional on $\mathcal{C}^n(a, b)$.[‡] A typical boundary condition is some linear combination of \mathbf{f} and its first $n - 1$ derivatives evaluated at the end points of the interval of definition. For example,

$$\beta_1(\mathbf{f}) \stackrel{\Delta}{=} \gamma_1 \mathbf{f}(a) + \gamma_2 \mathbf{f}'(a) + \gamma_3 \mathbf{f}(b) + \gamma_4 \mathbf{f}'(b) = \alpha_1 \quad (3.20)$$

(where the $\{\gamma_i\}$ are scalars) is as general a boundary condition as we would normally expect to encounter for a second-order differential operator acting on functions defined over $[a, b]$. The second boundary condition for the second-order differential equation, $\beta_2(\mathbf{f}) = \alpha_2$, would be of the same form, although the particular linear combination of derivatives which constitutes β_2 would have to be linearly independent of that specified by the coefficients $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ in β_1 . There is, of course, no reason why the boundary conditions could not involve evaluations of \mathbf{f} and its derivatives at interior points of the interval of definition. We refer to the boundary condition $\beta_i(\mathbf{f}) = 0$, where the boundary input α_i is zero, as a **homogeneous boundary condition**.

Consider the following n th-order differential system:

$$\begin{aligned} L\mathbf{f} &= \mathbf{u} \\ \beta_i(\mathbf{f}) &= \alpha_i, \quad i=1, \dots, m \end{aligned} \quad (3.21)$$

where L is defined in (3.18) and β_i is an n th-order version of (3.20); m is typically but not necessarily equal to n . We call a solution \mathbf{f}_p to (3.21) a

[†]See P&C 3.4.

[‡]Of course, it is possible for the boundary conditions associated with a physical system to be nonlinear functions of \mathbf{f} . We consider here only linear differential equations and linear boundary conditions.

particular solution for the differential system. A **completely homogeneous solution** \mathbf{f}_h for the differential system is a solution to the homogeneous differential equation with homogeneous boundary conditions (the homogeneous differential system):

$$\begin{aligned} \mathbf{L}\mathbf{f} &= \mathbf{0} \\ \beta_i(\mathbf{f}) &= 0, \quad i = 1, \dots, m \end{aligned} \quad (3.22)$$

Thus a completely homogeneous solution for the differential system is a solution with all inputs zero. Any solution \mathbf{f} to (3.21) can be written as $\mathbf{f} = \mathbf{f}_p + \mathbf{f}_h$, where \mathbf{f}_p is any particular solution and \mathbf{f}_h is some homogeneous solution. The set of completely homogeneous solutions constitutes the nullspace of the differential system (or the nullspace of the underlying differential operator).^{*} A system with a nonzero nullspace is not invertible.

Exercise 1. Suppose

$$(\mathbf{L}\mathbf{f})(t) \triangleq \mathbf{f}''(t) = \mathbf{u}(t) \quad (3.23)$$

with the boundary conditions

$$\beta_1(\mathbf{f}) \triangleq \mathbf{f}'(0) = \alpha_1 \quad \beta_2(\mathbf{f}) \triangleq \mathbf{f}'(1) = \alpha_2 \quad (3.24)$$

What is the completely homogeneous solution to (3.23)-(3.24)? Show that the general solution to (3.23)-(3.24) is

$$\mathbf{f}(t) = \int_0^t \int_0^\sigma \mathbf{u}(\tau) d\tau d\sigma + \alpha_1 t + \mathbf{f}(0) \quad (3.25)$$

where

$$\int_0^1 \mathbf{u}(\tau) d\tau = \alpha_1 - \alpha_2 \quad (3.26)$$

Note that the differential system (3.23)-(3.24) is not invertible. No solution exists unless the inputs \mathbf{u} and $\{\alpha_i\}$ satisfy (3.26).

The Role of the Homogeneous Differential System

The matrix analogue of the n th-order differential system (3.21) is the matrix equation $\mathbf{A}\mathbf{x} = \mathbf{y}$ (where \mathbf{A} is not necessarily square). Row reduction of \mathbf{A} determines the nullspace of \mathbf{A} (the solution to $\mathbf{A}\mathbf{x} = \mathbf{0}$); it also shows

^{*}See (3.4) and (3.9).

the dependencies in the rows of \mathbf{A} and the degree of degeneracy of the equation—the degree to which the range of the matrix transformation fails to fill the range of definition. To actually find the range of the matrix transformation (specific conditions on \mathbf{y} for which the equation is solvable), we can follow either of two approaches: (a) row reduce \mathbf{A}^T (the rows of \mathbf{A}^T span the range of \mathbf{A})*; or (b) row reduce $(\mathbf{A} \vdots \mathbf{I})$. If \mathbf{A} is square and invertible, approach (b) amounts to inversion of \mathbf{A} .

For the differential system (3.21), the analogue of row reduction of \mathbf{A} is the analysis of the completely homogeneous system (3.22). We focus first on this analysis, thereby determining the extent to which (3.21) is underdetermined or overdetermined. Then assuming the system (3.21) is invertible, we perform the analogue of row reduction of $(\mathbf{A} \vdots \mathbf{I})$ —inversion of the differential operator.

The solutions to the homogeneous differential equation, $\mathbf{L}\mathbf{f} = \mathbf{0}$, are expressed as the complementary function \mathbf{f}_c of (3.19). We apply the m homogeneous boundary conditions to \mathbf{f}_c , thereby eliminating some of the arbitrary constants in \mathbf{f}_c :

$$\begin{aligned} \beta_1(\mathbf{f}_c) &= c_1 \beta_1(\mathbf{v}_1) + \cdots + c_n \beta_1(\mathbf{v}_n) = 0 \\ &\vdots \\ \beta_m(\mathbf{f}_c) &= c_1 \beta_m(\mathbf{v}_1) + \cdots + c_n \beta_m(\mathbf{v}_n) = 0 \end{aligned}$$

or

$$\mathbf{B} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} \triangleq \begin{pmatrix} \beta_1(\mathbf{v}_1) \cdots \beta_1(\mathbf{v}_n) \\ \vdots \\ \beta_m(\mathbf{v}_1) \cdots \beta_m(\mathbf{v}_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad (3.27)$$

The nullspace of the differential system (3.21) consists in the functions $\mathbf{f}_c = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n$, where some of the arbitrary constants $\{c_i\}$ are eliminated by (3.27).

The key to the differential system (3.21) lies in \mathbf{B} , a **boundary condition matrix** (or **compatibility matrix**) for the system. In point of fact, \mathbf{B} completely characterizes (3.22). It describes not just the boundary conditions, but rather the effect of the boundary conditions on a set of fundamental solutions for \mathbf{L} . In general, the m boundary conditions, in concert with the

*see P&C 2.19. In Section 5.4 we introduce the adjoint operator, the analogue of \mathbf{A}^T . The orthogonal decomposition theorem (5.67) is the basis of a method for determining the range of an operator from the nullspace of its adjoint; this method is the analogue row reduction of \mathbf{A}^T .

n -order differential equation $\mathbf{L}\mathbf{f} = \mathbf{u}$, can specify either an underdetermined or overdetermined set of equations. Exercise 1 exhibits symptoms of both the underdetermined and overdetermined cases. Of course, \mathbf{B} is not unique since it can be based on any fundamental set of solutions. Yet the rank of \mathbf{B} is unique; $\text{rank}(\mathbf{B})$ tells much about the solutions to (3.21)*:

1. If $\text{rank}(\mathbf{B}) = m = n$, then (3.27) precisely eliminates the completely homogeneous solution, and (3.21) is then the analogue of an invertible square matrix equation; the system is invertible.

2. If $\text{rank}(\mathbf{B}) = p < n$, then $(n - p)$ of the constants $\{c_i\}$ in the characteristic function remain arbitrary and the nullspace of the system has dimension $(n - p)$. There are $(n - p)$ degrees of freedom in the solutions to (3.21); the system is singular.

3. If $\text{rank}(\mathbf{B}) = p < m$, then $(m - p)$ rows of \mathbf{B} are dependent on the rest. As demonstrated by Exercise 1, these dependencies in the rows of \mathbf{B} must be matched by $(m - p)$ scalar-valued relations among the boundary values $\{\alpha_i\}$ and the distributed input \mathbf{u} , or there can be no solutions to (3.21) (P&C 3.5). The system is not onto $\mathcal{C}(a, b)$.

The following example demonstrates the relationship between $\text{rank}(\mathbf{B})$ and the properties of the differential system.

Example 1. The Rank of the Boundary Condition Matrix. Let

$$(\mathbf{L}\mathbf{f})(t) \triangleq \mathbf{f}''(t) = \mathbf{u}(t)$$

for t in $[0, 1]$. The set $\{\mathbf{v}_1, \mathbf{v}_2\}$, where $\mathbf{v}_1(t) = 1$ and $\mathbf{v}_2(t) = t$, is a fundamental set of solutions for t . We apply several different sets of boundary conditions, demonstrating the three cases mentioned above.

1. $\beta_1(\mathbf{f}) \triangleq \mathbf{f}(0) = \alpha_1$, $\beta_2(\mathbf{f}) \triangleq \mathbf{f}(1) = \alpha_2$. In this case,

$$\mathbf{B} = \begin{pmatrix} \mathbf{v}_1(0) & \mathbf{v}_2(0) \\ \mathbf{v}_1(1) & \mathbf{v}_2(1) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Since $\text{rank}(\mathbf{B}) = 2 = m = n$, the system is invertible. We find the unique solution by direct integration:

$$\mathbf{f}(t) = \int_0^t \int_0^s \mathbf{u}(\tau) d\tau ds + \left[\alpha_2 - \alpha_1 - \int_0^1 \int_0^s \mathbf{u}(\tau) d\tau ds \right] t + \alpha_1$$

2. $\beta_1(\mathbf{f}) \triangleq \mathbf{f}(0) = \alpha_1$. For this single boundary condition,

$$\mathbf{B} = (\mathbf{v}_1(0) \quad \mathbf{v}_2(0)) = (1 \quad 0)$$

*Ince [3.6].

and $\text{rank}(\mathbf{B}) = 1$. Since $n = 2$, we should expect one degree of freedom in the solution. Since $m = \text{rank}(\mathbf{B})$, we should expect a solution to exist for all scalars α_1 and all continuous functions \mathbf{u} . By direct integration, the solution is

$$\mathbf{f}(t) = \int_0^t \int_0^s \mathbf{u}(\tau) d\tau ds + d_1 t + \alpha_1$$

where d_1 is an arbitrary constant.

3. $\beta_1(\mathbf{f}) \triangleq \mathbf{f}(0) = \alpha_1$, $\beta_2(\mathbf{f}) \triangleq \mathbf{f}(1) = \alpha_2$, $\beta_3(\mathbf{f}) \triangleq \mathbf{f}'(0) = \alpha_3$. Then,

$$\mathbf{B} = \begin{pmatrix} \mathbf{v}_1(0) & \mathbf{v}_2(0) \\ \mathbf{v}_1(1) & \mathbf{v}_2(1) \\ \mathbf{v}'_1(0) & \mathbf{v}'_2(0) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Because $\text{rank}(\mathbf{B}) = 2$ and $m = 3$, one scalar-valued function of \mathbf{u} , α_1 , α_2 , and α_3 must be satisfied in order that a solution exist. Since $\text{rank}(\mathbf{B}) = n$, if a solution exists for a given set of inputs $(\mathbf{u}, \alpha_1, \alpha_2)$, that solution is unique. We find the solution by direct integration and application of the three boundary conditions:

$$\mathbf{f}(t) = \int_0^t \int_0^s \mathbf{u}(\tau) d\tau ds + \alpha_3 t + \alpha_1$$

where \mathbf{u} , α_1 , α_2 , and α_3 must satisfy

$$\alpha_2 - \alpha_1 - \alpha_3 - \int_0^1 \int_0^s \mathbf{u}(\tau) d\tau ds = 0$$

4. $\beta_1(\mathbf{f}) \triangleq \mathbf{f}'(0) = \alpha_1$, $\beta_2(\mathbf{f}) \triangleq \mathbf{f}'(1) = \alpha_2$. This case is presented in Exercise 1.

$$\mathbf{B} = \begin{pmatrix} \mathbf{v}'_1(0) & \mathbf{v}'_2(0) \\ \mathbf{v}'_1(1) & \mathbf{v}'_2(1) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$\text{Rank}(\mathbf{B}) = 1$, but $m = n = 2$. We expect one scalar-valued condition on the inputs, and one degree of freedom in the solutions. The general solution and the restriction on the inputs are given in (3.25) and (3.26), respectively.

It is apparent from Example 1 that if $m < n$, the system is underdetermined; there are at least $(n - m)$ degrees of freedom in the solutions. On the other hand, if $m > n$, the system is usually overdetermined; since $\text{rank}(\mathbf{B}) < n$ for an n th-order differential system, the input data must satisfy at least $(m - n)$ different scalar-valued restrictions in order that the differential equation and boundary conditions be solvable.

Ordinarily, $m = n$; that is, the differential equation which represents a physical system usually has associated with it n independent boundary

conditions $\{\beta_i\}$. These n boundary conditions are independent in the sense that they represent independent linear combinations of \mathbf{f} , $\mathbf{f}^{(1)}$, . . . , $\mathbf{f}^{(n-1)}$ evaluated at one or more points of $[a, b]$. However, we see from the fourth case of Example 1 that a boundary condition matrix \mathbf{B} can be degenerate even if the boundary conditions are independent. Thus for a "square" differential operator, the condition for invertibility (or compatibility) is

$$\det(\mathbf{B}) \neq 0 \quad (3.28)$$

where \mathbf{B} is a boundary condition matrix as defined in (3.27).

It can be shown that (3.28) is satisfied for any differential operator for which $m = n$ and for which the boundary conditions are linearly independent and are all at one point (P&C 3.4). Only for multipoint boundary value problems can the test (3.28) fail. Exercise 1 is such a case.

For the rest of this chapter we assume (3.28) is satisfied, and proceed to determine the inverse of the differential system (3.21). In Section 4.3, where we determine eigenvalues and eigenfunctions of differential operators, we seek conditions under which (3.28) is **not** satisfied. These conditions occur, of course, only with multipoint boundary value problems.

The Green's Function and the Boundary Kernel

Our procedure for inverting the system (3.21) parallels the procedure used with the second-order example (3.1). Of course, the compatibility condition (3.28) must be satisfied. Assume $m = n$. We begin by splitting (3.21) into two parts, one involving only the distributed input, the other only the boundary inputs:

$$\begin{aligned} \mathbf{L}\mathbf{f} &= \mathbf{u} \\ \beta_i(\mathbf{f}) &= 0 \quad i = 1, \dots, n \end{aligned} \quad (3.29)$$

$$\begin{aligned} \mathbf{L}\mathbf{f} &= \theta \\ \beta_i(\mathbf{f}) &= \alpha_i \quad i = 1, \dots, n \end{aligned} \quad (3.30)$$

where \mathbf{L} is given in (3.18). The completely homogeneous equation (3.22) is a special case of both (3.29) and (3.30). Thus both are characterized by any boundary condition matrix \mathbf{B} derived from (3.22). If (3.28) is satisfied, both (3.29) and (3.30) are invertible. The inverse of (3.29) is an integral operator with a distributed kernel. The inverse of (3.30) is a summation operator involving a boundary kernel. These two kernels describe explicitly the dependence of $\mathbf{f}(t)$ on the input data $\mathbf{u}(t)$ and $\{\alpha_j\}$.

Assume the inverse of (3.29) is representable in the integral form

$$\mathbf{f}(t) = \int_a^b k(t, s) \mathbf{u}(s) ds \quad (3.31)$$

for all t in $\{a, b\}$. The kernel k is known as the Green's function for the system (3.21). If (3.31) is the correct inverse for (3.29), \mathbf{f} must satisfy (3.29):

$$\begin{aligned}(\mathbf{L}\mathbf{f})(t) &= \mathbf{L} \int_a^b k(t, s) \mathbf{u}(s) ds \\ &= \int_a^b \mathbf{L}k(t, s) \mathbf{u}(s) ds = \mathbf{u}(t) \\ (\beta_i \mathbf{f})(t) &= \beta_i \int_a^b k(t, s) \mathbf{u}(s) ds \\ &= \int_a^b \beta_i k(t, s) \mathbf{u}(s) ds = 0 \quad i=1, \dots, n\end{aligned}$$

for all \mathbf{u} in $\mathcal{C}(a, b)$. Both \mathbf{L} and β_i treat the variable s as a constant, *acting on $k(t, s)$ only as a function of t* . Each operator acts on the whole “ t ” function $k(\cdot, s)$. It is evident that $\mathbf{L}k(t, s)$ exhibits the “sifting” property of a delta function (see Appendix 2). On the other hand, $\beta_i k(t, s)$ acts like the zero function. Consequently, the Green's function k must satisfy

$$\begin{aligned}\mathbf{L}k(t, s) &\stackrel{\Delta}{=} g_0(t) \frac{d^n k(t, s)}{dt^n} + \dots + g_n(t) k(t, s) = \delta(t - s) \\ \beta_j k(t, s) &= 0 \quad j=1, \dots, n\end{aligned}\tag{3.32}$$

for all t and s in $[a, b]$. Because the delta function appears in (3.32), we cannot rigorously interchange the order of the differential operator \mathbf{L} and the integration without resorting to the theory of generalized functions (Appendix 2). However, we can justify the formal interchange for each specific problem by showing that the Green's function k derived from (3.32) does indeed lead to the solution of (3.29) for every continuous function \mathbf{u} .

Assume the inverse of (3.30) is representable as a summation operator of the form:

$$\mathbf{f}(t) = \sum_{j=1}^n \rho_j(t) \alpha_j\tag{3.33}$$

We can think of ρ as a kernel function of the two variables j and t . We call ρ the boundary kernel for (3.21). To find the equations which determine ρ ,

we substitute (3.33) into (3.30):

$$\begin{aligned}\mathbf{L}\mathbf{f} &= \mathbf{L} \sum_{j=1}^n \rho_j \alpha_j \\ &= \sum_{j=1}^n \alpha_j (\mathbf{L}\rho_j) = \boldsymbol{\theta} \\ \beta_i(\mathbf{f}) &= \beta_i \sum_{j=1}^n \rho_j \alpha_j \\ &= \sum_{j=1}^n \alpha_j \beta_i(\rho_j) = \alpha_i \quad i=1, \dots, n\end{aligned}$$

for all $\{\alpha_i\}$. Suppose we let $\alpha_k = 1$ and $\alpha_j = 0$ for $j \neq k$. It follows that for $k=1, \dots, n$, $\mathbf{L}\rho_k = \boldsymbol{\theta}$, $\beta_i \rho_k = 1$ for $k=i$, and $\beta_i \rho_k = 0$ for $k \neq i$. Thus the boundary kernel ρ must satisfy

$$\begin{aligned}(\mathbf{L}\rho_j)(t) &\triangleq g_0(t) \frac{d^n \rho_j(t)}{dt^n} + \dots + g_n(t) \rho_j(t) = 0 \quad j=1, \dots, n \\ \beta_i(\rho_j) &= \delta_{ij} \quad i=1, \dots, n; \quad j=1, \dots, n\end{aligned} \quad (3.34)$$

for all t in $[a, b]$, where δ_{ij} is the Kronecker delta (see A2.11 of Appendix 2). According to (3.34), the n components $\{\rho_j\}$ of the boundary kernel constitute a fundamental set of solutions for the operator \mathbf{L} ; furthermore, $\{\rho_j\}$ is a fundamental set for which the boundary condition matrix \mathbf{B} of (3.27) is the $n \times n$ identity matrix.

By solving (3.32) and (3.34), we can invert any regular n th-order differential system which has a nonsingular boundary condition matrix. The inverse of the differential system (3.21) (with $m = n$) consists in the sum of the inverses of (3.29) and (3.30), namely,

$$\mathbf{f}(t) = \int_a^b k(t, s) \mathbf{u}(s) ds + \sum_{j=1}^n \alpha_j \rho_j(t) \quad (3.35)$$

where k and ρ are determined by (3.32) and (3.34), respectively.

Theoretically, we can invert any linear differential operator, ordinary or partial, which has appropriate boundary conditions. That is, we can convert any invertible linear differential equation to an integral equation analogous to (3.35). As a model for a system, the integral equation is more

desirable than the differential equation from two standpoints. First, the boundary conditions are included automatically. Second, integral operators tend to “smooth” functions whereas differential operators introduce discontinuities and delta functions.* It is well known that numerical differentiation amplifies errors in empirical data, but numerical integration does not (Ralston [3.9, p. 791]). The rest of this chapter is devoted to techniques for determining the inverse (or integral) model for various types of ordinary differential operators. Techniques and examples which apply to partial differential operators can be found in Friedman [3.4], Stakgold [3.11], Morse and Feshbach [3.8], and Bergman and Schiffer [3.1].

3.3 Inversion of n th-Order Differential Systems

In Section 3.1 we determined the Green’s function and boundary kernel for a simple second-order system, (3.1). The Green’s function and boundary kernel for the general n th-order differential systems of Section 3.2 cannot be determined by the direct integration technique used for that simple system. In this section we describe general procedures for solving (3.32) and (3.34) to obtain \mathbf{k} and $\boldsymbol{\rho}$ for the n th-order differential system (3.21) with n independent boundary conditions. The procedures are demonstrated in detail for regular second-order variable-coefficient differential systems.

Obtaining a Complementary Function

Most techniques for determining particular solutions to differential systems are based on the complementary function (3.19). Techniques for determining the Green’s function \mathbf{k} and the boundary kernel $\boldsymbol{\rho}$ also depend heavily on the complementary function (or the equivalent, a fundamental set of solutions). In point of fact, the individual segments or components of \mathbf{k} and $\boldsymbol{\rho}$ are of the form of the complementary function.

It is well known that the complementary function for a **constant-coefficient** differential operator consists in sums of exponentials. Let \mathbf{L} of (3.18) be the constant-coefficient operator

$$\mathbf{L} \stackrel{\Delta}{=} \mathbf{D}^n + a_1 \mathbf{D}^{n-1} + \cdots + a_n \mathbf{I} \quad (3.36)$$

To find which exponentials are contained in the complementary function for \mathbf{L} , we insert a particular exponential $\mathbf{v}(t) = e^{\mu t}$ into the equation $\mathbf{L}\mathbf{f} = \boldsymbol{\theta}$

*Integral operators are continuous, whereas differential operators are not. See the discussion of continuous operators in Section 5.4.

and solve for μ . The result is

$$\mu^n + a_1 \mu^{n-1} + \cdots + a_n = 0 \quad (3.37)$$

This equation, known as the **characteristic equation for \mathbf{L}** , has n roots $\mu_1, \mu_2, \dots, \mu_n$. If the n roots are distinct, the complementary function is

$$\mathbf{f}_c(\mathbf{t}) = c_1 \exp(\mu_1 t) + \cdots + c_n \exp(\mu_n t) \quad (3.38)$$

Equation (3.38) can be verified by substituting \mathbf{f}_c into $\mathbf{L}\mathbf{f} = \mathbf{0}$. If two roots are equal, say, $\mu_1 = \mu_2$, then the corresponding fundamental solutions in (3.38) must be replaced by $c_1 \exp(\mu_1 t) + c_2 t \exp(\mu_1 t)$. This equal root case is discussed further in Section 4.4.

We are unable to deal with the variable-coefficient operator (3.18) with much generality. An approach that can be used to *seek* the complementary function for the variable-coefficient operator is the **power series method** (the method of Frobenius). The method consists in assuming a power series form for the complementary function, substituting the series into the homogeneous differential equation, equating the coefficient on each power of t to zero, and solving for the coefficients of the power series. The sum of the series, where it converges, is at least part of the complementary function. The sum will not, in general, consist of elementary functions. For example, Bessel functions arise as fundamental solutions to Bessel's equation (a second-order variable-coefficient differential equation); the power series method provides an expression for one of the two fundamental solutions to Bessel's equation. In the event that the power series method does not provide a full set of fundamental solutions for the differential equation, other methods must be used to complete the complementary function. See Ince [3.6] or Wiley [3.13, p. 255].

Example 1. Power Series Method—Variable Coefficients Suppose

$$(\mathbf{L}\mathbf{f})(t) \triangleq \mathbf{f}'(t) + t\mathbf{f}(t) \quad (3.39)$$

We find the complementary function for (3.39) by assuming a power series of the general form

$$\mathbf{f}_c(t) = t^a (c_0 + c_1 t + c_2 t^2 + \cdots)$$

where the constant a allows for noninteger powers of t . We first insert \mathbf{f}_c into the homogeneous equation and regroup terms:

$$\begin{aligned} \mathbf{f}'(t) + t\mathbf{f}(t) &= ac_0 t^{a-1} + (a+1)c_1 t^a + [(a+2)c_2 + c_0]t^{a+1} \\ &\quad + [(a+3)c_3 + c_1]t^{a+2} + [(a+4)c_4 + c_2]t^{a+3} + \cdots \\ &= 0 \end{aligned}$$

Equating each coefficient to zero, we obtain

$$\begin{aligned}ac_0 &= 0 \\(a+1)c_1 &= 0 \\(a+2)c_2 + c_0 &= 0 \\(a+3)c_3 + c_1 &= 0 \\(a+4)c_4 + c_2 &= 0 \\&\vdots\end{aligned}$$

We assume, without loss of generality, that $c_0 \neq 0$. It follows that $a = 0$ and c_0 is arbitrary; then

$$\begin{aligned}c_1 = c_3 = c_5 = \cdots &= 0, \\c_2 = -\frac{c_0}{2}, \quad c_4 = \frac{c_0}{4(2)}, \quad c_6 = -\frac{c_0}{6(4)(2)},\end{aligned}$$

and

$$\begin{aligned}f_c(t) &= c_0 \left[1 - \frac{t^2}{2} + \frac{1}{2!} \left(\frac{t^2}{2} \right)^2 - \frac{1}{3!} \left(\frac{t^2}{2} \right)^3 + \cdots \right] \\&= c_0 \exp\left(-\frac{t^2}{2} \right)\end{aligned}$$

Determination of the Green's Function and Boundary Kernel—An Example

We solved for the kernel functions k and ρ associated with (3.1) by direct integration of the differential equation. Unfortunately, that simple approach does not apply to most differential equations. In the following example we introduce a general technique for finding k and ρ .

The model for a particular armature-controlled dc motor and load is the differential equation

$$\ddot{\phi}(t) + \dot{\phi}(t) = \mathbf{u}(t) \quad (3.40)$$

where $\mathbf{u}(t)$ is the armature voltage at time t and $\phi(t)$ is the angular position of the motor shaft relative to some reference position. Let the boundary conditions be

$$\phi(0) = \alpha_1 \quad \text{and} \quad \phi(b) = \alpha_2 \quad (3.41)$$

That is, we seek the "trajectory" (or angular position versus time), of the

shaft in order that it be in position α_1 at time 0 and pass through position α_2 at time b . Comparing this problem to that of (3.21), we note that $\mathbf{L} = \mathbf{D}^2 + \mathbf{D}$, $\beta_1(\phi) = \phi(0)$, and $\beta_2(\phi) = \phi(b)$. The symbol ϕ replaces the symbol \mathbf{f} used earlier.

Finding the Green's function for the differential system (3.40)-(3.41) is equivalent to exploring the trajectory ϕ of the motor shaft for all possible applied voltages $\mathbf{u}(t)$, but for $\alpha_1 = \alpha_2 = 0$. The Green's function must satisfy (3.32):

$$\frac{d^2k(t,s)}{dt^2} + \frac{dk(t,s)}{dt} = \delta(t-s)$$

$$k(0,s) = k(b,s) = 0$$

Clearly $k(t,s)$ satisfies the homogeneous differential equation in each of the regions $[0,s)$ and $(s,b]$; that is, in the regions where $\delta(t-s)$ is zero. We let $k(t,s) = \mathbf{f}_c(t)$ for each of the two regions $[0,s)$ and $(s,b]$:

$$k(t,s) = c_1 + c_2e^{-t}, \quad t \text{ in } [0,s)$$

$$= d_1 + d_2e^{-t}, \quad t \text{ in } (s,b]$$

Since $k(t,s)$ is a function of s , the arbitrary constants must depend on s . We eliminate half of the arbitrary constants by applying the boundary conditions

$$k(0,s) = c_1 + c_2 = 0 \quad \Rightarrow \quad c_2 = -c_1$$

$$k(b,s) = d_1 + d_2e^{-b} = 0 \quad \Rightarrow \quad d_2 = -e^b d_1$$

It is the second (or highest) derivative of k that introduces the delta function in (3.32); for if the first derivative included a delta function, the second derivative would introduce the derivative of the delta function.* Since d^2k/dt^2 includes a unit impulse at $t = s$, dk/dt must include a unit step at $t = s$, and k itself must be continuous at $t = s$. We express these facts by the two "discontinuity" conditions:

$$k(s^+,s) = k(s^-,s) \quad (\text{continuity of } k \text{ at } t = s)$$

$$\frac{dk(s^+,s)}{dt} - \frac{dk(s^-,s)}{dt} = 1 \quad \left(\text{unit step in } \frac{dk}{dt} \text{ at } t = s \right)$$

*See Appendix 2 for a discussion of unit steps, delta functions, and derivatives of delta functions.

Applying these conditions to $k(t,s)$, we find

$$d_1 + d_2 e^{-s} = c_1 + c_2 e^{-s}$$

$$-d_2 e^{-s} - (-c_2 e^{-s}) = 1$$

A messy elimination procedure among the boundary condition equations and discontinuity condition equations yields

$$c_1(s) = \frac{e^s - e^b}{e^b - 1} \quad \text{and} \quad d_1(s) = \frac{e^s - 1}{e^b - 1}$$

It follows that

$$\begin{aligned} k(t,s) &= \frac{(1 - e^{-t})(e^s - e^b)}{e^b - 1} & t \leq s \\ &= \frac{(1 - e^b e^{-t})(e^s - 1)}{e^b - 1} & t \geq s \end{aligned} \quad (3.42)$$

To get a feel for the nature of this system (for which $\phi(0) = \phi(b) = 0$), we use k to determine the shaft trajectory ϕ and velocity profile $\dot{\phi}$ for a specific input $\mathbf{u}(t) = 1$:

$$\begin{aligned} \phi(t) &= \int_0^b k(t,s) \mathbf{u}(s) ds \\ &= \frac{1 - e^b e^{-t}}{e^b - 1} \int_0^t (e^s - 1) ds + \frac{1 - e^{-t}}{e^b - 1} \int_t^b (e^s - e^b) ds \\ &= t - \left(\frac{be^b}{e^b - 1} \right) (1 - e^{-t}) \\ \dot{\phi}(t) &= 1 - \left(\frac{be^b}{e^b - 1} \right) e^{-t} \end{aligned}$$

The trajectory ϕ and the velocity profile $\dot{\phi}$ are plotted in Figure 3.2 for $b = 1$. Observe that, in general, the motor shaft cannot be at rest at $t = 0$ and at $t = b$ if the shaft positions are specified; it is precisely the freedom in the initial and terminal velocities which allows us to choose both the end points, $\phi(0)$ and $\phi(b)$, and an arbitrary continuous input voltage \mathbf{u} .

The boundary kernel ρ for the system (3.40)-(3.41) describes the trajectory $\phi(t)$ as a function of the boundary conditions $\phi(0) = \alpha_1$ and

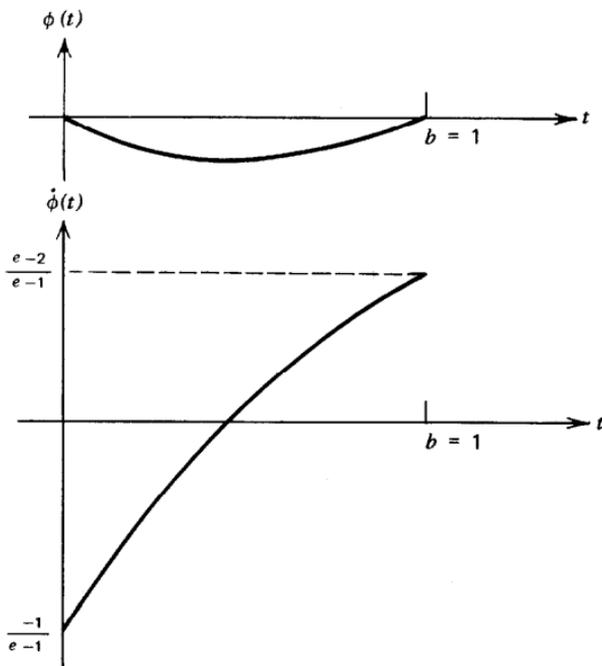


Figure 3.2. Shaft position and velocity for $\phi(0) = \phi(1) = 0$ and $u(t) = 1$.

$\phi(b) = \alpha_2$ with no voltage applied to the motor; that is,

$$\phi(t) = \rho_1(t)\alpha_1 + \rho_2(t)\alpha_2$$

Perhaps the most direct approach to the determination of $\rho_j(t)$ is to let $\phi(t) = c_1 \mathbf{v}_1(t) + c_2 \mathbf{v}_2(t)$, a linear combination of the fundamental solutions for (3.40), then apply the boundary conditions (3.41) to obtain the coefficients c_i as a function of α_1 and α_2 . Rather than use this approach, we attack the defining equations for $\rho_j(t)$ in a more formal manner which parallels the determination of the Green's function. The two approaches are equivalent in the amount of computation they require. The boundary kernel satisfies (3.34):

$$\begin{aligned} \ddot{\rho}_1(t) + \dot{\rho}_1(t) &= 0 & \ddot{\rho}_2(t) + \dot{\rho}_2(t) &= 0 \\ \beta_1(\rho_1) = \rho_1(0) &= 1 & \beta_1(\rho_2) = \rho_2(0) &= 0 \\ \beta_2(\rho_1) = \rho_1(b) &= 0 & \beta_2(\rho_2) = \rho_2(b) &= 1 \end{aligned}$$

The boundary condition statements are reminiscent of the boundary condition matrix (3.27). In point of fact, ρ_1 and ρ_2 each consist in a linear combination of the fundamental solutions $\mathbf{v}_1(t) = 1$ and $\mathbf{v}_2(t) = e^{-t}$. Apply-

ing the boundary conditions to $\rho_1(t) = c_1 + c_2 e^{-t}$, we get

$$\beta_1(\rho_1) = \rho_1(0) = c_1 + c_2 e^{-0} = 1$$

$$\beta_2(\rho_1) = \rho_1(b) = c_1 + c_2 e^{-b} = 0$$

or

$$\mathbf{B} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & e^{-b} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

where \mathbf{B} is, indeed, the boundary condition matrix of (3.27). Similarly, using $\rho_2(t) = d_1 + d_2 e^{-t}$, we find

$$\mathbf{B} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

We can combine the two coefficient equations into the single matrix equation

$$\mathbf{B} \begin{pmatrix} c_1 & d_1 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which has the solution

$$\begin{pmatrix} c_1 & d_1 \\ c_2 & d_2 \end{pmatrix} = \mathbf{B}^{-1} = \begin{pmatrix} 1 \\ e^b - 1 \end{pmatrix} \begin{pmatrix} -1 & e^b \\ e^b & -e^b \end{pmatrix}$$

The function ρ_j is a specific linear combination of the two fundamental solutions specified above; the j th column of \mathbf{B}^{-1} specifies the linear combination. Thus

$$\rho_1(t) = \frac{-1}{e^b - 1} + \frac{e^b}{e^b - 1} e^{-t} \tag{3.43}$$

$$\rho_2(t) = \frac{e^b}{e^b - 1} + \frac{-e^b}{e^b - 1} e^{-t}$$

The shaft position and velocity, as functions of the boundary conditions, are

$$\phi(t) = \frac{e^b e^{-t} - 1}{e^b - 1} \alpha_1 + \frac{e^b (1 - e^{-t})}{e^b - 1} \alpha_2$$

$$\dot{\phi}(t) = \frac{e^b e^{-t}}{e^b - 1} (\alpha_2 - \alpha_1)$$

Figure 3.3 shows the position and velocity of the motor shaft for $\alpha_1 = 0$ and $\alpha_2 = 1$. The shaft is already in motion at $t = 0$, and exhibits an “undriven” decay in velocity until it reaches the position $\phi(b) = 1$ rad. If the boundary conditions were $\alpha_1 = \alpha_2 = 1$, the shaft would sit at rest in the position $\phi(t) = 1$ rad; again an undriven trajectory.

The inverse of the system (3.40)-(3.41) is the sum of the separate solutions for the distributed and boundary inputs. That is,

$$\mathbf{f}(t) = \int_0^b k(t,s)\mathbf{u}(s)ds + \rho_1(t)\alpha_1 + \rho_2(t)\alpha_2$$

where k and ρ are given in (3.42) and (3.43), respectively. The nature of the system (3.40)-(3.41) does not seem in keeping with the nature of dynamic [real-time) systems. The motor must anticipate the input $\mathbf{u}(t)$ (or the impulse $\delta(t-s)$) and appropriately select its velocity at $t = 0$ in order to be able to meet the requirement on its position at $t = b$. We are more likely to meet such a two-point boundary value problem when the independent variable t represents not time, but rather a space variable. Yet a two-point

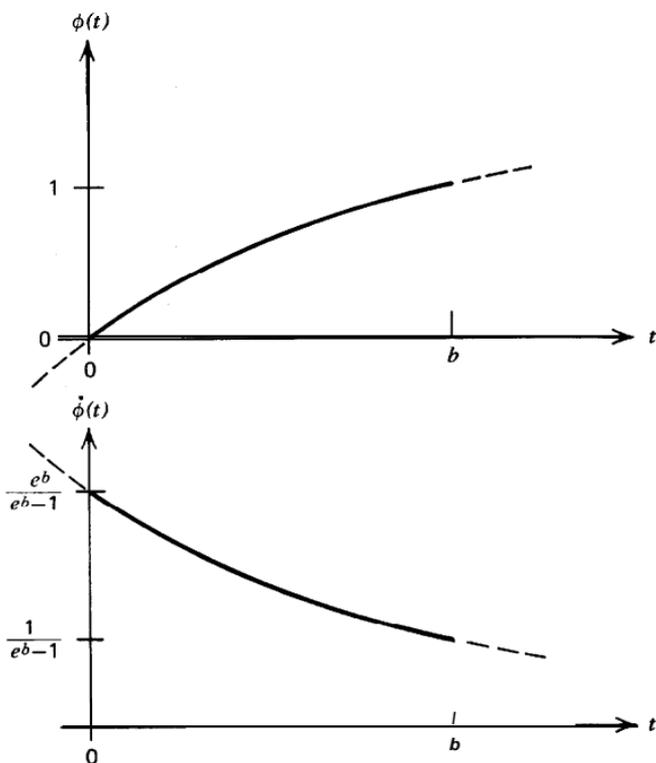


Figure 3.3. Undriven shaft position and velocity for $\alpha_1 = 0$ and $\alpha_2 = 1$.

boundary value problem can arise in a dynamic system if we impose requirements on the future behavior of the system as we did in (3.41).

Summary of the Technique

The technique demonstrated above for determining the Green's function and the boundary kernel depends upon knowledge of the complementary function. We can apply the technique to the regular n th-order system (3.21) if the corresponding complementary function can be determined. Assume \mathbf{L} of (3.21) has the complementary function $\mathbf{f} = c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n$. Further assume that the system is invertible (i.e., we have n independent boundary conditions for which (3.28) is satisfied). We obtain the Green's function \mathbf{k} and the boundary kernel ρ for the system (3.21) by following the technique used for the system (3.40)-(3.41).

Equation (3.32) determines the Green's function \mathbf{k} . The unit impulse $\delta(t-s)$ is zero for all $t \neq s$. Therefore, $\mathbf{k}(t,s)$ satisfies the homogeneous differential equation for $t \neq s$; $\mathbf{k}(t,s)$ is equal to the complementary function (3.19) in each of the two regions $[a,s)$ and $(s,b]$. Because the complementary function \mathbf{f}_c is used in two separate regions, we must determine two sets of n arbitrary constants:

$$\begin{aligned} \mathbf{k}(t,s) &= b_1 \mathbf{v}_1(t) + \dots + b_n \mathbf{v}_n(t), & t \text{ in } [a,s) \\ &= d_1 \mathbf{v}_1(t) + \dots + d_n \mathbf{v}_n(t), & t \text{ in } (s,b] \end{aligned} \quad (3.44)$$

Half of the $2n$ constants can be eliminated by the *homogeneous* boundary conditions of (3.21): $\beta_i \mathbf{k}(t,s) = \mathbf{0}$, $i = 1, \dots, n$. The rest are determined by appropriate "discontinuity" conditions at $t = s$. Only the highest derivative term, $\mathbf{g}_0(t) d^n \mathbf{k}(t,s) / dt^n$, can introduce the delta function into (3.32) (otherwise derivatives of delta functions would appear); therefore, we match the two halves of $\mathbf{k}(t,s)$ at $t = s$ in such a way that we satisfy the following n conditions:

$$\begin{aligned} \mathbf{k}, \frac{d\mathbf{k}}{dt}, \dots, \frac{d^{n-2}\mathbf{k}}{dt^{n-2}} & \text{ are continuous at } t = s \\ \frac{d^{n-1}\mathbf{k}(s^+,s)}{dt^{n-1}} - \frac{d^{n-1}\mathbf{k}(s^-,s)}{dt^{n-1}} &= \frac{1}{\mathbf{g}_0(s)} \end{aligned} \quad (3.45)$$

That is, $d^{n-1}\mathbf{k}(t,s)/dt^{n-1}$ must contain a step of size $1/\mathbf{g}_0(s)$ at $t = s$. Then $\mathbf{g}_0(t) d^n \mathbf{k}(t,s) / dt^n$ will include the term $\delta(t-s)$.*

*See Appendix 2 for a discussion of steps, delta functions, and derivatives of delta functions.

The boundary kernel ρ is specified by (3.34). Each component of ρ is a linear combination of the fundamental solutions for \mathbf{L} :

$$\rho_j = c_{1j}\mathbf{v}_1 + \cdots + c_{nj}\mathbf{v}_n \quad j = 1, \dots, n \quad (3.46)$$

Applying the n boundary conditions of (3.21) as required by (3.34), we find

$$\begin{pmatrix} \beta_1(\rho_j) \\ \vdots \\ \beta_n(\rho_j) \end{pmatrix} = \begin{pmatrix} \beta_1(\mathbf{v}_1) \cdots \beta_1(\mathbf{v}_n) \\ \vdots \\ \beta_n(\mathbf{v}_1) \cdots \beta_n(\mathbf{v}_n) \end{pmatrix} \begin{pmatrix} c_{1j} \\ \vdots \\ c_{nj} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1_j \\ \vdots \\ 0 \end{pmatrix}, \quad j = 1, \dots, n$$

These n sets of equations can be expressed as

$$\begin{pmatrix} \beta_1(\mathbf{v}_1) \cdots \beta_1(\mathbf{v}_n) \\ \vdots \\ \beta_n(\mathbf{v}_1) \cdots \beta_n(\mathbf{v}_n) \end{pmatrix} \begin{pmatrix} c_{11} \cdots c_{1n} \\ \vdots \\ c_{n1} \cdots c_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \quad (3.47)$$

It follows that the coefficients for ρ_j in (3.46) are the elements in the j th column of \mathbf{B}^{-1} , where \mathbf{B} is the boundary condition matrix defined in (3.27). Specifically, c_{ij} is the element in row i and column j of \mathbf{B}^{-1} .

Exercise 1. Let $\mathbf{f}'(t) + t\mathbf{f}(t) = \mathbf{u}(t)$ with $\mathbf{f}(0) = \alpha_1$. (The complementary function for this differential equation was determined in Example 1.) Show that the inverse of this differential system is

$$\mathbf{f}(t) = \exp(-t^2/2) \int_0^t \exp(s^2/2) \mathbf{u}(s) ds + \alpha_1 \exp(-t^2/2) \quad (3.48)$$

Second-Order Differential Systems

Many of the ordinary and partial differential equations that arise in the modeling of physical systems are second order. Some of the second-order partial differential equations can be reduced, by a substitution of variables or by integral transforms, to second-order ordinary differential equations.* Furthermore, use of the "separation of variables" technique in solving second-order partial differential equations produces sets of second-order

*See Kaplan [3.7].

ordinary differential equations. Thus the general second-order ordinary differential equation with variable coefficients is of considerable practical importance. We present explicit expressions for the Green's function and the boundary kernel for an arbitrary regular second-order differential system; these expressions are obtained in terms of a fundamental set of solutions for the differential operator.

The regular second-order differential system is[†]

$$\begin{aligned} g_0(t)\mathbf{f}''(t) + g_1(t)\mathbf{f}'(t) + g_2(t)\mathbf{f}(t) &= \mathbf{u}(t) \\ \beta_1(\mathbf{f}) &= \alpha_1 \quad \text{and} \quad \beta_2(\mathbf{f}) = \alpha_2 \end{aligned} \quad (3.49)$$

where g_i is continuous and $g_0(t) \neq 0$ in the region of interest. Assume \mathbf{v}_1 and \mathbf{v}_2 are independent solutions to the homogeneous differential equation. By (3.44), the Green's function is of the form

$$\begin{aligned} k(t, s) &= b_1\mathbf{v}_1(t) + b_2\mathbf{v}_2(t), \quad t < s \\ &= d_1\mathbf{v}_1(t) + d_2\mathbf{v}_2(t), \quad t > s \end{aligned}$$

The discontinuity conditions (3.45) become

$$d_1\mathbf{v}_1(s) + d_2\mathbf{v}_2(s) = b_1\mathbf{v}_1(s) + b_2\mathbf{v}_2(s) \quad (\text{continuity of } k)$$

$$d_1\mathbf{v}'_1(s) + d_2\mathbf{v}'_2(s) - b_1\mathbf{v}'_1(s) - b_2\mathbf{v}'_2(s) = \frac{1}{g_0(s)} \left(\text{step of size } \frac{1}{g_0(s)} \text{ in } \frac{dk}{dt} \right)$$

Since dk/dt has a step of size $1/g_0(s)$, then $g_0(t)d^2k/dt^2$ includes a *unit* impulse. These two discontinuity equations can be put in the matrix form

$$\begin{pmatrix} \mathbf{v}_1(s) & \mathbf{v}_2(s) \\ \mathbf{v}'_1(s) & \mathbf{v}'_2(s) \end{pmatrix} \begin{pmatrix} d_1 - b_1 \\ d_2 - b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1/g_0(s) \end{pmatrix}$$

The solution is

$$d_1 - b_1 = -\frac{\mathbf{v}_2(s)}{\mathbf{w}(s)g_0(s)}, \quad d_2 - b_2 = \frac{\mathbf{v}_1(s)}{\mathbf{w}(s)g_0(s)}$$

where $\mathbf{w}(s)$ is the *Wronskian determinant**:

$$\mathbf{w}(s) \triangleq \begin{vmatrix} \mathbf{v}_1(s) & \mathbf{v}_2(s) \\ \mathbf{v}'_1(s) & \mathbf{v}'_2(s) \end{vmatrix} \quad (3.50)$$

[†] In Section 5.5 we refer to the differential operator of (3.49) as a regular Sturm-Liouville operator.

*Note that the solution is undefined for $\mathbf{w}(s) = 0$. It can be shown that if \mathbf{v}_1 and \mathbf{v}_2 are independent solutions to the homogeneous differential equation, then $\mathbf{w}(s) \neq 0$ for all s in the interval of interest. See P&C 3.7.

The boundary conditions $\beta_1 k(t, s) = \beta_2 k(t, s) = 0$ provide two more linear algebraic equations which, together with the above pair of equations, determine the constants b_1 , b_2 , d_1 , and d_2 , and therefore, $k(t, s)$. However, without specific information about the nature of the boundary conditions, we can carry the solution no further. The solution for a dynamic system (initial conditions) is given in Exercise 2. Two-point boundary conditions are treated in Exercise 3.

Exercise 2. Let the boundary conditions of (3.49) be

$$\beta_1(\mathbf{f}) \stackrel{\Delta}{=} \mathbf{f}(a) \text{ and } \beta_2(\mathbf{f}) \stackrel{\Delta}{=} \mathbf{f}'(a) \quad (3.51)$$

Show that the corresponding Green's function is

$$\begin{aligned} k(t, s) &= 0, & t \text{ in } [a, s) \\ &= \frac{\Delta(s, t)}{g_0(s)w(s)}, & t \text{ in } (s, \infty) \end{aligned} \quad (3.52)$$

where w is given by (3.50), and

$$\Delta(s, t) \stackrel{\Delta}{=} \begin{vmatrix} \mathbf{v}_1(s) & \mathbf{v}_2(s) \\ \mathbf{v}_1(t) & \mathbf{v}_2(t) \end{vmatrix}$$

Show also that the corresponding boundary kernel is

$$\begin{aligned} \rho_1(t) &= \frac{\mathbf{v}'_2(a)\mathbf{v}_1(t) - \mathbf{v}'_1(a)\mathbf{v}_2(t)}{w(a)} \\ \rho_2(t) &= \frac{\mathbf{v}_1(a)\mathbf{v}_2(t) - \mathbf{v}_2(a)\mathbf{v}_1(t)}{w(a)} \end{aligned} \quad (3.53)$$

Exercise 3. Let the boundary conditions of (3.49) be $\beta_1(\mathbf{f}) \stackrel{\Delta}{=} \mathbf{f}(a)$ and $\beta_2(\mathbf{f}) \stackrel{\Delta}{=} \mathbf{f}(b)$. Show that for this two-point boundary value problem

$$\begin{aligned} k(t, s) &= \frac{1}{g_0(s)w(s)} \begin{cases} \frac{\Delta(b, s)\Delta(a, t)}{\Delta(a, b)}, & a \leq t \leq s \\ \frac{\Delta(b, s)\Delta(a, t)}{\Delta(a, b)} + \Delta(s, t), & s \leq t \leq b \end{cases} \\ \rho_1(t) &= -\frac{\Delta(b, t)}{\Delta(a, b)} \quad \text{and} \quad \rho_2(t) = \frac{\Delta(a, t)}{\Delta(a, b)} \end{aligned} \quad (3.54)$$

where $\Delta(s, t)$ is given beneath (3.52) and $w(s)$ is defined in (3.50).

Exercise 4. Use (3.54) to find k and ρ for the dc motor system (3.40)-(3.41). Compare the result with (3.42) and (3.43).

It is apparent that we could derive an explicit expression for the inverse of a regular n th-order linear differential system [assuming the boundary conditions satisfy the invertibility condition (3.28)]. The inverse would involve n independent fundamental solutions and the n th-order Wronskian determinant of these n solutions. Of course, as indicated by Exercise 3, the manipulation can be complicated. The determination of the Green's function for an n th-order two-point boundary value problem requires the solution of $2n$ simultaneous algebraic equations with coefficients which are functions of s . In contrast, the Green's function for the initial condition problem (or one-point boundary value problem) requires the solution of only n simultaneous equations because $k(t, s) = 0$ for $t < s$. Of particular interest is the constant-coefficient initial condition problem, for which determination of the Green's function reduces to inversion of an $n \times n$ matrix of constants.

3.4 Time-Invariant Dynamic Systems

The initial value problem is at the heart of dynamic systems—systems for which the variable t represents time. The linear time-invariant (or constant-coefficient) dynamic system merits special attention if only because its inversion is easily automated using standard computer programs for solving matrix equations. Furthermore, many dynamic systems are adequately represented as linear time-invariant systems. We examine these systems in detail in this section.

The Inverse of the n th-Order System

The general n th-order constant-coefficient differential equation with initial conditions is

$$\mathbf{f}^{(n)}(t) + a_1 \mathbf{f}^{(n-1)}(t) + \cdots + a_n \mathbf{f}(t) = \mathbf{u}(t) \quad (3.55)$$

$$\beta_i(\mathbf{f}) \triangleq \mathbf{f}^{(i-1)}(0) = \alpha_i \quad i = 1, \dots, n$$

for real scalars $\{a_i\}$ and $t \geq 0$. The characteristic equation for (3.55) is (3.37); assume it has n distinct roots μ_1, \dots, μ_n (the multiple root case is considered in Section 4.4). Then the fundamental solutions for (3.55) are $\mathbf{v}_i(t) \triangleq \exp(\mu_i t)$, $i = 1, \dots, n$.

The Green's function, as given by (3.44), is

$$k(t,s) = b_1 \exp(\mu_1 t) + \dots + b_n \exp(\mu_n t), \quad t \text{ in } [0,s)$$

$$= d_1 \exp(\mu_1 t) + \dots + d_n \exp(\mu_n t), \quad t \text{ in } (s, \infty)$$

All n boundary conditions apply to the first half of $k(t,s)$, the half involving the unknowns b_1, \dots, b_n . As a result,

$$\begin{pmatrix} k(0,s) \\ \frac{dk}{dt}(0,s) \\ \vdots \\ \frac{d^{n-1}k}{dt^{n-1}}(0,s) \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ \mu_1 & \dots & \mu_n \\ \vdots & & \vdots \\ \mu_1^{n-1} & \dots & \mu_n^{n-1} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

This boundary condition matrix is the **Wronskian matrix** of the functions $\{\exp(\mu_i t)\}$ at $t = 0$. The matrix is also known as the **Vandermond matrix** for the system (3.55). It is invertible if and only if the roots μ_1, \dots, μ_n are distinct as assumed.* Therefore, $b_1 = \dots = b_n = 0$, and $k(t,s) = 0$ for t in $[0,s)$. The discontinuity conditions (3.45) at $t = s$ are

$$d_1 \exp(\mu_1 s) + \dots + d_n \exp(\mu_n s) = 0 \quad (k \text{ continuous})$$

$$d_1 \mu_1 \exp(\mu_1 s) + \dots + d_n \mu_n \exp(\mu_n s) = 0 \quad (dk/dt \text{ continuous})$$

$$d_1 \mu_1^{n-2} \exp(\mu_1 s) + \dots + d_n \mu_n^{n-2} \exp(\mu_n s) = 0 \quad (d^{n-2}k/dt^{n-2} \text{ continuous})$$

$$d_1 \mu_1^{n-1} \exp(\mu_1 s) + \dots + d_n \mu_n^{n-1} \exp(\mu_n s) = 1 \quad (\text{unit step in } d^{n-1}k/dt^{n-1})$$

We substitute the new variables $\hat{d}_i \triangleq d_i \exp(\mu_i s)$, $i = 1, \dots, n$ into the discontinuity equations to obtain

$$\begin{pmatrix} 1 & \dots & 1 \\ \mu_1 & \dots & \mu_n \\ \vdots & & \vdots \\ \mu_1^{n-1} & \dots & \mu_n^{n-1} \end{pmatrix} \begin{pmatrix} \hat{d}_1 \\ \vdots \\ \hat{d}_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad (3.56)$$

*If the roots were not distinct, we would use a different set of fundamental solutions $\{\mathbf{v}_i\}$, and obtain a different boundary condition matrix. The Wronskian matrix is explored in P&C 3.7. The Vandermond matrix is examined in P&C 4.16.

Because the roots $\{\mu_i\}$ are distinct, the Vandermond matrix is invertible, and (3.56) can be solved by means of a standard computer program to obtain $\{d_i\}$. Notice that the new variables $\{d_i\}$ are independent of s . The s dependence of the variables $\{d_i\}$ has been removed by the substitution. In terms of the new variables, the Green's function becomes

$$k(t,s) = 0 \quad \text{for } 0 \leq t \leq s$$

$$= \hat{d}_1 \exp[\mu_1(t-s)] + \cdots + \hat{d}_n \exp[\mu_n(t-s)] \quad \text{for } t \geq s \quad (3.57)$$

The boundary kernel for the system (3.55) is found from (3.46) and (3.47). Equation (3.47) is

$$\begin{pmatrix} 1 & \cdots & 1 \\ \mu_1 & \cdots & \mu_n \\ \vdots & & \vdots \\ \mu_1^{n-1} & \cdots & \mu_n^{n-1} \end{pmatrix} \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} = \mathbf{I}$$

Then, by (3.46),

$$\rho_j(t) = c_{1j} \exp(\mu_1 t) + \cdots + c_{nj} \exp(\mu_n t), \quad j = 1, \dots, n \quad (3.58)$$

where the coefficients for ρ_j are obtained from the j th column of the inverse Vandermond matrix:

$$\begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \mu_1 & \cdots & \mu_n \\ \vdots & & \vdots \\ \mu_1^{n-1} & \cdots & \mu_n^{n-1} \end{pmatrix}^{-1} \quad (3.59)$$

The inverse of the differential equation and boundary conditions of (3.55) is

$$\mathbf{f}(t) = \int_0^\infty k(t,s) \mathbf{u}(s) ds + \sum_{j=1}^n \rho_j(t) \alpha_j$$

$$= \int_0^t \{ \hat{d}_1 \exp[\mu_1(t-s)] + \cdots + \hat{d}_n \exp[\mu_n(t-s)] \} \mathbf{u}(s) ds$$

$$+ \sum_{j=1}^n [c_{1j} \exp(\mu_1 t) + \cdots + c_{nj} \exp(\mu_n t)] \mathbf{f}^{(j-1)}(0) \quad (3.60)$$

where $\{\hat{d}_i\}$ and $\{c_{ij}\}$ are specified by (3.56) and (3.59), respectively. The computer program which produces (3.59) will simultaneously solve (3.56). Section 4.4 explores the computational difficulties which arise when the characteristic equation of the system has nearly equal roots.

The shape of the Green's function for a time-invariant (i.e., constant-coefficient) dynamic system depends only on $t-s$, the delay between the time s that an impulse is applied at the system input and the time t that the output $k(t,s)$ is observed. That is, $k(t,s) = k(t-s, 0)$. Therefore, actual measurement of the response of the physical system to an approximate impulse is a suitable method for determining the Green's function. The response of such a system, initially at rest, to an impulse input $\mathbf{u}(t) = \delta(t)$ is commonly referred to as the **impulse response** of the system. We denote the impulse response by \mathbf{g} , where $\mathbf{g}(t) \triangleq k(t, 0)$. Then the integral term in (3.60) can be rewritten as a convolution of \mathbf{u} and \mathbf{g} .*

$$\int_0^t k(t,s)\mathbf{u}(s) ds = \int_0^t \mathbf{g}(t-s)\mathbf{u}(s) ds$$

The components of the boundary kernel also can be measured physically; $\rho_j(t)$ is the response of the system with no distributed input \mathbf{u} , and with the initial conditions $\alpha_j = 1$, $\alpha_i = 0$, for $i \neq j$. Furthermore, we see from (3.56)-(3.59) that $\hat{d}_i = c_{in}$ for $i = 1, \dots, n$. Therefore, the impulse response is equal to one of the initial condition responses; specifically,

$$\rho_n(t) = k(t, 0) = \mathbf{g}(t) \quad (3.61)$$

Applying a unit impulse $\delta(t)$ is equivalent to instantaneously applying to the system (at rest) the unit initial condition $\mathbf{f}^{(n-1)}(0) = 1$ (all other initial conditions remaining zero); if we can apply this initial condition some other way, we do not need an approximate impulse in order to measure the impulse response of the system.

Exercise 1. The differential equation (3.40) for an armature-controlled dc motor is

$$\ddot{\phi}(t) + \dot{\phi}(t) = \mathbf{u}(t)$$

Show that for given initial conditions, $\phi(0)$ and $\dot{\phi}(0)$, the Green's function,

*See Appendix 2 for a discussion of convolution.

boundary kernel, and inverse equation are

$$\begin{aligned}
 k(t,s) &= 0, & 0 \leq t \leq s \\
 &= 1 - e^{-(t-s)}, & t \geq s \\
 \rho_1(t) &= 1 \\
 \rho_2(t) &= 1 - e^{-t} \\
 \phi(t) &= \int_0^t [1 - e^{-(t-s)}] \mathbf{u}(s) ds + \phi(0) + (1 - e^{-t}) \dot{\phi}(0)
 \end{aligned} \tag{3.62}$$

Compare (3.62) with (3.52) and (3.53).

The State-Space Model

The n th-order constant-coefficient differential equation with initial conditions, (3.59), can be expressed as a first-order vector differential equation by redefining the variables. If $\mathbf{u}(t) = 0$, the quantities $\mathbf{f}(0), \mathbf{f}^{(1)}(0), \dots, \mathbf{f}^{(n-1)}(0)$ determine the trajectory $\mathbf{f}(t)$ for all t ; these n quantities together form a more complete description of the state (or condition) of the system at $t = 0$ than does $\mathbf{f}(0)$ alone. Let $\mathbf{f}_1 \triangleq \mathbf{f}, \mathbf{f}_2 \triangleq \mathbf{f}^{(1)}, \dots, \mathbf{f}_n \triangleq \mathbf{f}^{(n-1)}$. Then (3.55) can be expressed as the following set of n first-order differential equations.

$$\begin{aligned}
 \dot{\mathbf{f}}_1(t) &= \mathbf{f}_2(t) \\
 \dot{\mathbf{f}}_2(t) &= \mathbf{f}_3(t) \\
 &\vdots \\
 \dot{\mathbf{f}}_{n-1}(t) &= \mathbf{f}_n(t) \\
 \dot{\mathbf{f}}_n(t) &= -a_n \mathbf{f}_1(t) - \dots - a_1 \mathbf{f}_n(t) + \mathbf{u}(t)
 \end{aligned}$$

By defining $\mathbf{x} \triangleq (\mathbf{f}_1 \dots \mathbf{f}_n)^T$ and $\dot{\mathbf{x}} \triangleq (\dot{\mathbf{f}}_1 \dots \dot{\mathbf{f}}_n)^T$, we write the n individual equations as

$$\dot{\mathbf{x}}(t) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_2 & -a_1 \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \mathbf{u}(t) \tag{3.63}$$

The square matrix of (3.63) is known as the **companion matrix** for the n th

order differential operator (3.55). The initial conditions of (3.55) become

$$\mathbf{x}(0) = \begin{pmatrix} \mathbf{f}(0) \\ \mathbf{f}^{(1)}(0) \\ \vdots \\ \mathbf{f}^{(n-1)}(0) \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \quad (3.64)$$

We call $\mathbf{x}(t)$ the *state vector* of the system at time t . Since the differential system (3.55) has a unique solution, the state at time t can be determined from the state at any time previous to t . The state provides precisely enough information concerning the condition of the system to determine the future behavior of the system for a given input. The vector $\mathbf{x}(t)$ is in $\mathcal{R}^n \times 1$. Therefore, we call $\mathcal{R}^n \times 1$ the *state space* of the system. The variables $\{\mathbf{f}_i(t)\}$ are known as *state variables*.

Equations (3.63) and (3.64) are of the general form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) \text{ given} \quad (3.65)$$

However, the notation of (3.65) is more general than that of (3.63) and (3.64). The input \mathbf{u} can include more than one function. A meaningful equation is defined by any $n \times n$ matrix \mathbf{A} and $n \times m$ matrix \mathbf{B} ; the resulting vector equation describes the evolution in time of a system with m inputs and n outputs. A general set of coupled linear time-invariant differential equations can be expressed in this state-space form (P&C 3.18). We refer to (3.65) as a **state equation**. We call $\mathbf{x}(t)$ the **state vector** and its elements the **state variables**; \mathbf{A} and \mathbf{B} are the **system matrix** and the **input matrix**, respectively.*

We should note that the description of a dynamic system by a state-space model is not unique. If we multiply both sides of (3.65) by an arbitrary invertible $n \times n$ matrix \mathbf{S} , we obtain

$$\mathbf{S}\dot{\mathbf{x}}(t) = \mathbf{S}\mathbf{A}\mathbf{x}(t) + \mathbf{S}\mathbf{B}\mathbf{u}(t)$$

Defining $\mathbf{y} = \mathbf{S}\mathbf{x}$, we find

$$\begin{aligned} \dot{\mathbf{y}}(t) &= \mathbf{S}\mathbf{A}\mathbf{S}^{-1}\mathbf{y}(t) + \mathbf{S}\mathbf{B}\mathbf{u}(t) \\ &= \hat{\mathbf{A}}\mathbf{y}(t) + \hat{\mathbf{B}}\mathbf{u}(t) \end{aligned}$$

*See Zadeh and Desoer [3.14] or DeRusso, Roy, and Close [3.2] for a more complete discussion of state-space models.

with $\mathbf{y}(0) = \mathbf{S}\mathbf{x}(0)$ given. This second state-space differential equation is equivalent to (3.65) as a representative of the system. The state vector $\mathbf{y}(t)$ is a representation of $\mathbf{x}(t)$ in new coordinates. Thus the state variables and system matrix which describe a given system are not unique. In Section 4.2 we explore the essential characteristics of a matrix, its eigenvalues. We find that the similarity transformation $\mathbf{S}\mathbf{A}\mathbf{S}^{-1}$ does not affect the eigenvalues. Consequently, all system matrices which represent the same system have the same essential characteristics. State space models of dynamic systems are analyzed in terms of their eigenvalues in Sections 4.3 and 4.5.

Example 1. A State Equation. The differential equation for the armature-controlled dc motor of Exercise 1 is

$$\ddot{\phi}(t) + \dot{\phi}(t) = \mathbf{u}(t), \quad \phi(0) = \alpha_1, \quad \dot{\phi}(0) = \alpha_2$$

Defining the state variables $\mathbf{f}_1(t) \triangleq \phi(t)$ and $\mathbf{f}_2(t) \triangleq \dot{\phi}(t)$, we obtain the following state equation

$$\dot{\mathbf{x}}(t) = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mathbf{u}(t), \quad \mathbf{x}(0) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

The system matrix is the companion matrix for the second-order differential equation.

Let us find an integral equation which is the inverse (or explicit solution) of the first-order vector-valued differential system (3.65). Although we work directly with the system in the specific form (3.65), we note that the equation can be expressed in terms of a general differential operator \mathbf{L} acting on a vector-valued function space. Let \mathbf{f} be in $\mathcal{C}^n(0, \infty)$; then $\mathbf{f}^{(k)}$ is in $\mathcal{C}^{n-k}(0, \infty)$ and \mathbf{x} is in the Cartesian product space:

$$\mathcal{V} = \mathcal{C}^n(0, \infty) \times \mathcal{C}^{n-1}(0, \infty) \times \cdots \times \mathcal{C}^1(0, \infty)$$

The system (3.65) is equivalent to the following operator equation on \mathcal{V} :

$$\mathbf{L}\mathbf{x} \triangleq \dot{\mathbf{x}} - \mathbf{A}\mathbf{x} = \mathbf{B}\mathbf{u} \quad \text{with } \mathbf{x}(0) \text{ given} \quad (3.66)$$

We express \mathbf{x} as an integral operation on the whole vector-valued function $\mathbf{B}\mathbf{u}$.

Inversion of the State Equation

The state equation for an n th-order time-invariant dynamic system is

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}_0 \quad (3.67)$$

where \mathbf{A} and \mathbf{B} are arbitrary $n \times n$ and $n \times m$ matrices, respectively. The state vector $\mathbf{x}(t)$ and the input vector $\mathbf{u}(t)$ are in $\mathfrak{R}^{n \times 1}$ and $\mathfrak{R}^{m \times 1}$, respectively. We invert (3.67) by the same approach we used for the n th-order differential system; we invert separately the two component equations

$$\dot{\mathbf{x}}(t) - \mathbf{A}\mathbf{x}(t) = \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) = \boldsymbol{\theta} \quad (3.68)$$

$$\dot{\mathbf{x}}(t) - \mathbf{A}\mathbf{x}(t) = \boldsymbol{\theta}, \quad \mathbf{x}(0) = \mathbf{x}_0 \quad (3.69)$$

Assume the inverse of the “boundary input” system (3.69) is of the form

$$\mathbf{x}(t) = \boldsymbol{\Phi}(t)\mathbf{x}(0) \quad (3.70)$$

where the boundary kernel $\boldsymbol{\Phi}(t)$ is a $n \times n$ matrix commonly referred to as the **state transition** matrix. (It describes the “undriven” transition from the state at “0” to the state at t .) In order that (3.70) be the correct inverse, $\mathbf{x}(t)$ must satisfy (3.69),

$$\dot{\mathbf{x}}(t) - \mathbf{A}\mathbf{x}(t) = \frac{d\boldsymbol{\Phi}(t)}{dt}\mathbf{x}(0) - \mathbf{A}\boldsymbol{\Phi}(t)\mathbf{x}(0) = \boldsymbol{\theta}, \quad \mathbf{x}(0) = \boldsymbol{\Phi}(0)\mathbf{x}(0)$$

for any initial condition vector $\mathbf{x}(0)$. Therefore, the state transition matrix must satisfy

$$\frac{d\boldsymbol{\Phi}(t)}{dt} - \mathbf{A}\boldsymbol{\Phi}(t) = \boldsymbol{\Theta}, \quad \boldsymbol{\Phi}(0) = \mathbf{I} \quad (3.71)$$

Rather than treat the system (3.71) one element at a time, we work with the whole $n \times n$ matrix-valued system. We use the power series method to find the complementary function for the system. Assume

$$\boldsymbol{\Phi}(t) = \mathbf{C}_0 + \mathbf{C}_1 t + \mathbf{C}_2 t^2 + \dots$$

where each \mathbf{C}_i is a constant $n \times n$ matrix. We substitute $\boldsymbol{\Phi}(t)$ into the differential equation of (3.71) and equate the coefficient on each power of t to the zero matrix $\boldsymbol{\Theta}$ to find

$$\mathbf{C}_1 = \mathbf{A}\mathbf{C}_0, \quad \mathbf{C}_2 = \left(\frac{1}{2!}\right)\mathbf{A}^2\mathbf{C}_0, \quad \mathbf{C}_3 = \left(\frac{1}{3!}\right)\mathbf{A}^3\mathbf{C}_0, \quad \dots$$

It follows that \mathbf{C}_0 is arbitrary and

$$\boldsymbol{\Phi}(t) = \left(\mathbf{I} + \mathbf{A}t + \frac{\mathbf{A}^2 t^2}{2!} + \dots \right) \mathbf{C}_0 \stackrel{\Delta}{=} e^{\mathbf{A}t} \mathbf{C}_0 \quad (3.72)$$

We have used the symbol $e^{\mathbf{A}t}$ to represent the sum of the “exponential-looking” matrix series of (3.72):

$$e^{\mathbf{A}t} \triangleq \mathbf{I} + \mathbf{A}t + \frac{\mathbf{A}^2 t^2}{2!} \dots$$

We call $e^{\mathbf{A}t}$ a **fundamental matrix** for the state equation of (3.67); the matrix is analogous to a fundamental set of solutions for an n th-order differential equation. Applying the boundary conditions of (3.71) to (3.72), we find $\Phi(0) = e^{\mathbf{A}0} \mathbf{C}_0 = \mathbf{I}$. It is clear from the definition of $e^{\mathbf{A}t}$ that $e^{\mathbf{A}0} = \mathbf{I}$; therefore, $\mathbf{C}_0 = \mathbf{I}$ and the state transition matrix (or boundary kernel) for the state-space system (3.67) is

$$\Phi(t) = e^{\mathbf{A}t} \quad (3.73)$$

Example 2. A State Transition Matrix. In Example 1 we found the system matrix \mathbf{A} for the differential equation $\ddot{\phi}(t) + \dot{\phi}(t) = \mathbf{u}(t)$:

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

To find the fundamental matrix for this system, we sum the defining infinite series:

$$\begin{aligned} \Phi(t) &= e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \frac{\mathbf{A}^2 t^2}{2!} + \dots \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} t + \begin{pmatrix} 0 & -1 \\ 0 & 1 \end{pmatrix} \frac{t^2}{2!} + \dots \\ &= \begin{pmatrix} 1 & \left(t - \frac{t^2}{2!} + \frac{t^3}{3!} - \dots \right) \\ 0 & \left(1 - t + \frac{t^2}{2!} - \dots \right) \end{pmatrix} = \begin{pmatrix} 1 & 1 - e^{-t} \\ 0 & e^{-t} \end{pmatrix} \end{aligned}$$

If the matrix \mathbf{A} of Example 2 were not simple, it would be difficult to sum the infinite series for $e^{\mathbf{A}t}$ by the method of that example. It would not be easy to recognize the function to which each scalar series converges. Arbitrary functions of matrices are examined in detail in Section 4.6, and practical techniques for computing functions of matrices are developed. These techniques can be used to compute $e^{\mathbf{A}t}$ for an arbitrary square matrix \mathbf{A} .

Exercise 2. Show that for the fundamental matrix $e^{\mathbf{A}t}$ of Example 2,

$$\begin{aligned} e^{\mathbf{A}t}e^{\mathbf{A}\tau} &= e^{\mathbf{A}(t+\tau)} \\ (e^{\mathbf{A}t})^{-1} &= e^{-\mathbf{A}t} \end{aligned} \quad (3.74)$$

Properties (3.74) apply to all time-invariant systems, that is, all systems which have a constant system matrix (P&C 3.19).

We can view \mathbf{Bu} as a vector-valued distributed input to (3.67). Therefore, we assume the inverse of the distributed-input state equation (3.68) is an integral equation of the form

$$\mathbf{x}(t) = \int_0^\infty \mathbf{K}(t,s)\mathbf{Bu}(s)ds \quad (3.75)$$

where the $n \times n$ matrix $\mathbf{K}(t,s)$ is called the **matrix Green's function** for the system (3.67). (By the integral of a matrix we mean the matrix of integrals.) We substitute (3.75) into (3.68) to determine the equations which describe \mathbf{K} :

$$\begin{aligned} \dot{\mathbf{x}}(t) - \mathbf{Ax}(t) &= \int_0^\infty \left[\frac{d}{dt} \mathbf{K}(t,s) - \mathbf{AK}(t,s) \right] \mathbf{Bu}(s) ds = \mathbf{Bu}(t) \\ \mathbf{x}(0) &= \int_0^\infty \mathbf{K}(0,s)\mathbf{Bu}(s) ds = \boldsymbol{\theta} \end{aligned} \quad (3.76)$$

for all vectors \mathbf{u} with elements which are continuous functions. To see more clearly the conditions on $\mathbf{K}(t,s)$ which follow from (3.76), note that

$$\int_0^\infty \delta(t-s)\mathbf{I} \begin{pmatrix} g_{11}(s) & \cdots & g_{1m}(s) \\ \vdots & & \vdots \\ g_{n1}(s) & \cdots & g_{nm}(s) \end{pmatrix} ds = \begin{pmatrix} g_{11}(t) & \cdots & g_{1m}(t) \\ \vdots & & \vdots \\ g_{n1}(t) & \cdots & g_{nm}(t) \end{pmatrix}$$

In other words, if we let $\mathbf{G}(s)$ denote the matrix with elements $g_{ij}(s)$, then the equation $\int_0^\infty \mathbf{K}(t,s)\mathbf{G}(s)ds = \mathbf{G}(t)$ is satisfied by $\mathbf{K}(t,s) = \delta(t-s)\mathbf{I}$. Thus in order to satisfy (3.76), it is sufficient that \mathbf{K} meet the following requirements:

$$\begin{aligned} \frac{d\mathbf{K}(t,s)}{dt} - \mathbf{AK}(t,s) &= \delta(t-s)\mathbf{I} \\ \mathbf{K}(0,s) &= \boldsymbol{\theta} \end{aligned} \quad (3.77)$$

The approach we use to solve (3.77) for \mathbf{K} is essentially the same as that used for the n th-order scalar system (3.55). For $t \neq s$, $\mathbf{K}(t, s)$ satisfies the same $n \times n$ differential equation, (3.71), as does the state transition matrix. Thus, using the general solution to (3.71) found earlier,

$$\begin{aligned}\mathbf{K}(t, s) &= e^{\mathbf{A}t} \mathbf{B}_0, & t \text{ in } [0, s) \\ &= e^{\mathbf{A}t} \mathbf{D}_0, & t \text{ in } (s, \infty)\end{aligned}$$

where \mathbf{B}_0 and \mathbf{D}_0 are $n \times n$ constant matrices. The boundary conditions of (3.76) require $\mathbf{K}(0, s) = e^{\mathbf{A}0} \mathbf{B}_0 = \mathbf{\Theta}$; since $e^{\mathbf{A}0} = \mathbf{I}$, $\mathbf{B}_0 = \mathbf{\Theta}$. From (3.77), we also note that \mathbf{K} must satisfy a discontinuity condition at $t = s$. The delta functions on the right-hand side of (3.77) must be introduced by the highest derivative, $d\mathbf{K}/dt$; otherwise derivatives of delta functions would appear. Consequently, the diagonal elements of \mathbf{K} contain a unit step at $t = s$, whereas off-diagonal elements are continuous:

$$\mathbf{K}(s^+, s) - \mathbf{K}(s^-, s) = e^{\mathbf{A}s} \mathbf{D}_0 - \mathbf{\Theta} = \mathbf{I}$$

Then, using (3.74), $\mathbf{D}_0 = (e^{\mathbf{A}s})^{-1} = e^{-\mathbf{A}s}$, and

$$\begin{aligned}\mathbf{K}(t, s) &= \mathbf{\Theta}, & t < s \\ &= e^{\mathbf{A}(t-s)}, & t > s\end{aligned}\tag{3.78}$$

The inverse of the state-space system (3.67) is the sum of (3.75) and (3.70); Φ and \mathbf{K} are given by (3.73) and (3.78), respectively:

$$\mathbf{x}(t) = \int_0^t e^{\mathbf{A}(t-s)} \mathbf{B} \mathbf{u}(s) ds + e^{\mathbf{A}t} \mathbf{x}(0)\tag{3.79}$$

The inverse system is fully determined by the state transition matrix $e^{\mathbf{A}t}$ and the input matrix \mathbf{B} . In Section 4.6 we determine how to evaluate $e^{\mathbf{A}t}$ by methods other than summing of the series (3.72).

At the heart of the solution (3.60) for the n th-order dynamic system (3.55) is the Vandermonde matrix for the system. If the state equation is derived from the n th-order differential equation as in (3.63), we would expect the Vandermonde matrix to be involved in the solution (3.79) of the state equation. We find in P&C 4.16 and (4.98) that if the system matrix \mathbf{A} is the companion matrix for an n th-order dynamic system, the Vandermonde matrix is intimately related to both \mathbf{A} and $e^{\mathbf{A}t}$.

Exercise 3. Show that for the system of Examples 1 and 2,

$$\mathbf{x}(t) = \begin{pmatrix} \phi(t) \\ \dot{\phi}(t) \end{pmatrix} = \int_0^t \begin{pmatrix} 1 - e^{-(t-s)} \\ e^{-(t-s)} \end{pmatrix} \mathbf{u}(s) ds + \begin{pmatrix} \phi(0) + \dot{\phi}(0) - \dot{\phi}(0)e^{-t} \\ \dot{\phi}(0)e^{-t} \end{pmatrix}\tag{3.80}$$

Equation (3.80) should be compared with its second-order scalar equivalent (3.62). The state-space solution usually contains more information than its scalar counterpart—information about derivatives of the solution is stated explicitly.

Exercise 4. Use the solution (3.79) at $t = a$ to determine the form of the solution to the state-space system (3.67) if the initial conditions are given at $t = a$ instead of $t = 0$; that is, show that

$$\mathbf{f}(t) = \int_a^t e^{\mathbf{A}(t-s)} \mathbf{B}\mathbf{u}(s) ds + e^{\mathbf{A}(t-a)} \mathbf{x}(a)$$

The discussion beneath the n th-order scalar solution (3.60) extends to the more general state-space solution (3.79). We can interpret $\mathbf{K}(t, 0) = e^{\mathbf{A}t}$ as the matrix impulse response of the state-space system. Since the matrix \mathbf{A} is constant, it is appropriate to measure physically the state transition matrix $e^{\mathbf{A}t}$. By (3.70), the j th column of $\Phi(t)$ (or $e^{\mathbf{A}t}$) consists in the “undriven” decay of $\mathbf{x}(t)$ from the initial condition $\mathbf{x}(0) = \boldsymbol{\varepsilon}_j$, the j th standard basis vector for $\mathfrak{R}^n \times 1$. From measurements of the n columns of $e^{\mathbf{A}t}$ we can determine the full inverse equation (3.79) without explicit determination of the system matrix \mathbf{A} (P&C 3.20).

The techniques used to invert the first-order state-space system (3.67) are applied to a *second-order* vector differential system in P&C 4.32. As with the state-space system, the Green’s function for this system can be obtained from the boundary kernel; the latter can be measured physically. The inverse for this second-order vector system involves several functions of matrices. We discuss methods for evaluating general functions of matrices in Section 4.6.

3.5 Problems and Comments

3.1 *Forward integration:* the differential system $\mathbf{f}''(t) + \frac{1}{4}\mathbf{f}(t) + (1/400)\mathbf{f}^3(t) = 0$, $\mathbf{f}(0) = 10$, $\mathbf{f}'(0) = 0$ describes the unforced oscillations of a mass hanging on a spring. The spring has a nonlinear force-elongation characteristic; $\mathbf{f}(t)$ denotes the position of the mass at time t . There are many numerical integration techniques for obtaining an approximate solution to such a nonlinear differential equation with initial conditions (see [3.9]). The following technique is one of the simplest. We concern ourselves only with integer values of t , and replace the derivatives by the finite-difference approximations $\mathbf{f}'(n) \approx \mathbf{f}(n+1) - \mathbf{f}(n)$ and $\mathbf{f}''(n) \approx \mathbf{f}(n+1) - 2\mathbf{f}(n) + \mathbf{f}(n-1)$. Use these finite-difference approximations and the differential system

to express $\mathbf{f}(n+1)$ in terms of $\mathbf{f}(n)$ and $\mathbf{f}(n-1)$. Compute $\mathbf{f}(1), \mathbf{f}(2), \dots, \mathbf{f}(8)$. How might the above finite-difference approximation be modified to obtain a more accurate solution to the differential equation?

- 3.2 *Backward integration:* a (nonlinear) differential equation with final end-point conditions (rather than initial conditions) can be solved by backward numerical integration. Backward integration can be carried out by means of any forward integration routine. Suppose the differential system is of the form $\mathbf{f}^{(n)}(t) + \mathbf{F}(\mathbf{f}(t), \mathbf{f}'(t), \dots, \mathbf{f}^{(n-1)}(t), t) = \mathbf{0}$ with $\mathbf{f}(t_f), \mathbf{f}'(t_f), \dots, \mathbf{f}^{(k-1)}(t_f)$ specified. Show that the change of variables $\mathbf{f}(t) = \mathbf{f}(t_f - s) = \mathbf{g}(s)$ converts the final conditions on \mathbf{f} to initial conditions on \mathbf{g} and produces a differential equation in \mathbf{g} which differs from the differential equation in \mathbf{f} in the sign on the odd-order derivatives.
- 3.3 *Relaxation:* the finite-difference approximation to a two-point boundary value problem can be solved by a simple iterative technique known as relaxation [3.3]. Suppose $\mathbf{f}''(s) = 1$ with $\mathbf{f}(0) = \mathbf{f}(5) = \mathbf{0}$. Consider the values of \mathbf{f} only at integer values of s . Replace the second derivative by the approximation $\mathbf{f}''(n) \approx \mathbf{f}(n+1) - 2\mathbf{f}(n) + \mathbf{f}(n-1)$, and express $\mathbf{f}(n)$ in terms of $\mathbf{f}(n-1)$ and $\mathbf{f}(n+1)$. Let the initial values of $\mathbf{f}(1), \dots, \mathbf{f}(4)$ be zero. A single step in the iteration consists in solving successively for each of the values $\mathbf{f}(1), \dots, \mathbf{f}(4)$ in terms of current values of \mathbf{f} at the two neighboring points. Repetitive improvement of the set of values $\{\mathbf{f}(k)\}$ results in convergence of this set of values to the solution of the set of difference equations, regardless of numerical errors, and regardless of the order in which the values are improved during each iteration.
- Carry out six iterations for the above problem.
 - Find the exact solution to the set of difference equations by solving the equations simultaneously. Compare the results of the iteration of (a) with the exact solution for the differential system.
- *3.4 An intuitive understanding of the following properties of differential systems can be gained by examining a finite-difference approximation to the second-order case. See [3.6] for a rigorous discussion of these statements.
- A regular n th-order linear differential equation has n independent solutions.
 - A boundary condition consisting in a linear combination of values of $\mathbf{f}, \mathbf{f}', \dots, \mathbf{f}^{(n-1)}$ need not be independent of the regular n th-order differential equation; consider, for example, $\mathbf{f}''(s) = 0$ with $\mathbf{f}'(0) - \mathbf{f}'(1) = 0$.

- (c) If the boundary conditions associated with a regular n th-order differential equation consist in n independent linear combinations of the values $\mathbf{f}(\mathbf{a}), \mathbf{f}'(\mathbf{a}), \dots, \mathbf{f}^{(n-1)}(\mathbf{a})$, at a single point \mathbf{a} in the domain of \mathbf{f} , then the differential system has a unique solution.

3.5 The following differential system is degenerate:

$$\phi'' + \phi' = \mathbf{u} \quad \text{with} \quad \begin{cases} \phi(0) - \phi(1) = \alpha_1 \\ 2\phi(1) - 2\phi(0) + \phi'(0) - \phi'(1) = \alpha_2 \\ \phi'(1) - \phi'(0) = \alpha_3 \end{cases}$$

Find the solutions to the differential system in terms of the inputs \mathbf{u} , α_1 , α_2 , and α_3 . Also find the relations among the inputs that must be satisfied in order that solutions exist. (Hint: the solution to the differential equation is expressed in terms of $\phi(0)$ and $\phi'(0)$ in (3.80).) What relationship exists between the number of dependent rows in a boundary condition matrix for a system and the number of different relations which must be satisfied by the inputs to that system?

3.6 Let \mathbf{L} be a regular n th-order differential operator and $\{\beta_i(\mathbf{f}) = 0, i=1, \dots, m\}$ a set of homogeneous boundary conditions. Let \mathcal{V} be the space of functions in $\mathcal{C}^n(a, b)$ which satisfy the homogeneous differential equation $\mathbf{L}\mathbf{f} = \theta$. Let $\mathcal{F} \stackrel{\Delta}{=} \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a fundamental set of solutions for \mathbf{L} ; \mathcal{F} is a basis for \mathcal{V} . Define $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{R}^m$ by $\mathbf{T}\mathbf{f} \stackrel{\Delta}{=} (\beta_1(\mathbf{f}), \dots, \beta_m(\mathbf{f}))$ for all \mathbf{f} in \mathcal{V} . Let \mathcal{E} be the standard basis for \mathcal{R}^m . Show that the matrix $[\mathbf{T}]_{\mathcal{F}\mathcal{E}}$ is a boundary condition matrix for the differential system $\{\mathbf{L}, \beta_1, \dots, \beta_m\}$.

*3.7 The Wronskian: let $\mathbf{f}_1, \dots, \mathbf{f}_n$ be in $\mathcal{C}^n(a, b)$. The Wronskian matrix of $\mathbf{f}_1, \dots, \mathbf{f}_n$ at t is defined by

$$\mathbf{W}(t) \stackrel{\Delta}{=} \begin{pmatrix} \mathbf{f}_1(t) & \cdots & \mathbf{f}_n(t) \\ \mathbf{f}_1^{(1)}(t) & \cdots & \mathbf{f}_n^{(1)}(t) \\ \vdots & & \vdots \\ \mathbf{f}_1^{(n-1)}(t) & \cdots & \mathbf{f}_n^{(n-1)}(t) \end{pmatrix}$$

The Wronskian determinant is $w(t) \stackrel{\Delta}{=} \det(\mathbf{W}(t))$.

(a) Show that $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ cannot be linearly dependent unless $w(t) = 0$ for all t in $[a, b]$.

- (b) The fact that $\mathbf{w}(t) = 0$ for *some* t does not ordinarily imply that the set $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ is dependent; try, for example, $\mathbf{f}_1(t) = t^2$ and $\mathbf{f}_2(t) = t^3$ at $t = 0$. Suppose, however, that $\mathbf{f}_1, \dots, \mathbf{f}_n$ are solutions to an n th-order homogeneous differential equation defined on $[a, b]$. Then if $\mathbf{w}(t) = 0$ for *any* t in $[a, b]$, $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ is a linearly dependent set.

3.8 *Difference equations*: an arbitrary linear constant-coefficient difference equation can be expressed in the form

$$a_0(\mathbf{E}^n \mathbf{f})(k) + a_1(\mathbf{E}^{n-1} \mathbf{f})(k) + \dots + a_n \mathbf{f}(k) = \mathbf{u}(k), \quad k=0, 1, 2, \dots$$

where \mathbf{E} is the shift operator defined by $(\mathbf{E}\mathbf{f})(k) \triangleq \mathbf{f}(k+1)$; we concern ourselves only with integer values of the argument of \mathbf{f} . The order of the difference equation is the number of boundary conditions needed to specify a unique solution to the equation; that is, the order is $n-p$, where p is the lowest power of \mathbf{E} to appear in the equation. (See [3.2].)

- (a) The solutions to the homogeneous difference equation (the equation with $\mathbf{u}(k) = 0$) usually consist of combinations of geometric sequences. Substitution of the sequence $\mathbf{f}(k) = r^k$, $k=0, 1, 2, \dots$, into the homogeneous equation shows that non-trivial sequences must satisfy the following characteristic equation: $a_0 r^n + a_1 r^{n-1} + \dots + a_n = 0$. Find a basis for the nullspace of the difference operator \mathbf{T} defined by

$$\begin{aligned} (\mathbf{T}\mathbf{f})(k) &\triangleq 2(\mathbf{E}^2 \mathbf{f})(k) - 3(\mathbf{E}\mathbf{f})(k) + \mathbf{f}(k) \\ &= 2\mathbf{f}(k+2) - 3\mathbf{f}(k+1) + \mathbf{f}(k) \end{aligned}$$

What is the dimension of the nullspace of an n th-order difference operator?

- (b) Let $\mathbf{f}_1, \dots, \mathbf{f}_n$ be infinite sequences of the form $\mathbf{f}_i(k)$, $k=0, 1, 2, \dots$. The *Casorati matrix* of $\mathbf{f}_1, \dots, \mathbf{f}_n$ is defined by

$$\mathbf{C}(k) \triangleq \begin{pmatrix} \mathbf{f}_1(k) & \dots & \mathbf{f}_n(k) \\ \mathbf{E}\mathbf{f}_1(k) & \dots & \mathbf{E}\mathbf{f}_n(k) \\ \vdots & & \vdots \\ \mathbf{E}^{n-1}\mathbf{f}_1(k) & \dots & \mathbf{E}^{n-1}\mathbf{f}_n(k) \end{pmatrix}$$

The infinite sequences $\mathbf{f}_1, \dots, \mathbf{f}_n$ are linearly independent if and

only if $c(k) \neq 0$ for $k = 0, 1, 2, \dots$, where $c(k)$ is the Casorati determinant, $\det(\mathbf{C}(k))$. Use the Casorati determinant to show the independence of the basis vectors found in (a).

- 3.9 Use the power series method to find the complementary function for the differential operator $(\mathbf{D} - 1)^2$.
- 3.10 Define $\mathbf{L}: \mathcal{C}^1(0, 1) \rightarrow \mathcal{C}(0, 1)$ by $\mathbf{L} \stackrel{\Delta}{=} -\mathbf{D} - a\mathbf{I}$. Find the Green's function k and the inverse equation for the differential system $\mathbf{L}\mathbf{f} = \mathbf{u}$, $\mathbf{f}(0) = \mathbf{f}(1)$.
- 3.11 Define $\mathbf{L}: \mathcal{C}^2(0, b) \rightarrow \mathcal{C}(0, b)$ by $\mathbf{L} \stackrel{\Delta}{=} \mathbf{D}^2 - 3\mathbf{D} + 2\mathbf{I}$. Find the Green's function k , the boundary kernel ρ , and the inverse equation for the differential system $\mathbf{L}\mathbf{f} = \mathbf{u}$, $\mathbf{f}(0) = \alpha_1$, $\mathbf{f}(b) = \alpha_2$.
- 3.12 Find the inverse equation for each of the following differential systems:
- (a) $\mathbf{f}'' + 6\mathbf{f}' + 5\mathbf{f} = \mathbf{u}$, $\mathbf{f}(0) = \alpha_1$, $\mathbf{f}'(0) = \alpha_2$
- (b) $\mathbf{f}'' + 2\mathbf{f}' + 2\mathbf{f} = \mathbf{u}$, $\mathbf{f}(0) = \alpha_1$, $\mathbf{f}'(0) = \alpha_2$
- (c) $\mathbf{f}''' + 6\mathbf{f}'' + 5\mathbf{f}' = \mathbf{u}$, $\mathbf{f}(0) = \alpha_1$, $\mathbf{f}'(0) = \alpha_2$, $\mathbf{f}''(0) = \alpha_3$
- 3.13 The following differential system describes the steady-state temperature distribution along an insulated bar of length b : $-\mathbf{f}'' = \mathbf{u}$, $\mathbf{f}(0) = \alpha_1$, $\mathbf{f}'(b) + \mathbf{f}(b) = \alpha_2$. (The second boundary condition implies that heat is removed by convection at point b .) Show that the inverse equation for this system is

$$\mathbf{f}(t) = \left(1 - \frac{t}{1+b}\right) \int_0^t s\mathbf{u}(s) ds + t \int_t^b \left(1 - \frac{s}{1+b}\right) \mathbf{u}(s) ds + \alpha_1 \left(1 - \frac{t}{1+b}\right) + \alpha_2 \left(\frac{t}{1+b}\right)$$

- 3.14 For the differential system $t\mathbf{f}'(t) - \mathbf{f}(t) = \mathbf{u}(t)$, $\mathbf{f}'(t_1) = \alpha$, $t_1 > 0$,
- (a) Find the complementary function by the power series method;
- (b) Find the Green's function $k(t, s)$;
- (c) Find the boundary kernel $\rho_j(t)$;
- (d) State explicitly the inverse equation.
- *3.15 Let μ_1 and μ_2 be the roots of the characteristic equation for the differential system $\mathbf{f}'' + a_1\mathbf{f}' + a_2\mathbf{f} = \mathbf{u}$, $\mathbf{f}(0) = \mathbf{f}'(0) = 0$.
- (a) Use (3.56) and (3.57) to find the Green's function k for this system. If $\mu_2 \approx \mu_1$, computed values of $\mu_2 - \mu_1$ and $\exp(\mu_2 t) - \exp(\mu_1 t)$ will be badly in error. What is the effect of near equality of the roots on the numerical computation of $k(t, s)$ and $\int k(t, s)\mathbf{u}(s) ds$?

- (b) If $\mu_2 \approx \mu_1$, the fundamental set $\{\exp(\mu_1 t), \exp(\mu_2 t)\}$ is nearly dependent. A better fundamental set (not nearly dependent) in this circumstance is

$$\mathbf{v}_1(t) = \frac{\exp(\mu_1 t) + \exp(\mu_2 t)}{2} \quad \mathbf{v}_2(t) = \frac{\exp(\mu_1 t) - \exp(\mu_2 t)}{\mu_1 - \mu_2}$$

Derive a power series expansion of \mathbf{v}_2 which can be used to compute values of \mathbf{v}_2 without numerical division by the inaccurate quantity $\mu_2 - \mu_1$. Show that as $\mu_2 \rightarrow \mu_1$, $\{\mathbf{v}_1(t), \mathbf{v}_2(t)\} \rightarrow \{\exp(\mu_1 t), t \exp(\mu_1 t)\}$.

- (c) Equation (3.52) expresses the Green's function k in terms of the functions $\{\mathbf{v}_i\}$ of (b). Evaluate the Wronskian determinant w in this expression in terms of exponentials. Values of $k(t, s)$ and $\int k(t, s) \mathbf{u}(s) ds$ can be computed accurately by using this expression for $k(t, s)$ together with computed values of $\mathbf{v}_1, \mathbf{v}_2$, and w . Show that this expression for $k(t, s)$ is a rearrangement of the expression for $k(t, s)$ found in (a).

- 3.16 One method for obtaining the Green's function for a constant-coefficient differential system is to solve (3.32) by means of one-sided Laplace transforms. Use this technique to show that the inverse of the differential equation $\ddot{\mathbf{f}} + \omega^2 \mathbf{f} = \mathbf{u}$, with constant ω and given values of $\mathbf{f}(0)$ and $\dot{\mathbf{f}}(0)$, is

$$\mathbf{f}(t) = \mathbf{f}(0) \cos \omega t + \frac{\dot{\mathbf{f}}(0)}{\omega} \sin \omega t + \frac{1}{\omega} \int_0^t \sin \omega(t-s) \mathbf{u}(s) ds$$

- 3.17 The approximation of derivatives by finite-differences leads to the approximate representation of differential equations by difference equations. For instance, the use of a second-central difference plus a forward difference converts the second-order differential system $\phi'' + \phi' = \mathbf{u}$, $\phi(0) = \alpha_1$, $\phi'(0) = \alpha_2$ to the approximately equivalent second-order difference system $2\phi(i+2) - 3\phi(i+1) + \phi(i) = \mathbf{u}(i+1)$, $\phi(0) = \alpha_1$, $\phi(1) = \alpha_2 + \alpha_1$. A general form for the n th-order constant-coefficient difference system with initial conditions is

$$\begin{aligned} \mathbf{f}(i+n) + a_1 \mathbf{f}(i+n-1) + \cdots + a_n \mathbf{f}(i) &= \mathbf{v}(i) \\ \mathbf{f}(0) = \gamma_1, \quad \mathbf{f}(1) = \gamma_2 \quad \dots, \quad \mathbf{f}(n-1) &= \gamma_n \end{aligned}$$

for $i=0, 1, 2, \dots$.

By analogy to the inverse equation for the n th-order differential system, we assume the inverse of the n th-order difference system is

of the form

$$\mathbf{f}(i) = \sum_{j=0}^{\infty} k(i,j)\mathbf{v}(j) + \sum_{m=1}^n \rho_m(i)\mathbf{f}(m-1)$$

for $i=0, 1, 2, \dots$

(a) Show that the discrete Green's function $k(i,j)$ is specified by the difference system

$$k(i+n,j) + a_1 k(i+n-1,j) + \dots + a_n k(i,j) = \delta_{ij}$$

$$k(0,j) = k(1,j) = \dots = k(n-1,j) = 0$$

for $i=0, 1, 2, \dots$ and $j=0, 1, 2, \dots$

(b) Show that the discrete boundary kernel $\rho_m(i)$ is specified by the difference system

$$\rho_m(i+n) + a_1 \rho_m(i+n-1) + \dots + a_n \rho_m(i) = 0$$

$$\rho_m(p) = \delta_{m,p+1}$$

for $i=0, 1, 2, \dots, m=1, \dots, n$, and $\dots, n-1$.

(c) Find the inverse of the second-order difference system mentioned above by solving the difference systems corresponding to those in (a) and (b). Hint: solutions to homogeneous constant-coefficient difference equations consist in sums of geometric sequences of the form $\mathbf{f}(i) = r^i$, $i = 0, \pm 1, \pm 2, \dots$

3.18 The following pair of coupled differential equations relates a pair of system outputs $\{\mathbf{f}_i(t)\}$ to a pair of inputs $\{\mathbf{u}_i(t)\}$:

$$\mathbf{f}_1'' + 3\mathbf{f}_1' + 2\mathbf{f}_2 = \mathbf{u}_1,$$

$$\mathbf{f}_2'' + \mathbf{f}_1' + \mathbf{f}_2 = \mathbf{u}_2,$$

$$\mathbf{f}_1(0), \mathbf{f}_1'(0), \mathbf{f}_2(0), \mathbf{f}_2'(0) \text{ specified.}$$

(a) Find a first-order state equation of the form (3.65) which is equivalent to the set of coupled equations. (Hint: use as state variables the output functions and their first derivatives.) Is the state equation unique?

(b) The solution to the state equation is determined by the state transition matrix (3.73). How could this matrix function be computed for the system in (a)?

3.19 *Properties of state transition matrices:* the concept of a state transition matrix extends to time-varying dynamic systems [3.14]. sup-

pose a dynamic system satisfies $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t)$, where $\mathbf{x}(t_0)$ is given and $\mathbf{A}(t)$ is an $n \times n$ matrix. We can express the solution in the form $\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}(t_0)$. We refer to the $n \times n$ matrix $\Phi(t_0, t)$ as the **state transition matrix**. The state transition matrix has the following properties:

- (a) $\frac{d}{dt}\Phi(t, t_0) = \mathbf{A}(t)\Phi(t, t_0)$, $\Phi(t_0, t_0) = I$;
- (b) $\Phi(t_0, t_1)\Phi(t_1, t_2) = \Phi(t_0, t_2)$ for all t_0, t_1 , and t_2 ;
- (c) $\Phi(t_1, t_0)^{-1} = \Phi(t_0, t_1)$;
- (d) If $\mathbf{A}(t)\int_{t_0}^t \mathbf{A}(s)ds = \int_{t_0}^t \mathbf{A}(s)ds \mathbf{A}(t)$, then $\Phi(t, t_0) = \exp \int_{t_0}^t \mathbf{A}(s)ds$ (see P&C 4.29);
- (e) $\det \Phi(t, t_0) = \exp \int_{t_0}^t \text{trace}[\mathbf{A}(s)]ds$, where $\text{trace}[\mathbf{A}(s)]$ is the sum of the diagonal elements of $\mathbf{A}(s)$.

3.20 A certain system can be represented by a differential equation of the form $\ddot{\mathbf{f}} + a_1\dot{\mathbf{f}} + a_2\mathbf{f} = \mathbf{u}$. The values of the coefficients a_1 and a_2 are unknown. However, we have observed the response of the undriven system ($\mathbf{u}(t) = 0$ for $t > 0$) with various initial conditions. In particular, for $\mathbf{f}(0) = 1$ and $\dot{\mathbf{f}}(0) = 0$, we find that $\mathbf{f}(t) = 2e^{-t} - e^{-2t}$ and $\dot{\mathbf{f}}(t) = 2(e^{-2t} - e^{-t})$ for $t \geq 0$. Also, for $\mathbf{f}(0) = 0$ and $\dot{\mathbf{f}}(0) = 1$, we find that $\mathbf{f}(t) = e^{-t} - e^{-2t}$ and $\dot{\mathbf{f}}(t) = 2e^{-2t} - e^{-t}$ for $t \geq 0$.

- (a) Determine the state equation in terms of a_1 and a_2 .
- (b) Use the transient measurements to determine the state transition matrix and the precise inverse of the state equation.

3.21 *Discrete-time state equations*: by using finite-difference approximations for derivatives, an arbitrary n th-order linear constant-coefficient differential equation with initial conditions can be approximated by an n th-order linear constant-coefficient difference equation of the form

$$\mathbf{f}((k+n)\tau) + a_1\mathbf{f}((k+n-1)\tau) + \cdots + a_n\mathbf{f}(k\tau) = \mathbf{u}(k\tau)$$

for $k = 0, 1, 2, \dots$, with $\mathbf{f}(0), \mathbf{f}(\tau), \dots, \mathbf{f}((n-1)\tau)$ given. The quantity τ is the time increment used in the finite-difference approximation.

- (a) Put this n th-order difference equation in state-space form; that is, develop an equivalent first-order vector difference equation.
- (b) Determine the form of the inverse of the discrete-time state equation.

3.6 References

- [3.1] Bergman, Stefan and M. Schiffer, *Kernel Functions and Elliptic Differential Equations in Mathematical Physics*, Academic Press, New York, 1953.

- [3.2] DeRusso, Paul M., Rob J. Roy, and Charles M. Close, *State Variables for Engineers*, Wiley, New York, 1966.
- [3.3] Forsythe, George E. and Wolfgang R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, Wiley, New York, 1960.
- *[3.4] Friedman, Bernard, *Principles and Techniques of Applied Mathematics*, Wiley, New York, 1966.
- [3.5] Greenberg, Michael D., *Applications of Green's Functions in Science and Engineering*, Prentice-Hall, Englewood Cliffs, N.J., 1971.
- [3.6] Ince, E. L., *Ordinary Differential Equations*, Dover, New York, 1956.
- [3.7] Kaplan, Wilfred, *Advanced Calculus*, Addison-Wesley, Reading, Mass., 1959.
- [3.8] Morse, Philip M. and Herman Feshbach, *Methods of Theoretical Physics*, Parts I and II, McGraw-Hill, New York, 1953.
- [3.9] Ralston, Anthony, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1965.
- [3.10] Schwartz, L., *Théorie des Distributions*, Vols. 1 and 2, Hermann & Cie, Paris, 1951, 1957.
- *[3.11] Stakgold, Ivar, *Boundary Value Problems of Mathematical Physics*, Volume I, Macmillan, New York, 1968.
- [3.12] Varga, Richard S., *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.
- [3.13] Wylie, C. R., Jr., *Advanced Engineering Mathematics*, 3rd ed., McGraw-Hill, New York, 1966.
- [3.14] Zadeh, Lotfi A. and Charles A. Desoer, *Linear System Theory*, McGraw-Hill, New York, 1963.

Spectral Analysis of Linear Systems

In this chapter the central theme is the decomposition of the abstract linear equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ into sets of simple linear equations which can be solved independently. Our initial purpose for exploring this decomposition is to obtain conceptual simplification of the system model. It is easier to think about the behavior of one scalar variable at a time than to think about the behavior of a vector variable. Furthermore, the solutions to the decomposed pieces of the original equation usually have physical meanings which provide insight into the behavior of the system. (See for example, P&C 4.7 or the discussion of the analysis of three-phase power systems by the method of symmetrical components.)

There are also computational reasons for examining the decomposition process. Generally speaking, decomposition provides an alternative to inversion as a technique for solving or analyzing the equations which describe a system. In particular, decomposition provides a practical technique for computing solutions to linear differential equations with arbitrary inputs (Section 5.5). In some instances decomposition provides both solutions and insight at no additional computational expense as compared to inversion. (Again, see the discussion of symmetrical components mentioned above.)

The ability to combine the solutions to small subproblems into a solution for the full system equation depends on the principle of linearity. Consequently, we restrict ourselves to linear models in this chapter in order to be able to fully develop the decomposition principle. We find that we can decompose most linear systems into sets of simple scalar multiplications. We refer to such “completely decomposable” systems as “diagonalizable” systems. A few systems are not diagonalizable or are so nearly nondiagonalizable that we cannot accurately compute fully decomposed solutions. We still split them into as small pieces as possible. Nondiagonalizable finite-dimensional systems are discussed in Sections 4.4 and 4.5. In Section 4.6 we explore the concept of functions of matrices for

both the diagonalizable and nondiagonalizable cases. We encountered several such matrix functions in Chapter 3; we find the need for others in later chapters. The discussion of diagonalization of infinite-dimensional systems and of functions of linear operators on infinite dimensional spaces is begun in Section 4.6, but is not completed until Section 5.5.

4.1 System Decomposition

In this section we explore the subdivision of the system equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ into a set of “smaller” equations which can be solved independently. Our ability to subdivide a linear equation in this manner is based partly on the fact that the effect of a linear transformation \mathbf{T} on a basis determines the effect of \mathbf{T} on all vectors in the space. In finding the matrix of a transformation, for instance, we simplified the process of determining the matrix elements by examining the effect of the transformation on the basis vectors. Consequently, we begin our investigation of decomposition by subdividing the vector space on which the transformation \mathbf{T} acts. We can think of the space as a sum of smaller subspaces.

Definition. Let \mathcal{W}_1 and \mathcal{W}_2 be subspaces of the vector space \mathcal{V} . We call \mathcal{V} the **direct sum** $\mathcal{W}_1 \oplus \mathcal{W}_2$ of \mathcal{W}_1 and \mathcal{W}_2 if*

- (a) $\mathcal{V} = \mathcal{W}_1 + \mathcal{W}_2$ (\mathcal{W}_1 and \mathcal{W}_2 **span** \mathcal{V}) and
- (b) $\mathcal{W}_1 \cap \mathcal{W}_2 = \theta$ (\mathcal{W}_1 and \mathcal{W}_2 are **linearly independent**)

Example 1. Direct Sum in Arrow Space. The two-dimensional arrow space is the direct sum of two different lines which intersect at the origin (Figure 4.1). If the two lines are identical, they are not independent and do not span the arrow space.

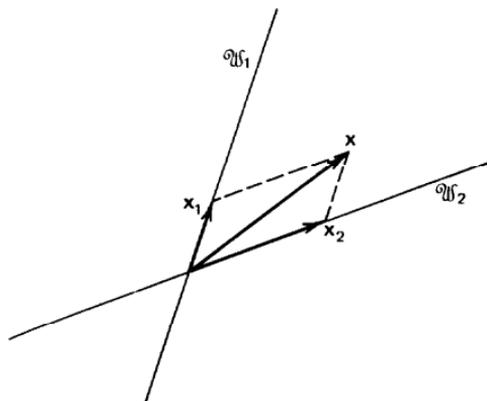


Figure 4.1. Direct sum in arrow space.

*See P&C 2.1 for definitions of the sum and intersection of subspaces.

This arrow space is also the sum of *three* lines which intersect at the origin. However, that sum is not direct; only two of the lines can be independent.

It is apparent from Figure 4.1 that for any finite-dimensional space every splitting of a basis into two parts determines a direct sum; that is, if $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a basis for \mathcal{V} , $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_n\} = \text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_k\} \oplus \text{span}\{\mathbf{x}_{k+1}, \dots, \mathbf{x}_n\}$. It is apparent that the two subspaces can also be subdivided. Although we have not yet defined a basis for an infinite-dimensional space, the concept of splitting a basis applies as well to direct sums in infinite-dimensional spaces (Sections 5.3-5.5).

Example 2. Direct Sum in a Function Space. Let $\mathcal{C}(-1, 1)$ be the space of continuous functions defined on $[-1, 1]$. Let \mathcal{W}_1 be the even functions in $\mathcal{C}(-1, 1)$; $\mathbf{f}_e(-t) = \mathbf{f}_e(t)$. Let \mathcal{W}_2 be the odd functions in $\mathcal{C}(-1, 1)$; $\mathbf{f}_o(-t) = -\mathbf{f}_o(t)$. Any function \mathbf{f} in $\mathcal{C}(-1, 1)$ decomposes into even and odd components:

$$\mathbf{f}(t) = \frac{\mathbf{f}(t) + \mathbf{f}(-t)}{2} + \frac{\mathbf{f}(t) - \mathbf{f}(-t)}{2}$$

Thus \mathcal{W}_1 and \mathcal{W}_2 span $\mathcal{C}(-1, 1)$. The even and odd components of \mathbf{f} are unique; for if \mathbf{f}_e and \mathbf{f}_o are even and odd functions, respectively, such that $\mathbf{f} = \mathbf{f}_e + \mathbf{f}_o$, then

$$\frac{\mathbf{f}(t) + \mathbf{f}(-t)}{2} = \frac{[\mathbf{f}_e(t) + \mathbf{f}_o(t)] + [\mathbf{f}_e(-t) + \mathbf{f}_o(-t)]}{2} = \mathbf{f}_e(t)$$

$$\frac{\mathbf{f}(t) - \mathbf{f}(-t)}{2} = \frac{[\mathbf{f}_e(t) + \mathbf{f}_o(t)] - [\mathbf{f}_e(-t) + \mathbf{f}_o(-t)]}{2} = \mathbf{f}_o(t)$$

Only the zero function is both even and odd; therefore, $\mathcal{W}_1 \cap \mathcal{W}_2 = \mathbf{0}$, and $\mathcal{C}(-1, 1) = \mathcal{W}_1 \oplus \mathcal{W}_2$.

Example 2 demonstrates an important property of the direct sum. Using bases for \mathcal{W}_1 and \mathcal{W}_2 , it is easily shown that $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$ if and only if each \mathbf{x} in \mathcal{V} decomposes *uniquely* into a sum, $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, with \mathbf{x}_1 in \mathcal{W}_1 and \mathbf{x}_2 in \mathcal{W}_2 .

It is a small step to extend the direct sum concept to several subspaces. We merely redefine independence of subspaces: $\mathcal{W}_1, \dots, \mathcal{W}_p$ are linearly independent if each subspace is disjoint from the sum of the rest,

$$\mathcal{W}_i \cap \left(\sum_{j \neq i} \mathcal{W}_j \right) = \mathbf{0} \quad (4.1)$$

With the modification (4.1) we say \mathcal{V} is the direct sum of $\{\mathcal{W}_i\}$ if the subspaces $\{\mathcal{W}_i\}$ are linearly independent and span \mathcal{V} . We denote the direct sum by

$$\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \dots \oplus \mathcal{W}_p \quad (4.2)$$

The previous comments concerning splitting of bases and unique decomposition of vectors also extend to the direct sum of several subspaces.

Exercise 1. Demonstrate in the two-dimensional arrow space that pairwise disjointness is not sufficient to guarantee independence of $\mathcal{W}_1, \dots, \mathcal{W}_p$.

Example 3. Direct Sum of Three Subspaces. Let $\mathbf{f}_1(t) = 1 + t$, $\mathbf{f}_2(t) = t + t^2$, and $\mathbf{f}_3(t) = 1 + t^2$ be a basis for \mathcal{P}^3 . Define $\mathcal{W}_i = \text{span}\{\mathbf{f}_i\}$, $i = 1, 2, 3$. Then $\mathcal{P}^3 = \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3$. Let $\mathbf{f}(t) \triangleq \eta_1 + \eta_2 t + \eta_3 t^2$ be a specific vector in \mathcal{P}^3 . By the process of determining coordinates of \mathbf{f} relative to the basis $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ for \mathcal{P}^3 , we decompose \mathbf{f} uniquely into

$$\mathbf{f} = \left(\frac{\eta_1 + \eta_2 - \eta_3}{2} \right) \mathbf{f}_1 + \left(\frac{-\eta_1 + \eta_2 + \eta_3}{2} \right) \mathbf{f}_2 + \left(\frac{\eta_1 - \eta_2 + \eta_3}{2} \right) \mathbf{f}_3,$$

a sum of vectors from \mathcal{W}_1 , \mathcal{W}_2 , and \mathcal{W}_3 , respectively.

Projection Operators

We can express the direct-sum decomposition of a space in terms of linear operators on the space. Suppose $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$; any vector \mathbf{x} in \mathcal{V} can be written uniquely as $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ with \mathbf{x}_i in \mathcal{W}_i . We define the **projector** (or **projection operator**) \mathbf{P}_1 on \mathcal{W}_1 along \mathcal{W}_2 by $\mathbf{P}_1 \mathbf{x} \triangleq \mathbf{x}_1$ (see Figure 4.1). We call the vector \mathbf{x}_1 the **projection of \mathbf{x} on \mathcal{W}_1 along \mathcal{W}_2** . Similarly $\mathbf{P}_2 \mathbf{x} \triangleq \mathbf{x}_2$ defines the projector on \mathcal{W}_2 along \mathcal{W}_1 .

Example 4. Projector on \mathcal{P}^3 . Let \mathbf{f}_1 , \mathbf{f}_2 , and \mathbf{f}_3 be the functions defined in Example 3. Redefine $\mathcal{W}_1 \triangleq \text{span}\{\mathbf{f}_1\}$ and $\mathcal{W}_2 \triangleq \text{span}\{\mathbf{f}_2, \mathbf{f}_3\}$. Then $\mathcal{P}^3 = \mathcal{W}_1 \oplus \mathcal{W}_2$. In Example 3, the general vector $\mathbf{f}(t) = \eta_1 + \eta_2 t + \eta_3 t^2$ in \mathcal{P}^3 is decomposed into a linear combination of \mathbf{f}_1 , \mathbf{f}_2 , and \mathbf{f}_3 . From that decomposition we see that the projections of \mathbf{f} on \mathcal{W}_1 and \mathcal{W}_2 , respectively, are

$$\begin{aligned} \mathbf{P}_1 \mathbf{f} &= \left(\frac{\eta_1 + \eta_2 - \eta_3}{2} \right) \mathbf{f}_1 \\ \mathbf{P}_2 \mathbf{f} &= \left(\frac{-\eta_1 + \eta_2 + \eta_3}{2} \right) \mathbf{f}_2 + \left(\frac{\eta_1 - \eta_2 + \eta_3}{2} \right) \mathbf{f}_3 \end{aligned}$$

The bases for \mathcal{W}_1 and \mathcal{W}_2 combine to provide a basis which is particularly appropriate for matrix representation of the projectors. Using (2.48), the matrix of the projector \mathbf{P}_1 relative to the basis $\mathcal{F} \triangleq \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ is

$$\begin{aligned} [\mathbf{P}_1]_{\mathcal{F}\mathcal{F}} &= ([\mathbf{P}_1 \mathbf{f}_1]_{\mathcal{F}} \ : \ [\mathbf{P}_1 \mathbf{f}_2]_{\mathcal{F}} \ : \ [\mathbf{P}_1 \mathbf{f}_3]_{\mathcal{F}}) \\ &= ([\mathbf{f}_1]_{\mathcal{F}} \ : \ [\mathbf{0}]_{\mathcal{F}} \ : \ [\mathbf{0}]_{\mathcal{F}}) \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

Similarly, the matrix of \mathbf{P}_2 with respect to \mathfrak{F} is

$$[\mathbf{P}_2]_{\mathfrak{F}\mathfrak{F}} = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

Example 4 emphasizes the fact that a projector acts like the identity operator on its “own” subspace, the one *onto* which it projects, but like the zero operator on the subspace *along* which it projects. The following properties of projectors can be derived from the definition and verified by the matrices of Example 4. Assume $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$. Let \mathbf{P}_i be the projector on \mathcal{W}_i along \mathcal{W}_j ($j \neq i$), and $\mathbf{x}_i = \mathbf{P}_i \mathbf{x}$. Then

$$\begin{aligned} (a) \quad & \mathbf{P}_i \text{ is linear} \\ (b) \quad & \mathbf{P}_i^2 = \mathbf{P}_i \text{ (i.e., } \mathbf{P}_i \mathbf{x}_i = \mathbf{x}_i) \\ (c) \quad & \mathbf{P}_i \mathbf{P}_j = \mathbf{0} \text{ (i.e., } \mathbf{P}_i \mathbf{x}_j = \mathbf{0} \text{ for } j \neq i) \\ (d) \quad & \text{range } (\mathbf{P}_i) = \mathcal{W}_i \\ (e) \quad & \sum_i \mathbf{P}_i = \mathbf{I} \text{ (i.e., } \sum_i \mathbf{P}_i \mathbf{x} = \mathbf{x}) \end{aligned} \tag{4.3}$$

If $\mathcal{V} = \mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_k$, we can define the projector \mathbf{P}_i on \mathcal{W}_i along $\sum_{j \neq i} \mathcal{W}_j$, for $i = 1, \dots, k$. The properties (4.3) apply to this set of projectors as well.

Reduced Operators

The projectors in Example 4 act like scalar multiplication on certain vectors in \mathcal{V} ; \mathbf{P}_i acts like multiplication by 1 on all vectors in the subspace \mathcal{W}_i , and like multiplication by zero on \mathcal{W}_j , $j \neq i$. Other operators also act in a simple manner on certain subspaces. Define the nonlinear operator $\mathbf{G}: \mathcal{R}^2 \rightarrow \mathcal{R}^2$ by

$$\mathbf{G}(\xi_1, \xi_2) \triangleq ((\xi_1 - \xi_2)^2 + 2\xi_2, 2\xi_2)$$

On the subspace $\mathcal{W}_1 \triangleq \text{span}\{(1,0)\}$, \mathbf{G} acts like the simple “squaring” operation, $\mathbf{G}(a,0) = (a^2,0)$. On the subspace $\mathcal{W}_2 \triangleq \text{span}\{(1,1)\}$, \mathbf{G} acts like the “doubling” operation $\mathbf{G}(b,b) = (2b,2b)$. In point of fact, as far as vectors in \mathcal{W}_1 and \mathcal{W}_2 are concerned we can replace \mathbf{G} by the “simpler” operators $\mathbf{G}_1: \mathcal{W}_1 \rightarrow \mathcal{W}_1$ defined by $\mathbf{G}_1(\xi, 0) \triangleq (\xi^2, 0)$ and $\mathbf{G}_2: \mathcal{W}_2 \rightarrow \mathcal{W}_2$ defined by $\mathbf{G}_2(\xi, \xi) \triangleq 2(\xi, \xi)$. We are able to reduce \mathbf{G} to these simpler operators because the action of \mathbf{G} on \mathcal{W}_1 produces only vectors in \mathcal{W}_1 and the action of \mathbf{G} on \mathcal{W}_2 produces only vectors in \mathcal{W}_2 .

Definition. Let \mathbf{G} be an operator (perhaps nonlinear) on \mathcal{V} . The subspace \mathcal{W} (of \mathcal{V}) is **invariant under \mathbf{G}** if for each \mathbf{x} in \mathcal{W} , $\mathbf{G}\mathbf{x}$ is also in \mathcal{W} ; that is, if $\mathbf{G}(\mathcal{W})$ is contained in \mathcal{W} .

Example 5. Invariance of the Nullspace and Range. Let $\mathbf{G}: \mathcal{V} \rightarrow \mathcal{V}$. Then $\text{range}(\mathbf{G})$ is invariant under \mathbf{G} , for \mathbf{G} takes all vectors in \mathcal{V} , including those in $\text{range}(\mathbf{G})$, into $\text{range}(\mathbf{G})$. By definition, \mathbf{G} takes $\text{nullspace}(\mathbf{G})$ into $\mathbf{0}$. If $\mathbf{G}(\mathbf{0}) = \mathbf{0}$, then $\mathbf{0}$ is in $\text{nullspace}(\mathbf{G})$. In this case, $\text{nullspace}(\mathbf{G})$ is also invariant under \mathbf{G} . These subspaces are pictured abstractly in Figure 2.6.

If $\mathbf{G}: \mathcal{V} \rightarrow \mathcal{V}$, and \mathcal{W} is a subspace of \mathcal{V} which is invariant under \mathbf{G} , then we can define a **reduced operator $\mathbf{G}_{\mathcal{W}}$** : $\mathcal{W} \rightarrow \mathcal{W}$ by $\mathbf{G}_{\mathcal{W}} \mathbf{x} \triangleq \mathbf{G}\mathbf{x}$ for all \mathbf{x} in \mathcal{W} . The operators \mathbf{G}_1 and \mathbf{G}_2 discussed earlier are examples of reduced operators. The following illustration shows that the reduced operator $\mathbf{G}_{\mathcal{W}}$ is truly different from \mathbf{G} .

Example 6. Reduced Linear Operators. We define $\mathbf{T}: \mathcal{R}^2 \rightarrow \mathcal{R}^2$ by

$$\mathbf{T}(\xi_1, \xi_2) \triangleq (2\xi_1 + 3\xi_2, 4\xi_2) \tag{4.4}$$

The matrix of \mathbf{T} relative to the standard basis \mathcal{E} is

$$\begin{aligned} [\mathbf{T}]_{\mathcal{E}\mathcal{E}} &= ([\mathbf{T}\mathbf{e}_1]_{\mathcal{E}} : [\mathbf{T}\mathbf{e}_2]_{\mathcal{E}}) \\ &= \begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix} \end{aligned}$$

The subspaces $\mathcal{W}_1 \triangleq \text{span}\{(1,0)\}$ and $\mathcal{W}_2 \triangleq \text{span}\{(3,2)\}$ are invariant under \mathbf{T} . Therefore, we can define the reduced operators $\mathbf{T}_1: \mathcal{W}_1 \rightarrow \mathcal{W}_1$ by $\mathbf{T}_1(\xi, 0) \triangleq \mathbf{T}(\xi, 0) = 2(\xi, 0)$ and $\mathbf{T}_2: \mathcal{W}_2 \rightarrow \mathcal{W}_2$ by $\mathbf{T}_2(3\xi, 2\xi) \triangleq \mathbf{T}(3\xi, 2\xi) = 4(3\xi, 2\xi)$. Using $\mathcal{X} \triangleq \{(1,0)\}$ as a basis for \mathcal{W}_1 and $\mathcal{Y} \triangleq \{(3,2)\}$ as a basis for \mathcal{W}_2 we find

$$\begin{aligned} [\mathbf{T}_1]_{\mathcal{X}\mathcal{X}} &= ([\mathbf{T}_1(1,0)]_{\mathcal{X}}) = (2) \\ [\mathbf{T}_2]_{\mathcal{Y}\mathcal{Y}} &= ([\mathbf{T}_2(3,2)]_{\mathcal{Y}}) = (4) \end{aligned}$$

The reduced operators \mathbf{T}_1 and \mathbf{T}_2 are scalar operators, represented by 1×1 matrices. They are very different from \mathbf{T} , which is represented by a 2×2 matrix. Clearly the domain and range of definition of a transformation are necessary parts of its definition.

Solution of Equations by Decomposition

The combination of three basic concepts—direct sum, invariance, and linearity—leads to the spectral decomposition, a decomposition of an

operator or an equation into a set of scalar multipliers or scalar single-variable equations. The decomposition provides considerable insight into the nature of linear models. It also provides a technique for solving equations which is an alternative to inverting the equations.

If \mathbf{T} is a linear operator on \mathcal{V} , if $\mathcal{V} = \mathcal{W}_1 \oplus \dots \oplus \mathcal{W}_p$, and if each \mathcal{W}_i is invariant under \mathbf{T} , then the set $\{\mathcal{W}_i\}$ of subspaces **decomposes** \mathbf{T} into a set of reduced linear operators $\mathbf{T}_i: \mathcal{W}_i \rightarrow \mathcal{W}_i$ defined by $\mathbf{T}_i \mathbf{x} \triangleq \mathbf{T} \mathbf{x}$ for all \mathbf{x} in \mathcal{W}_i . Analysis of a system represented by \mathbf{T} reduces to analysis of a set of *independent subsystems* represented by $\{\mathbf{T}_i\}$; that is, we can solve the equation $\mathbf{T} \mathbf{x} = \mathbf{y}$ by the following process.

The Spectral Decomposition Process (4.5)

- Using the direct sum, decompose \mathbf{y} into the unique combination

$$\mathbf{y} = \mathbf{y}_1 + \dots + \mathbf{y}_p \quad \text{with } \mathbf{y}_i \text{ in } \mathcal{W}_i$$

- Using the invariance of \mathcal{W}_i under \mathbf{T} , solve the subsystems

$$\mathbf{T} \mathbf{x}_i = \mathbf{y}_i \quad i = 1, 2, \dots, p$$

(in effect solving the reduced equations $\mathbf{T}_i \mathbf{x}_i = \mathbf{y}_i$).

- Using the linearity of \mathbf{T} , get the solution \mathbf{x} by adding

$$\mathbf{x} = \mathbf{x}_1 + \dots + \mathbf{x}_p$$

If the reduced operators \mathbf{T}_i are simple scalar multipliers like those of Example 6, then solution of the subsystem equations is trivial; that is, if $\mathbf{T} \mathbf{x}_i = \lambda_i \mathbf{x}_i$ for each \mathbf{x}_i in \mathcal{W}_i , then $\lambda_i \mathbf{x}_i = \mathbf{y}_i$ and parts (2) and (3) of (4.5) can be expressed as

$$\mathbf{x} = \left(\frac{1}{\lambda_1} \right) \mathbf{y}_1 + \dots + \left(\frac{1}{\lambda_p} \right) \mathbf{y}_p \quad (4.6)$$

If we know the invariant subspaces \mathcal{W}_i and the scalars λ_i , the primary effort required to carry out this procedure is that in decomposing \mathbf{y} .

Example 7. Solution of an Equation by Decomposition. Let $\mathbf{T}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be as in (4.4):

$$\mathbf{T}(\xi_1, \xi_2) \triangleq (2\xi_1 + 3\xi_2, 4\xi_2)$$

From Example 6, we know the subspaces $\mathcal{W}_1 \triangleq \text{span}\{(1,0)\}$ and $\mathcal{W}_2 \triangleq \text{span}\{(3,2)\}$ are invariant under \mathbf{T} ; furthermore, \mathbf{T} acts like $\mathbf{T}_1 \mathbf{x} \triangleq 2\mathbf{x}$ for \mathbf{x} in \mathcal{W}_1 , and like $\mathbf{T}_2 \mathbf{x} \triangleq 4\mathbf{x}$ for \mathbf{x} in \mathcal{W}_2 . Also $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$. Therefore, we can solve the

equation

$$\mathbf{T}\mathbf{x} = \mathbf{y} \triangleq (\eta_1, \eta_2)$$

by the process (4.5). We decompose \mathbf{y} by solving $(\eta_1, \eta_2) = c_1(1, 0) + c_2(3, 2)$ to find

$$\begin{aligned} (\eta_1, \eta_2) &= \left(\eta_1 - \frac{3\eta_2}{2} \right) (1, 0) + \left(\frac{\eta_2}{2} \right) (3, 2) \\ &\triangleq \mathbf{y}_1 + \mathbf{y}_2 \end{aligned}$$

By (4.6)

$$\begin{aligned} (\xi_1, \xi_2) &= \left(\frac{1}{2} \right) \left(\eta_1 - \frac{3\eta_2}{2} \right) (1, 0) + \left(\frac{1}{4} \right) \left(\frac{\eta_2}{2} \right) (3, 2) \\ &= \left(\frac{\eta_1}{2} - \frac{3\eta_2}{8}, \frac{\eta_2}{4} \right) \end{aligned}$$

The procedure (4.5) is essentially the one we use to determine the steady-state solution of a constant-coefficient differential equation by Fourier series. It is well known that a continuous function \mathbf{f} can be expanded uniquely as a Fourier series of complex exponentials of the form $e^{i2\pi kt/b}$, where $i = \sqrt{-1}$ and b is the length of the interval over which \mathbf{f} is defined. Each such exponential spans a subspace \mathcal{W}_k . The Fourier series expansion is possible because the space of continuous functions is in some sense the direct sum of $\{ \mathcal{W}_k \}$. But each subspace \mathcal{W}_k is invariant under any linear constant-coefficient differential operator; for instance, $(\mathbf{D}^2 + \mathbf{D})e^{\mu t} = (\mu^2 + \mu)e^{\mu t}$, a scalar multiple of $e^{\mu t}$. Thus the solution to certain differential equations can be found by an extension of (4.6). See P&C 5.35.

The Spectrum

The real goal of most systems analyses is insight into the system structure. Most linear models have a structure which permits decomposition into a set of scalar operations. It is not yet clear what effect the subdivision of a linear operator \mathbf{T} has on the overall computation. In fact, since one result of the decomposition is valuable insight into the structure of the system represented by \mathbf{T} , perhaps we should expect an increase in total computation. Although this expectation is justified, we find that under certain circumstances the decomposition information is known a priori. Then decomposition can also lead to reduced computation (Section 5.2).

Definition. An **eigenvalue** (or **characteristic value**) of a linear operator \mathbf{T} on a vector space \mathcal{V} is a scalar λ such that $\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$ for some nonzero

vector \mathbf{x} in \mathcal{V} . Any nonzero \mathbf{x} for which $\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$ is called an **eigenvector** of \mathbf{T} associated with the eigenvalue λ .

The eigenvector \mathbf{x} spans a subspace of \mathcal{V} . Each member of this subspace (or **eigenspace**) is also an eigenvector for the same eigenvalue. In fact, because \mathbf{T} is linear, any one-dimensional subspace which is invariant under \mathbf{T} must be an eigenspace of \mathbf{T} . The identity operator \mathbf{I} clearly has only one eigenvalue; the whole space \mathcal{V} is the eigenspace for $\lambda = 1$. Similarly, for the zero operator $\mathbf{0}$, \mathcal{V} is the eigenspace for $\lambda = 0$. If $\mathcal{V} = \mathcal{W}_1 \oplus \mathcal{W}_2$, then for the projector \mathbf{P}_i of (4.3), \mathcal{W}_i is the eigenspace for $\lambda = 1$ and \mathcal{W}_j is the eigenspace for $\lambda = 0$.

The eigenvectors of an operator which acts on a function space are often called **eigenfunctions**. We will refer to the eigenvalues and eigenvectors (or eigenfunctions) of \mathbf{T} as the **eigendata for \mathbf{T}** . The eigendata usually have some significant physical interpretation in terms of the system represented by \mathbf{T} .

Example 8. Eigendata for a Transformation in \mathcal{R}^2 . The operator $\mathbf{T}: \mathcal{R}^2 \rightarrow \mathcal{R}^2$ of (4.4) is

$$\mathbf{T}(\xi_1, \xi_2) \triangleq (2\xi_1 + 3\xi_2, 4\xi_2)$$

It has two eigenvalues: $\lambda_1 = 2$ and $\lambda_2 = 4$. The corresponding eigenspaces are $\text{span}\{(1,0)\}$ for λ_1 and $\text{span}\{(3,2)\}$ for λ_2 .

Example 9. Eigendata for Differential Operators. The exponential function $e^{\mu t}$ and its multiples form an eigenspace for any linear constant-coefficient differential operator *without boundary conditions*. For instance, since

$$\frac{d^n}{dt^n} e^{\mu t} + a_1 \frac{d^{n-1}}{dt^{n-1}} e^{\mu t} + \cdots + a_n e^{\mu t} = (\mu^n + a_1 \mu^{n-1} + \cdots + a_n) e^{\mu t}$$

for any complex scalar μ , the differential operator $\mathbf{D}^n + a_1 \mathbf{D}^{n-1} + \cdots + a_n \mathbf{I}$ has the eigenfunction $e^{\mu t}$ corresponding to the eigenvalue $\lambda = \mu^n + a_1 \mu^{n-1} + \cdots + a_n$. A differential operator without boundary conditions possesses a continuum of eigenvalues.

Example 10. An Operator Without Eigenvalues. A linear differential operator with homogeneous boundary conditions need not have any eigenvalues. For example, the only vector that satisfies

$$\frac{d\mathbf{f}(t)}{dt} = \lambda\mathbf{f}(t), \quad \mathbf{f}(0) = 0$$

is the zero function, regardless of the value we try for the eigenvalue λ . Thus the operator \mathbf{D} acting on the space of differentiable functions \mathbf{f} which satisfy $\mathbf{f}(0) = 0$ has no eigenvalues. Furthermore, any n th order linear differential operator with n independent one-point homogeneous boundary conditions is without eigenvalues. [See the discussion following (3.28).]

The problem of finding eigenvalues for a linear operator $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{V}$ is basically the problem of determining values of λ for which the equation

$$(\mathbf{T} - \lambda \mathbf{I})\mathbf{x} = \boldsymbol{\theta} \quad (4.7)$$

has nonzero solutions \mathbf{x} ; that is, we seek the values of λ for which the operator $\mathbf{T} - \lambda \mathbf{I}$ is singular. Once we have a specific eigenvalue, say, λ_1 , obtaining the corresponding eigenvectors involves the determination of $\text{nullspace}(\mathbf{T} - \lambda_1 \mathbf{I})$ —the solution of (4.7) with $\lambda = \lambda_1$. The determination of eigendata and the use of eigendata in practical analysis are explored for finite-dimensional systems in Section 4.2 and for infinite-dimensional systems in Section 4.3.

4.2 Spectral Analysis in Finite-Dimensional Spaces

In this section we convert (4.7) to a matrix equation for the case where \mathcal{V} is finite-dimensional. We also examine the spectral (eigendata) properties of matrix equations. Practical computation of eigendata for finite-dimensional problems, a more difficult task than appears on the surface, is discussed at the end of the section.

In Section 2.5 we found we could convert any equation involving a linear operator on a finite-dimensional space into an equivalent matrix equation. If $\mathbf{T}: \mathcal{V} \rightarrow \mathcal{V}$, we simply pick a basis \mathcal{Z} for \mathcal{V} . The basis converts the equation $\mathbf{T}\mathbf{x} = \mathbf{y}$ into the equation $[\mathbf{T}]_{\mathcal{Z}\mathcal{Z}}[\mathbf{x}]_{\mathcal{Z}} = [\mathbf{y}]_{\mathcal{Z}}$. We generally define $\mathbf{A} \triangleq [\mathbf{T}]_{\mathcal{Z}\mathcal{Z}}$, and use the simpler matrix notation $\mathbf{A}[\mathbf{x}]_{\mathcal{Z}} = [\mathbf{y}]_{\mathcal{Z}}$. The eigenvalues and eigenvectors for \mathbf{T} are then specified by the matrix equivalent of (4.7):

$$(\mathbf{A} - \lambda \mathbf{I})[\mathbf{x}]_{\mathcal{Z}} = [\boldsymbol{\theta}]_{\mathcal{Z}} \quad (4.8)$$

The values of λ for which (4.8) has nonzero solutions constitute the eigenvalues of \mathbf{T} . We also refer to them as the **eigenvalues of the matrix \mathbf{A}** .

From Section 1.5 we know that the square-matrix equation (4.8) has nonzero solutions if and only if

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \quad (4.9)$$

Equation (4.9) is known as the **characteristic equation of the matrix \mathbf{A}** (or of the operator \mathbf{T} which \mathbf{A} represents). If \mathbf{A} is an $n \times n$ matrix, then

$$c(\lambda) \triangleq \det(\lambda \mathbf{I} - \mathbf{A}) = (-1)^n \det(\mathbf{A} - \lambda \mathbf{I}) \quad (4.10)$$

is an n th order polynomial in λ called the **characteristic polynomial of \mathbf{A}** (or of \mathbf{T}). An n th order polynomial has precisely n (possibly complex) roots. (This fact follows from the fundamental theorem of algebra.) The set $\{\lambda_1, \dots, \lambda_n\}$ of roots of $c(\lambda)$ constitutes the complete set of eigenvalues of \mathbf{A} (or \mathbf{T}); the set is called the **spectrum of \mathbf{A}** (or \mathbf{T}). We often refer to an analysis which involves eigenvalues as a *spectral analysis*. Since $\lambda = \lambda_i$ makes $\mathbf{A} - \lambda\mathbf{I}$ singular, there must be at least one nonzero eigenvector for each different eigenvalue. A solution $[\mathbf{x}]_{\mathcal{E}}$ of (4.8) for $\lambda = \lambda_i$ is an eigenvector of \mathbf{A} for λ_i . The corresponding vector \mathbf{x} is an eigenvector of \mathbf{T} for λ_i .

Example 1. Finding Eigendata from $[\mathbf{T}]$. Let $\mathbf{T}: \mathcal{R}^2 \rightarrow \mathcal{R}^2$ be defined as in (4.4) by

$$\mathbf{T}(\xi_1, \xi_2) \triangleq (2\xi_1 + 3\xi_2, 4\xi_2)$$

Using the standard basis \mathcal{E} for \mathcal{R}^2 as in Example 6, (4.8) becomes

$$\left(\begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) [\mathbf{x}]_{\mathcal{E}} = [\mathbf{0}]_{\mathcal{E}}$$

or

$$\begin{pmatrix} 2-\lambda & 3 \\ 0 & 4-\lambda \end{pmatrix} [\mathbf{x}]_{\mathcal{E}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The characteristic equation is

$$\begin{vmatrix} 2-\lambda & 3 \\ 0 & 4-\lambda \end{vmatrix} = (2-\lambda)(4-\lambda) = 0$$

The eigenvalues of \mathbf{A} (and \mathbf{T}) are $\lambda_1 = 2$ and $\lambda_2 = 4$. We find the eigenvectors of \mathbf{A} for λ_i by solving (4.8) with $\lambda = \lambda_i$:

$$(\mathbf{A} - 2\mathbf{I}) [\mathbf{x}_1]_{\mathcal{E}} = \begin{pmatrix} 0 & 3 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow [\mathbf{x}_1]_{\mathcal{E}} = \begin{pmatrix} c_1 \\ 0 \end{pmatrix}$$

$$(\mathbf{A} - 4\mathbf{I}) [\mathbf{x}_2]_{\mathcal{E}} = \begin{pmatrix} -2 & 3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow [\mathbf{x}_2]_{\mathcal{E}} = \begin{pmatrix} 3d_1 \\ 2d_1 \end{pmatrix}$$

The scalars c_1 and d_1 are arbitrary; there is a one-dimensional eigenspace for each eigenvalue. The eigenvectors of \mathbf{T} for λ_i are found from the relationship between a vector and its coordinates relative to the basis \mathcal{E} :

$$[\mathbf{x}]_{\mathcal{E}} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \Leftrightarrow \mathbf{x} = c_1(1, 0) + c_2(0, 1)$$

Therefore, the eigenvectors of \mathbf{T} corresponding to λ_1 and λ_2 are

$$\mathbf{x}_1 = c_1(1, 0) + 0(0, 1) = c_1(1, 0)$$

$$\mathbf{x}_2 = 3d_1(1, 0) + 2d_1(0, 1) = d_1(3, 2)$$

In our previous discussions of vector spaces we have been able to allow freedom in the type of scalars which we use. We have thought primarily in terms of real numbers. However, in the discussion of eigenvalues this freedom in choice of scalars can cause difficulty. A real polynomial need not have real roots. Thus an operator on a space with real scalars may not have real eigenvalues; on the other hand, a complex eigenvalue has no meaning for such a space. The usual engineering practice is to accept the complex scalars whenever they appear, and assign them an appropriate meaning if necessary. We follow this approach, and assume, whenever we speak of eigenvalues, that the characteristic equation has a full set of roots.

Exercise 1. Define the operator \mathbf{T} on \mathcal{R}^2 by

$$\mathbf{T}(\xi_1, \xi_2) = (\xi_1 \cos \phi - \xi_2 \sin \phi, \xi_2 \cos \phi + \xi_1 \sin \phi) \quad (4.11)$$

This operator describes “rotation through the angle ϕ ” in \mathcal{R}^2 . Show that the eigendata for \mathbf{T} are

$$\lambda_1 = \cos \phi + i \sin \phi = e^{i\phi}, \quad \mathbf{x}_1 = (1, -i)$$

$$\lambda_2 = \cos \phi - i \sin \phi = e^{-i\phi}, \quad \mathbf{x}_2 = (1, i)$$

where $i = \sqrt{-1}$. The vector $(1, \pm i)$ is not a real 2-tuple; it is not in \mathcal{R}^2 .

We could have used any basis in Example 1. The eigenvalues and eigenvectors of \mathbf{T} are properties of \mathbf{T} ; they do not depend upon the basis. Suppose we use the invertible change of coordinate matrix \mathbf{S}^{-1} to convert (4.8) from the \mathcal{Z} coordinate system to a new coordinate system \mathcal{X} as in (2.54):

$$[\mathbf{x}]_{\mathcal{X}} = \mathbf{S}^{-1}[\mathbf{x}]_{\mathcal{Z}}$$

The effect of the change of coordinates on the matrix of \mathbf{T} is represented by the similarity transformation (2.62): $[\mathbf{T}]_{\mathcal{X}\mathcal{X}} = \mathbf{S}[\mathbf{T}]_{\mathcal{Z}\mathcal{Z}}\mathbf{S}^{-1}$. Recalling that $\mathbf{A} = [\mathbf{T}]_{\mathcal{Z}\mathcal{Z}}$, we find that (4.8) can be expressed as $([\mathbf{T}]_{\mathcal{Z}\mathcal{Z}} - \lambda\mathbf{I})[\mathbf{x}]_{\mathcal{Z}} = (\mathbf{S}[\mathbf{T}]_{\mathcal{Z}\mathcal{Z}}\mathbf{S}^{-1} - \lambda\mathbf{I})[\mathbf{x}]_{\mathcal{Z}} = \mathbf{S}([\mathbf{T}]_{\mathcal{X}\mathcal{X}} - \lambda\mathbf{I})\mathbf{S}^{-1}[\mathbf{x}]_{\mathcal{Z}} = [\boldsymbol{\theta}]_{\mathcal{Z}}$. Multiplying by the invertible matrix \mathbf{S}^{-1} , we find

$$([\mathbf{T}]_{\mathcal{X}\mathcal{X}} - \lambda\mathbf{I})[\mathbf{x}]_{\mathcal{X}} = [\boldsymbol{\theta}]_{\mathcal{X}} \quad (4.12)$$

Clearly, any λ which is an eigenvalue of \mathbf{A} is also an eigenvalue of any other matrix $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$ which represents \mathbf{T} . The similarity transformation, $[\mathbf{T}]_{\mathcal{X}\mathcal{X}} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$, results in a change in the *coordinates* of the eigenvectors of \mathbf{T} corresponding to λ , but it does not change either the eigenvectors of \mathbf{T} or the characteristic polynomial of \mathbf{T} .

Example 2. Invariance of Eigenvalues under a Change of Coordinates. The transformation $T: \mathcal{R}^2 \rightarrow \mathcal{R}^2$ of Example 1 is

$$T(\xi_1, \xi_2) \triangleq (2\xi_1 + 3\xi_2, 4\xi_2)$$

The eigenvectors $(1,0)$ and $(3,2)$ found for T in Example 1 form a basis for \mathcal{R}^2 ; denote this basis by \mathcal{X} . With respect to this basis,

$$\begin{aligned} [T]_{\mathcal{X}\mathcal{X}} &= \left([T(1,0)]_{\mathcal{X}} \ : \ [T(3,2)]_{\mathcal{X}} \right) \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} \end{aligned}$$

Then

$$\begin{aligned} \det([T]_{\mathcal{X}\mathcal{X}} - \lambda I) &= \begin{vmatrix} 2-\lambda & 0 \\ 0 & 4-\lambda \end{vmatrix} \\ &= (2-\lambda)(4-\lambda) \end{aligned}$$

The characteristic polynomial and the eigenvalues are those found in Example 1.

Diagonalization

It is apparent that the matrix of any linear operator T with respect to a basis of eigenvectors for T is of the form demonstrated in Example 2. If \mathcal{X} is a basis of eigenvectors, $[T]_{\mathcal{X}\mathcal{X}}$ has the eigenvalues of T on its diagonal; the rest of the matrix is zero. We call a linear operator $T: \mathcal{V} \rightarrow \mathcal{V}$ **diagonalizable** if there is a basis \mathcal{X} for \mathcal{V} which is composed of eigenvectors of T . We refer to the diagonal matrix $[T]_{\mathcal{X}\mathcal{X}}$ as the **spectral matrix** of T , and denote it by the symbol Λ . If \mathbf{A} is the matrix of T relative to some other basis, say \mathcal{Z} , for \mathcal{V} , we will also refer to Λ as the **diagonal form of \mathbf{A}** .

A basis of eigenvectors converts the operator equation $T\mathbf{x} = \mathbf{y}$ to the matrix equation

$$\Lambda[\mathbf{x}]_{\mathcal{X}} = [\mathbf{y}]_{\mathcal{X}} \tag{4.13}$$

Equation (4.13) is actually a matrix version of the process (4.5) for solving an equation by decomposition. Finding an eigenvector basis \mathcal{X} corresponds to finding a direct-sum decomposition of the space into subspaces \mathcal{W}_i which are invariant under T . Finding a coordinate matrix $[\mathbf{y}]_{\mathcal{X}}$ is equivalent to the decomposition of \mathbf{y} in (4.5). Inverting the diagonal (or "uncoupled") matrix Λ amounts to solving the reduced equations, $T_i \mathbf{x}_i = \lambda_i \mathbf{x}_i = \mathbf{y}_i$. When we find \mathbf{x} from the coordinates $[\mathbf{x}]_{\mathcal{X}}$, we are merely

combining the subsystem solutions as in (4.6). The process of computing eigenvalues and eigenvectors of matrices has been automated using a digital computer. Furthermore, the process of diagonalizing a matrix equation is more mnemonic than the decomposition process (4.5); the visual manner in which the eigenvalues and eigenvectors interact is easy to remember. Equation (4.13) is a clear and simple model for the system it represents.

What types of linear operators are diagonalizable? That is, for what finite-dimensional systems is there a basis of eigenvectors for the space? Since the existence of an eigenvalue λ_i implies the existence of a corresponding eigenvector \mathbf{x}_i , we expect the eigenvectors of an operator \mathbf{T} on an n -dimensional space \mathcal{V} to form a basis if its n eigenvalues are distinct. We verify that the n eigenvectors are independent if the eigenvalues are distinct by the test (2.11). Let

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n = \mathbf{0}$$

where \mathbf{x}_i is an eigenvector of \mathbf{T} for the eigenvalue λ_i . Operating with $(\mathbf{T} - \lambda_1\mathbf{I})$ we obtain

$$c_1(\lambda_1 \underset{0}{\downarrow} - \lambda_1)\mathbf{x}_1 + c_2(\lambda_2 - \lambda_1)\mathbf{x}_2 + \cdots + c_n(\lambda_n - \lambda_1)\mathbf{x}_n = \mathbf{0}$$

Successively operating with $(\mathbf{T} - \lambda_2\mathbf{I}), \dots, (\mathbf{T} - \lambda_{n-1}\mathbf{I})$ eliminates all terms but

$$c_n(\lambda_n - \lambda_1)(\lambda_n - \lambda_2) \cdots (\lambda_n - \lambda_{n-1})\mathbf{x}_n = \mathbf{0}$$

since $\lambda_i \neq \lambda_j$, $c_n = 0$. By backtracking, we can successively show that $c_{n-1} = \cdots = c_1 = 0$; the eigenvectors are independent and form a basis for the n -dimensional space.

In the above proof we applied the operator $(\mathbf{T} - \lambda_1\mathbf{I})(\mathbf{T} - \lambda_2\mathbf{I}) \cdots (\mathbf{T} - \lambda_{n-1}\mathbf{I})$ to a general vector in the space $\mathcal{N}^{n \times 1}$ (i.e., to a linear combination, $\mathbf{x} = \sum c_i\mathbf{x}_i$, of the eigenvectors in the basis). Suppose we operate once more, using the factor $(\mathbf{T} - \lambda_n\mathbf{I})$. Then, for any \mathbf{x} , we obtain

$$c_n(\lambda_n - \lambda_1)(\lambda_n - \lambda_2) \cdots (\lambda_n \underset{0}{\downarrow} - \lambda_n)\mathbf{x}_n = \mathbf{0}$$

That is,

$$(\mathbf{T} - \lambda_1\mathbf{I})(\mathbf{T} - \lambda_2\mathbf{I}) \cdots (\mathbf{T} - \lambda_n\mathbf{I}) = \mathbf{0} \quad (4.14)$$

Recall from (4.10) that if \mathbf{A} is a matrix of \mathbf{T} , the characteristic polynomial

for \mathbf{T} is $c(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$. Thus (4.14) is an operator analogue of $c(\lambda)$ which we denote by $c(\mathbf{T})$. The characteristic polynomial in \mathbf{T} annihilates all vectors in the space. This fact is commonly known as the **Cayley-Hamilton theorem**. It applies as well to matrices—a square matrix satisfies its own characteristic equation:

$$c(\mathbf{A}) = \mathbf{0} \quad (4.15)$$

Although we have proved the theorem only for an operator which is diagonalizable, it holds for all square matrices [see (4.85)].

Example 3. A Nondiagonalizable Matrix. Suppose

$$[\mathbf{T}]_{\mathfrak{X}\mathfrak{X}} = \mathbf{A} = \begin{pmatrix} \lambda_1 & 2 \\ 0 & \lambda_1 \end{pmatrix}$$

Then

$$\begin{aligned} c(\lambda) &= \det(\lambda\mathbf{I} - \mathbf{A}) \\ &= (\lambda - \lambda_1)^2 \end{aligned}$$

The only eigenvalue for \mathbf{A} is $\lambda = \lambda_1$. Using (4.8) we solve for the associated eigenvectors of \mathbf{A} :

$$(\mathbf{A} - \lambda_1\mathbf{I})[\mathbf{x}_1]_{\mathfrak{X}} \triangleq \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

or

$$[\mathbf{x}_1]_{\mathfrak{X}} = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$$

There are not enough independent eigenvectors of \mathbf{A} to form a basis for $\mathfrak{N}^{2 \times 1}$. The characteristic polynomial in \mathbf{A} is

$$\begin{aligned} c(\mathbf{A}) &= (\mathbf{A} - \lambda_1\mathbf{I})^2 \\ &= \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}^2 = \mathbf{0} \end{aligned}$$

It is apparent that the Cayley-Hamilton theorem also applies to matrices which do not possess distinct eigenvalues.

Although repeated eigenvalues can signal difficulty, it is possible for the eigenvectors to form a basis even though the eigenvalues are not distinct. A notable example is the identity operator; any vector in the space is an

eigenvector for the eigenvalue $\lambda = 1$. In Section 4.4 we discuss further those operators that are not diagonalizable.

Most matrices have distinct eigenvalues, and are thus diagonalizable. For a diagonalizable matrix \mathbf{A} , the eigenvalues by themselves (or the equivalent spectral matrix $\mathbf{\Lambda}$) give a rough idea of the manner in which the system operates. However, in order to be specific about the operation of the system, we need to know what \mathbf{A} does to specific vectors $[\mathbf{x}]_{\mathcal{X}}$ on which it operates. Thus we need the eigenvectors of \mathbf{A} . In the process of finding the eigenvectors, we relate \mathbf{A} and $\mathbf{\Lambda}$. A change of basis is the key. Let \mathbf{T} act on a finite-dimensional space \mathcal{V} . Assume $\mathbf{A} = [\mathbf{T}]_{\mathcal{X}\mathcal{X}}$. Let $\mathcal{X} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a basis for \mathcal{V} composed of eigenvectors of \mathbf{T} . Let $\{[\mathbf{x}_1]_{\mathcal{X}}, \dots, [\mathbf{x}_n]_{\mathcal{X}}\}$ be the corresponding basis for $\mathfrak{N}^n \times 1$ composed of eigenvectors of \mathbf{A} . Define the change of basis matrix \mathbf{S} by

$$\mathbf{S}[\mathbf{x}]_{\mathcal{X}} = [\mathbf{x}]_{\mathcal{X}} \quad (4.16)$$

Then, by (2.55),

$$\mathbf{S} = \left([\mathbf{x}_1]_{\mathcal{X}} \quad \cdots \quad [\mathbf{x}_n]_{\mathcal{X}} \right) \quad (4.17)$$

Furthermore, by (2.62),

$$\begin{aligned} [\mathbf{T}]_{\mathcal{X}\mathcal{X}} &= \mathbf{S}^{-1}[\mathbf{T}]_{\mathcal{X}\mathcal{X}}\mathbf{S} \\ &= \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \\ &= \mathbf{\Lambda} \end{aligned} \quad (4.18)$$

We call the matrix \mathbf{S} , the columns of which are eigenvectors of \mathbf{A} , a **modal matrix** for \mathbf{A} .^{*} Of course, the definition of \mathbf{S} in (4.16) is arbitrary; the roles of \mathbf{S} and \mathbf{S}^{-1} can be reversed. In order to help keep in mind which of the matrices \mathbf{S} and \mathbf{S}^{-1} is the modal matrix, we note that \mathbf{A} in (4.18) multiplies the eigenvectors of \mathbf{A} in the modal matrix.

An engineer often generates a system model directly in matrix form. The matrix form follows naturally from the use of standard models and standard physical units. When the underlying transformation is not explicitly stated, it becomes cumbersome to carry the coordinate notation $[\mathbf{x}]_{\mathcal{X}}$ for the vectors on which the $n \times n$ matrix \mathbf{A} operates. Under these circumstances, we will change the notation in (4.8) to

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \quad (4.19)$$

^{*}In some contexts the eigenvectors are referred to as modes of the system.

where \mathbf{x} is a vector in $\mathfrak{N}^{n \times 1}$. This new notation can cause confusion—we are using the same notation \mathbf{x} for both a vector (on which \mathbf{T} operates) and its coordinate matrix (which \mathbf{A} multiplies.) We must keep in mind that \mathbf{A} and \mathbf{x} may be representatives of an underlying transformation \mathbf{T} and a vector \mathbf{x} on which it operates.

Example 4. Diagonalization of a Matrix. Let

$$\mathbf{A} = \begin{pmatrix} 4 & -2 & 1 \\ -2 & 1 & 2 \\ 1 & 2 & 4 \end{pmatrix}$$

Then $c(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - 5)^2 (\lambda + 1) = 0$. The eigenvalues of \mathbf{A} are $\lambda_1 = 5$, $\lambda_2 = 5$, $\lambda_3 = -1$. The eigenvectors for $\lambda = 5$ satisfy

$$(\mathbf{A} - 5\mathbf{I})\mathbf{x} = \begin{pmatrix} -1 & -2 & 1 \\ -2 & -4 & 2 \\ 1 & 2 & -1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

or $\xi_3 = \xi_1 + 2\xi_2$. The eigenspace of \mathbf{A} for $\lambda = 5$ is two-dimensional; one basis for this space is

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$$

The eigenvectors for $\lambda = -1$ satisfy

$$(\mathbf{A} + \mathbf{I})\mathbf{x} = \begin{pmatrix} 5 & -2 & 1 \\ -2 & 2 & 2 \\ 1 & 2 & 5 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

or, by row reduction, $\xi_1 = -\xi_3$ and $\xi_2 = -2\xi_3$. The eigenspace of \mathbf{A} for $\lambda = -1$ is one-dimensional. We choose

$$\mathbf{x}_3 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$$

as a basis for this eigenspace. We use the eigenvectors \mathbf{x}_1 , \mathbf{x}_2 , and \mathbf{x}_3 of the matrix \mathbf{A} as the columns of a modal matrix \mathbf{S} for \mathbf{A} . We find \mathbf{S}^{-1} from \mathbf{S} by row reduction:

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & -1 \end{pmatrix} \quad \mathbf{S}^{-1} = \frac{1}{6} \begin{pmatrix} 5 & -2 & 1 \\ -2 & 2 & 2 \\ 1 & 2 & -1 \end{pmatrix}$$

The diagonal form of \mathbf{A} is:

$$\mathbf{\Lambda} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

The eigenvalues appear on the diagonal of $\mathbf{\Lambda}$ in the same order as their corresponding eigenvectors appear in the modal matrix.

Eigendata and Inverse Operators

If \mathbf{T} is an invertible operator and \mathbf{x} is an eigenvector of \mathbf{T} for the eigenvalue λ , it follows from the definition ($\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$) that

$$\mathbf{T}^{-1}\mathbf{x} = \left(\frac{1}{\lambda}\right)\mathbf{x} \quad (4.20)$$

That is, \mathbf{x} is also an eigenvector for \mathbf{T}^{-1} corresponding to the eigenvalue $1/\lambda$. Furthermore, \mathbf{T} is invertible if and only if $\lambda = 0$ is not an eigenvalue of \mathbf{T} . This fact is easily seen if \mathbf{T} acts on a finite-dimensional space: suppose \mathbf{A} is a matrix of \mathbf{T} (relative to some basis). Then $\lambda = 0$ is an eigenvalue of \mathbf{T} if and only if

$$\det(\mathbf{A} - 0\mathbf{I}) = 0 \quad (4.21)$$

But (4.21) is just the condition for noninvertibility of \mathbf{A} (and \mathbf{T}). If $\mathbf{\Lambda}$ is a diagonal form of \mathbf{A} , the relationship between the eigenvalues and invertibility is even more transparent. If $\lambda = 0$ is an eigenvalue of \mathbf{A} , then $\mathbf{\Lambda}$ has a zero row, and \mathbf{A} and \mathbf{T} are not invertible.

Example 5. Eigendata for an Inverse Matrix. The inverse of the matrix \mathbf{A} of Example 4 is

$$\mathbf{A}^{-1} = \frac{1}{5} \begin{pmatrix} 0 & -2 & 1 \\ -2 & -3 & 2 \\ 1 & 2 & 0 \end{pmatrix}$$

Using the spectral matrix $\mathbf{\Lambda}$ and the modal matrix \mathbf{S} for \mathbf{A} (from Example 4), we find the spectral matrix for \mathbf{A}^{-1} by

$$\mathbf{\Lambda}_{\mathbf{A}^{-1}} = \mathbf{S}^{-1}\mathbf{A}^{-1}\mathbf{S} = (\mathbf{S}^{-1}\mathbf{A}\mathbf{S})^{-1} = \mathbf{\Lambda}_{\mathbf{A}}^{-1}$$

or

$$\mathbf{\Lambda}_{\mathbf{A}^{-1}} = \begin{pmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Thus \mathbf{A} and \mathbf{A}^{-1} have inverse eigenvalues, but the same eigenvectors (modal matrices).

Computation of Eigendata for Matrices

Computation of the eigenvalues and eigenvectors of a square matrix appears straightforward. We need only solve for the roots λ_i of the characteristic polynomial, $c(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A})$, then solve the equation $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{x} = \mathbf{0}$ for the eigenvectors associated with λ_i . For the selected low-order matrices used in the examples and in the Problems and Comments, the eigendata can be computed exactly using this approach. As a practical matter, however, the process is difficult for an arbitrary diagonalizable matrix. For a matrix larger than, say, 3×3 , we resort to the digital computer.

Determination of the characteristic polynomial of the matrix by computing the determinant of $\lambda\mathbf{I} - \mathbf{A}$ is an expensive process. Computation of a simple $n \times n$ determinant requires $n^3/3$ multiplications, without the complication of the unspecified variable λ .^{*} A more efficient approach for finding $c(\lambda)$ is **Krylov's method**, which is based on the Cayley-Hamilton theorem (4.15).[†] The characteristic equation for the $n \times n$ matrix \mathbf{A} can be written

$$c(\lambda) = \lambda^n + b_1\lambda^{n-1} + \dots + b_n = 0 \quad (4.22)$$

where the coefficients $\{b_i\}$ are, as yet, unknown. By (4.15),

$$c(\mathbf{A}) = \mathbf{A}^n + b_1\mathbf{A}^{n-1} + \dots + b_n\mathbf{A} = \mathbf{0}$$

Then for an arbitrary vector \mathbf{x} in $\mathfrak{N}^{n \times 1}$,

$$\mathbf{A}^n\mathbf{x} + b_1\mathbf{A}^{n-1}\mathbf{x} + \dots + b_n\mathbf{x} = \mathbf{0} \quad (4.23)$$

For a specific \mathbf{x} , the vector equation (4.23) can be solved by row reduction to obtain the coefficients $\{b_i\}$. Note that the powers of \mathbf{A} need not be formed. Rather, \mathbf{x} is multiplied by \mathbf{A} n times. The method requires approximately n^3 multiplications to compute (4.23), then $n^3/3$ multiplications to solve for the coefficients $\{b_i\}$ by Gaussian elimination.

Example 6. Computing $c(\lambda)$ by Krylov's Method Let \mathbf{A} be the system matrix of Example 1, Section 3.4:

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

^{*}See Appendix 1 for a discussion of determinants and their evaluation.

[†]Ralston [4.13]. Refer also to P&C 1.3c.

The characteristic equation is second order:

$$c(\lambda) = \lambda^2 + b_1\lambda + b_2 = 0$$

$$c(\mathbf{A}) = \mathbf{A}^2 + b_1\mathbf{A} + b_2\mathbf{I} = \mathbf{0}$$

Let $\mathbf{x} = (1 \ 1)^T$. Then

$$\mathbf{A}^2\mathbf{x} + b_1\mathbf{A}\mathbf{x} + b_2\mathbf{x} = \mathbf{0}$$

or

$$\begin{pmatrix} -1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} -1 \\ -1 \end{pmatrix} + b_2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The solution to these equations is $b_1 = 1$, $b_2 = 0$. Therefore,

$$c(\lambda) = \lambda^2 + \lambda$$

Suppose that in Example 6 we had let $\mathbf{x} = (1 \ -1)^T$, the eigenvector of \mathbf{A} for $\lambda = -1$. Then (4.23) would have been

$$\begin{pmatrix} -1 \\ -1 \end{pmatrix} + b_1 \begin{pmatrix} -1 \\ 1 \end{pmatrix} + b_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

an underdetermined set of equations. The difficulty arises because $\mathbf{A} + \mathbf{I}$, one of the two factors of $c(\mathbf{A})$, is sufficient to annihilate \mathbf{x} . If we use an eigenvector of \mathbf{A} in (4.23), we can determine only those factors of $c(\mathbf{A})$ that annihilate the eigenvector. Thus is it possible to make a poor choice for \mathbf{x} in (4.23); try another! If the eigenvalues are not distinct, similar difficulties arise. (Try Krylov's method for $\mathbf{A} = \mathbf{I}$.)

Once we have $c(\lambda)$, we still need a scheme for finding its roots. A suitable method for finding the real roots is the iterative technique known as Newton's method. This method is discussed in detail in Section 8.1. If we need only the eigenvalues of \mathbf{A} [as in evaluating functions of matrices by (4.108)], and if these eigenvalues are real, Krylov's method together with Newton's method is a reasonable approach to obtaining them.

Denote the eigenvalue of \mathbf{A} which is of largest magnitude by λ_L . If λ_L is real, the **power method** obtains directly from \mathbf{A} both its largest eigenvalue λ_L and a corresponding eigenvector \mathbf{x}_L . The method relies on the "dominance" of the eigenvalue λ_L . Suppose eigenvectors of an $n \times n$ matrix \mathbf{A} form a basis $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ for $\mathfrak{R}^{n \times 1}$. Then any vector \mathbf{x} in $\mathfrak{R}^{n \times 1}$ can be expressed as $\mathbf{x} = \sum_{i=1}^n c_i \mathbf{x}_i$. Repeated multiplication of \mathbf{x} by \mathbf{A} yields $\mathbf{A}^k \mathbf{x} = \sum_{i=1}^n c_i \mathbf{A}^k \mathbf{x}_i = \sum_{i=1}^n c_i \lambda_i^k \mathbf{x}_i$. If one of the eigenvalues λ_L is larger in magni-

tude that the rest, then for large enough k , $\mathbf{A}^k \mathbf{x} \approx c_L \lambda_L^k \mathbf{x}_L$, an eigenvector for λ_L . Furthermore, λ_L is approximately equal to the ratio of the elements of $\mathbf{A}^{k+1} \mathbf{x}$ to those of $\mathbf{A}^k \mathbf{x}$. We explore the use of the power method in P&C 4.17. The method can be extended, by a process known as deflation, to obtain all the eigendata for \mathbf{A} . However, computational errors accumulate; the method is practical only for a few dominant eigenvalues. See Wilkinson [4.19].

Practical computation of the full set of eigenvectors of an arbitrary matrix is more difficult than is computation of the eigenvalues. The eigenvalues $\{\lambda_i\}$, by whatever method they are obtained, will be inexact, if only because of computer roundoff. Therefore, $(\mathbf{A} - \lambda_i \mathbf{I})$ is not quite singular; we need to compute the “near nullspace” of $(\mathbf{A} - \lambda_i \mathbf{I})$ (i.e., the “near solution” to $(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x} = \mathbf{0}$). In Section 2.4 we describe the **inverse iteration method** for determining a vector in the “near nullspace” of a nearly singular matrix. We now justify that method. If a matrix \mathbf{B} is nearly singular, its near nullspace is precisely the eigenspace for its smallest (least dominant) eigenvalue, λ_s . Then the near nullspace of \mathbf{B} is also the eigenspace for the largest (dominant) eigenvalue $1/\lambda_s$ of \mathbf{B}^{-1} . If λ_s is real, we can determine an eigenvector \mathbf{x}_s corresponding to λ_s by applying the power method to \mathbf{B}^{-1} . We pick an arbitrary vector \mathbf{z}_0 , and repetitively determine $\mathbf{z}_{k+1} = \mathbf{B}^{-1} \mathbf{z}_k$; for large enough k , the vector \mathbf{z}_k is a good approximation to \mathbf{x}_s ; the ratio of the components of \mathbf{z}_k to those of \mathbf{z}_{k+1} , is essentially λ_s . Thus the inverse iteration method is just the power method applied to the inverse matrix. In practice, rather than explicitly computing \mathbf{B}^{-1} , we would repetitively solve $\mathbf{B} \mathbf{z}_{k+1} = \mathbf{z}_k$, a less expensive operation.

The inverse iteration method can be used to obtain the eigenvectors of a matrix \mathbf{A} which correspond to a previously computed real eigenvalue 4. Just repetitively solve $(\mathbf{A} - \lambda_i \mathbf{I}) \mathbf{z}_{k+1} = \mathbf{z}_k$ for some initial vector \mathbf{z}_0 ; after several iterations, \mathbf{z}_k will approximate an eigenvector \mathbf{x}_i corresponding to 4. The ratio of the elements of \mathbf{z}_{k+1} to those of \mathbf{z}_k will approximate $1/\lambda_s$ where λ_s is the smallest eigenvalue of the matrix $\mathbf{B} = \mathbf{A} - \lambda_i \mathbf{I}$. The eigenvalue λ_s is a measure of the nonsingularity of \mathbf{B} and, therefore, the inaccuracy in λ_i ; a better approximation to the eigenvalue of \mathbf{A} is $\lambda_i + \lambda_s$. A highly accurate value of λ_i implies a low value of λ_s and, consequently, rapid convergence. Of course, small λ_s also implies an ill-conditioned matrix $(\mathbf{A} - \lambda_i \mathbf{I})$; yet, as discussed in Section 2.4, the resulting uncertainty in the solution will be a vector in nullspace $(\mathbf{A} - \lambda_i \mathbf{I})$. The inverse iteration method works well as long as the eigenvalue 4 is “isolated.” Any method will have trouble distinguishing between eigenvectors corresponding to nearly equal eigenvalues. *

*Wilkinson [4.19].

Example 7. Computing Eigenvectors by Inverse Iteration. Let \mathbf{A} be the following matrix

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

The exact eigendata of \mathbf{A} are

$$\lambda_1 = 1, \quad \mathbf{x}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \lambda_2 = -1, \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Suppose we have computed the eigenvalue $\hat{\lambda}_1 = 1 + \epsilon$, perhaps by means of Krylov's method and Newton's method. The equation $(\mathbf{A} - \hat{\lambda}_1 \mathbf{I})\mathbf{x} = \boldsymbol{\theta}$ has no nonzero solution. We use inverse iteration with the matrix $(\mathbf{A} - \hat{\lambda}_1 \mathbf{I})$ to approximate the true eigenvector \mathbf{x}_1 . Denote $\mathbf{z}_k = (\eta_1 \eta_2)^T$ and $\mathbf{z}_{k+1} = (\xi_1 \xi_2)^T$. Then

$$(\mathbf{A} - \hat{\lambda}_1 \mathbf{I})\mathbf{z}_{k+1} = \begin{pmatrix} -\epsilon & 0 \\ 1 & -2 - \epsilon \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \mathbf{z}_k$$

has the exact solution

$$\mathbf{z}_{k+1} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = -\frac{1}{\epsilon} \begin{pmatrix} 1 & 0 \\ 1 & \epsilon \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = (\mathbf{A} - \hat{\lambda}_1 \mathbf{I})^{-1} \mathbf{z}_k$$

Let $\mathbf{z}_0 = (1 \ 1)^T$. Then

$$\mathbf{z}_1 = -\frac{1}{\epsilon} \begin{pmatrix} 1 \\ \frac{\epsilon+1}{\epsilon+2} \end{pmatrix}, \quad \mathbf{z}_2 = \left(-\frac{1}{\epsilon}\right)^2 \begin{pmatrix} 1 \\ \frac{\epsilon^2+2\epsilon+2}{(\epsilon+2)^2} \end{pmatrix}$$

This sequence rapidly approaches a true eigenvector for λ_1 even if the approximate eigenvalue $\hat{\lambda}_1$ contains significant error. If $\epsilon = 0.1$, for instance, $\mathbf{z}_1 = -10 (1 \ .52)^T$ and $\mathbf{z}_2 = 100 (1 \ .501)^T$. The smallest eigenvalue of $(\mathbf{A} - \hat{\lambda}_1 \mathbf{I})$ is clearly $\hat{\lambda}_2 = -\epsilon$, which approaches zero as the error in $\hat{\lambda}_1$ approaches zero. It is apparent that for small ϵ , the elements of \mathbf{z}_k would soon become very large. Practical computer implementations of the inverse iteration method avoid large numbers by normalizing \mathbf{z}_k at each iteration.

If \mathbf{A} is symmetric, the eigenvalues of \mathbf{A} are real (P&C 5.28) and there is a basis of eigenvectors for the space.* The most efficient and accurate algorithms for determination of the full set of eigendata for a symmetric matrix avoid computation of the characteristic polynomial altogether. Rather, they perform a series of similarity transformations on \mathbf{A} , reducing the matrix to its diagonal form $\boldsymbol{\Lambda}$; the eigenvalues appear on the diagonal. Since $\boldsymbol{\Lambda} = \mathbf{S}^{-1} \mathbf{A} \mathbf{S}$, where \mathbf{S} is a matrix of eigenvectors, the sequence of

*See Section 5.4.

similarity transformations determines the eigenvectors of \mathbf{A} . See P&C 4.11 for an example of such a method.

Because methods that produce the full set of eigendata for a matrix must, in effect, determine both \mathbf{S} and \mathbf{S}^{-1} , we should expect the accuracy of the results to be related to the invertibility of the modal matrix \mathbf{S} . In point of fact, it can be shown that if \mathbf{S} is ill-conditioned, the eigenvalues are difficult to compute accurately; some of the eigenvalues are sensitive functions of the elements of \mathbf{A} . As a general rule, symmetric matrices have easily determined eigenvalues, whereas unsymmetric matrices do not. For a full discussion of computer techniques for computing eigendata, see Wilkinson [4.19] and Forsythe [4.6].

Application of Spectral Decomposition-Symmetrical Components

Since a sinusoid of specified frequency is completely determined by two real numbers, its amplitude and phase, we can represent it by a single complex number; for example, the function $2 \sin(\omega t + \phi)$ is equivalent to the complex number $2e^{i\phi}$, where $i = \sqrt{-1}$. Therefore, complex numbers adequately represent the steady-state 60-Hz sinusoidal voltages and currents in an electric power system (assuming physical units of volts and amperes, respectively).

Figure 4.2 is a simplified description of a three-phase electric power system. The complex amplitudes of the generated voltages, load voltages, and load currents are denoted by E_i , V_i , and I_i , respectively. These voltages and currents are related by the following matrix equations:

$$\mathbf{E} - \mathbf{V} = \mathbf{Z}\mathbf{I} \quad (4.24)$$

$$\mathbf{V} = \mathbf{W}\mathbf{I} \quad (4.25)$$

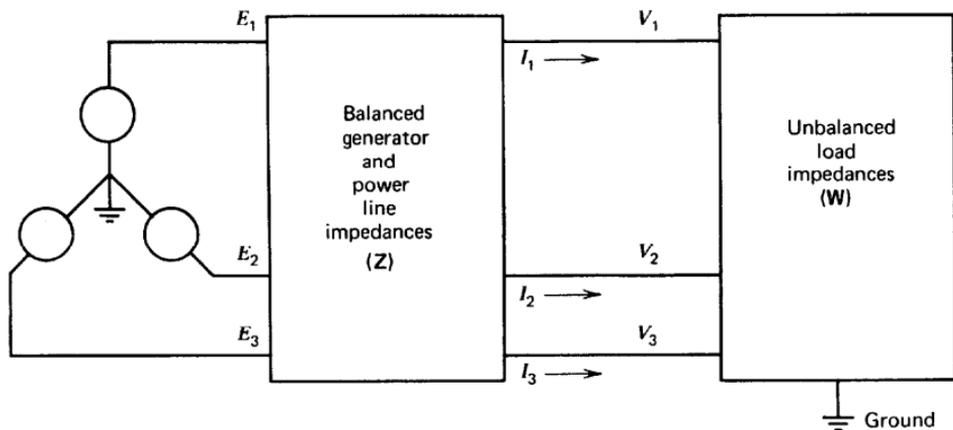


Figure 4.2. A three-phase electric power system.

where $\mathbf{E} = (E_1 \ E_2 \ E_3)^T$, $\mathbf{V} = (V_1 \ V_2 \ V_3)^T$, $\mathbf{I} = (I_1 \ I_2 \ I_3)^T$, and \mathbf{Z} and \mathbf{W} are 3×3 impedance matrices. In a typical power system, the generating system is balanced; that is, \mathbf{Z} has the form

$$\mathbf{Z} = \begin{pmatrix} z_1 & z_2 & z_2 \\ z_2 & z_1 & z_2 \\ z_2 & z_2 & z_1 \end{pmatrix} \quad (4.26)$$

A useful approach to analyzing a three-phase power system is to change coordinates in (4.24)-(4.25) in order to diagonalize (4.24). The method is known to power system engineers as the **method of symmetrical components**.

Exercise 2. Show (or verify) that the eigenvalues λ_i and corresponding eigenvectors \mathbf{x}_i of \mathbf{Z} are

$$\lambda_0 = z_1 + 2z_2 \quad \lambda_+ = z_1 - z_2 \quad \lambda_- = z_1 - z_2 \quad (4.27)$$

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \mathbf{x}_+ = \begin{pmatrix} 1 \\ a \\ a^2 \end{pmatrix} \quad \mathbf{x}_- = \begin{pmatrix} 1 \\ a^2 \\ a \end{pmatrix} \quad (4.28)$$

where $a = e^{i2\pi/3}$, a 120° counterclockwise rotation in the complex plane. (Note that $a^2 + a + 1 = 0$.) Let $\mathbf{S} = (\mathbf{x}_0 \ : \ \mathbf{x}_+ \ : \ \mathbf{x}_-)$. Show (or verify) that

$$\mathbf{S}^{-1} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{pmatrix} \quad (4.29)$$

Each of the eigenvectors (4.28) represents the complex amplitudes of a symmetrical three-phase sinusoidal quantity (voltage or current). The subscripts indicate the relative placement of the elements of each vector in the complex plane. The generated voltage vector \mathbf{E} typically has the form of \mathbf{x}_+ . The eigenvalues (4.27) can be interpreted as impedances associated with the symmetrical (eigenvector) components of the voltage and current vectors.

The engineer usually needs to analyze the generation and distribution system under various loads. If the load impedance matrix \mathbf{W} is an arbitrary matrix, it need not simplify during diagonalization. However, system loads are usually of a more specialized nature. For example, if the load is balanced (a goal of system planners), \mathbf{W} is of the same form as \mathbf{Z} , both (4.24) and (4.25) diagonalize simultaneously, only positive sequence quantities appear in the equations, and the matrix equations reduce to two scalar

equations. Certain unbalanced loads (such as a line-to-line fault) also lead to specialized forms of \mathbf{W} for which symmetrical component analysis is useful. A more complete discussion of symmetrical component analysis can be found in Rothe [4.15].

4.3 Spectral Analysis in Function Spaces

Spectral analysis is at least as helpful for understanding differential systems as it is for matrix equations. Furthermore, for many distributed systems (those described by partial differential equations) it provides the only reasonable approach to the determination of solutions. This section is devoted primarily to a discussion of spectral analysis of differential systems. We found in Example 9 of Section 4.1 that for a differential operator without boundary conditions, every scalar is an eigenvalue. The differential operators of real interest, however, are the ones we use in modeling systems. These ordinarily possess an appropriate number of boundary conditions. Suppose

$$\mathbf{L}\mathbf{f} \triangleq g_0(t) \frac{d^n \mathbf{f}(t)}{dt^n} + \cdots + g_n(t) \mathbf{f}(t) = \mathbf{u}(t) \quad (4.30)$$

$$\beta_i(\mathbf{f}) = \alpha_i \quad i = 1, \dots, n$$

It is convenient to decompose this differential system into two pieces:

$$\mathbf{L}\mathbf{f} = \mathbf{u} \quad \text{with } \beta_i(\mathbf{f}) = 0, \quad i = 1, \dots, n \quad (4.31)$$

and

$$\mathbf{L}\mathbf{f} = \mathbf{\theta} \quad \text{with } \beta_i(\mathbf{f}) = \alpha_i, \quad i = 1, \dots, n \quad (4.32)$$

Equation (4.32) is essentially finite dimensional in nature-by substituting for \mathbf{f} the complementary function $\mathbf{f}_c = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n$ of (3.19), we convert (4.32) to the matrix equation

$$\begin{pmatrix} \beta_1(\mathbf{v}_1) & \cdots & \beta_1(\mathbf{v}_n) \\ \vdots & & \vdots \\ \beta_n(\mathbf{v}_1) & \cdots & \beta_n(\mathbf{v}_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \quad (4.33)$$

We examined the eigendata for matrix operators in Section 4.2. We focus now on the infinite-dimensional problem (4.31).

We seek the eigenvalues and eigenfunctions for the system \mathbf{T} defined by \mathbf{L} together with the homogeneous boundary conditions of (4.31). That is,

we only allow \mathbf{L} to operate on functions which satisfy these boundary conditions. The equation which defines the eigendata is (4.7); thus

$$\begin{aligned}(\mathbf{L} - \lambda \mathbf{I})\mathbf{f} &= \boldsymbol{\theta} \\ \beta_i(\mathbf{f}) &= 0 \quad i = 1, \dots, n\end{aligned}\quad (4.34)$$

We introduce, by means of an example, a procedure for obtaining from (4.34) the eigenvalues and eigenfunctions associated with (4.31). The armature-controlled motor of (3.40)-(3.41) is modeled by $\mathbf{L}\phi \triangleq \mathbf{D}^2\phi + \mathbf{D}\phi$, with $\beta_1(\phi) \triangleq \phi(0)$ and $\beta_2(\phi) \triangleq \phi(b)$. For this specific \mathbf{L} and $\{\beta_i\}$, (4.34) becomes

$$\begin{aligned}\frac{d^2\phi(t)}{dt^2} + \frac{d\phi(t)}{dt} - \lambda\phi(t) &= 0 \\ \phi(0) = \phi(b) &= 0\end{aligned}\quad (4.35)$$

We first obtain a fundamental set of solutions for $(\mathbf{L} - \lambda \mathbf{I})$. The characteristic equation for $(\mathbf{L} - \lambda \mathbf{I})$, found by inserting $\phi(t) = e^{\mu t}$, is

$$\mu^2 + \mu - \lambda = 0$$

with roots

$$\mu = \frac{-1 \pm \sqrt{1 + 4\lambda}}{2}$$

If $\lambda = -\frac{1}{4}$, then the fundamental solutions are

$$\mathbf{v}_1(t) = e^{-t/2} \quad \mathbf{v}_2(t) = te^{-t/2}$$

Any nonzero solutions to (4.35) for $\lambda = -\frac{1}{4}$ must be of the form $\mathbf{f} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2$ and must satisfy the boundary conditions:

$$\begin{pmatrix} \beta_1(\mathbf{f}) \\ \beta_2(\mathbf{f}) \end{pmatrix} = \begin{pmatrix} \beta_1(\mathbf{v}_1) & \beta_1(\mathbf{v}_2) \\ \beta_2(\mathbf{v}_1) & \beta_2(\mathbf{v}_2) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ e^{-b/2} & be^{-b/2} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The boundary condition matrix is invertible; $c_1 = c_2 = 0$. There are no nonzero solutions for $\lambda = -\frac{1}{4}$, and $\lambda = -\frac{1}{4}$ is not an eigenvalue.

If $\lambda \neq -\frac{1}{4}$, a pair of fundamental solutions is

$$\mathbf{g}_1(t) = e^{-t/2} \exp\left(\frac{(1+4\lambda)^{1/2}t}{2}\right), \quad \mathbf{g}_2(t) = e^{-t/2} \exp\left(\frac{-(1+4\lambda)^{1/2}t}{2}\right)$$

A different but equivalent pair is

$$\mathbf{h}_1(t) = e^{-t/2} \cos\left(\frac{-i(1+4\lambda)^{1/2}t}{2}\right), \quad \mathbf{h}_2(t) = e^{-t/2} \sin\left(\frac{-i(1+4\lambda)^{1/2}t}{2}\right)$$

We let $\mathbf{g} = c_1\mathbf{g}_1 + c_2\mathbf{g}_2$, and again invoke the boundary conditions:

$$\begin{aligned} \begin{pmatrix} \beta_1(\mathbf{g}) \\ \beta_2(\mathbf{g}) \end{pmatrix} &= \begin{pmatrix} \beta_1(\mathbf{g}_1) & \beta_1(\mathbf{g}_2) \\ \beta_2(\mathbf{g}_1) & \beta_2(\mathbf{g}_2) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ e^{-b/2} \exp\left(\frac{(1+4\lambda)^{1/2}b}{2}\right) & e^{-b/2} \exp\left(\frac{-(1+4\lambda)^{1/2}b}{2}\right) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

There is a nonzero solution \mathbf{g} (or nonzero coefficients $\{c_i\}$) if and only if the boundary condition matrix is singular; thus, denoting the boundary condition matrix by $\mathbf{B}(\lambda)$,

$$\det(\mathbf{B}(\lambda)) = e^{-b/2} \exp\left(\frac{-(1+4\lambda)^{1/2}b}{2}\right) - e^{-b/2} \exp\left(\frac{(1+4\lambda)^{1/2}b}{2}\right) = 0$$

or

$$\exp\left[(1+4\lambda)^{1/2}b\right] = 1 \tag{4.36}$$

By analogy with the finite-dimensional case, we are inclined to refer to $\det(\mathbf{B}(\lambda)) = 0$ as the characteristic equation for the operator \mathbf{T} (\mathbf{L} with the homogeneous boundary conditions). However, the term characteristic equation is commonly used in reference to the equation (in the variable μ) used earlier to determine the fundamental solutions for \mathbf{L} . Therefore, we call $\det(\mathbf{B}(\lambda)) = 0$ the **eigenvalue equation** for \mathbf{T} . We may also refer to it as the eigenvalue equation for \mathbf{L} if it is clear which homogeneous boundary conditions are intended. The eigenvalue equation (4.36) is a transcendental equation in λ . To find the roots, recall from the theory of complex variables that*

$$\ln(e^{\alpha+i\gamma}) = \alpha + i\gamma + i2\pi k, \quad k=0, \pm 1, \pm 2, \dots$$

for real scalars α and γ . Thus (4.36) becomes

$$(1+4\lambda)^{1/2}b + i2\pi k = 0 \quad k=0, \pm 1, \pm 2, \dots$$

*See Chapter 14 of Wylie [4.18].

and the eigenvalues (for which nonzero solutions exist) are

$$\lambda_k = -\frac{1}{4} - \left(\frac{k\pi}{b}\right)^2 \quad k = 1, 2, 3, \dots \quad (4.37)$$

Note that $k = 0$ has been deleted; it corresponds to $\lambda = -\frac{1}{4}$, for which case \mathbf{g}_1 and \mathbf{g}_2 are not a fundamental set of solutions. Since k is squared, the positive and negative values of k yield identical values of λ ; thus, the positive values are sufficient.

We obtain the eigenfunctions ϕ_k corresponding to the eigenvalue λ_k by solving (4.35) with $\lambda = \lambda_k$. The solutions involve the roots μ_k of the characteristic equation:

$$\mu_k = \frac{-1 \pm (1 + 4\lambda_k)^{1/2}}{2} = -\frac{1}{2} \pm i \frac{k\pi}{b}$$

Since these roots are complex, we use the sinusoidal form $\{\mathbf{h}_i\}$ for the fundamental solutions:

$$\phi_k(t) = c_1 e^{-t/2} \cos\left(\frac{k\pi t}{b}\right) + c_2 e^{-t/2} \sin\left(\frac{k\pi t}{b}\right)$$

The boundary conditions yield

$$\begin{pmatrix} \beta_1(\phi_k) \\ \beta_2(\phi_k) \end{pmatrix} = \mathbf{B}(\lambda_k) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} e^0 \cos(0) & e^0 \sin(0) \\ e^{-b/2} \cos(k\pi) & e^{-b/2} \sin(k\pi) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

It follows that $c_1 = 0$ and c_2 is arbitrary. Letting $c_2 = 1$, we obtain the eigenfunction

$$\phi_k(t) = e^{-t/2} \sin\left(\frac{k\pi t}{b}\right) \quad (4.38)$$

corresponding to the eigenvalue λ_k .

The eigenfunctions for the two-point boundary value operator of (4.35) are analogous to the modes of oscillation of a string which is tied at both ends. The modes are harmonics of the fundamental or lowest-order mode, $e^{-t/2} \sin(\pi t / b)$; that is, the frequencies of oscillation are integral multiples of the lowest-order frequency. The number μ_k is the complex "natural frequency" of the k th mode. The eigenvalue λ_k can be thought of as a "characteristic number" for the k th mode. It is not clear whether or not \mathbf{T} is a diagonalizable operator. The eigenvalues are distinct; the set of eigenfunctions are suggestive of the terms of a Fourier series; however, we

wait until Chapter 5 to determine that there are sufficient eigenfunctions $\{\phi_k, k = 1, 2, \dots\}$ to form a basis for the space of functions \mathbf{f} on which \mathbf{T} (or \mathbf{L}) operates. (See Example 3, Section 5.3.)

Finding Eigendata for Differential Operators

For general differential equations of the form (4.30) we find eigendata by following the procedure used for the specific operator of (4.35). We first seek values of λ (or **eigenvalues**) for which (4.34) has nonzero solutions (**eigenfunctions**). Then we determine the corresponding eigenfunctions. We occasionally refer to the eigendata for the differential equation when we really mean the eigendata for the differential operator which determines the equation. Let the functions $\mathbf{v}_1(\lambda), \dots, \mathbf{v}_n(\lambda)$ be a fundamental set of solutions for $(\mathbf{L} - \lambda\mathbf{I})$; note that the functions depend on λ . The solutions to (4.34) consist in linear combinations

$$\mathbf{f}_c = c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n$$

which satisfy the boundary conditions. The coefficients are determined by the boundary condition matrix, whose λ dependency we denote explicitly by $\mathbf{B}(\lambda)$:

$$\mathbf{B}(\lambda) \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \beta_1(\mathbf{v}_1) & \cdots & \beta_1(\mathbf{v}_n) \\ \vdots & & \vdots \\ \beta_n(\mathbf{v}_1) & \cdots & \beta_n(\mathbf{v}_n) \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad (4.39)$$

There are nonzero solutions to (4.34) [or nonzero coefficients $\{c_i\}$ in (4.39)] only for λ such that

$$\det(\mathbf{B}(\lambda)) = 0 \quad (4.40)$$

As discussed beneath (4.36), we call (4.40) the **eigenvalue equation for \mathbf{T}** (or for \mathbf{L} with its boundary conditions). Its roots constitute the spectrum of \mathbf{T} (or of \mathbf{L} with its boundary conditions).

Determining the complementary function for $\mathbf{T} - \lambda\mathbf{I}$ is not necessarily a simple task. But it is the fundamental problem of differential equation analysis—standard techniques apply. The eigenvalue equation (4.40) is generally transcendental. Its solution, perhaps difficult, is a matter of algebra. Once we have determined a specific eigenvalue λ_k we return to (4.39) to determine those combinations of the fundamental solutions which are eigenfunctions for λ_k . The eigenfunctions are

$$\mathbf{f}_k = c_1 \mathbf{v}_1(\lambda_k) + \dots + c_n \mathbf{v}_n(\lambda_k) \quad (4.41)$$

where the scalars c_1, \dots, c_n satisfy

$$\mathbf{B}(\lambda_k) \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \theta$$

As noted in the discussion following (3.28), the boundary condition matrix for a one-point boundary value problem is always invertible. Thus if the boundary conditions for \mathbf{L} are all initial conditions, (4.40) has no roots, and the system \mathbf{T} has no eigenvalues.

Exercise 1. Seek the eigenvalues for the operator \mathbf{L} of (4.35) with the initial conditions $\phi(0) = \phi'(0) = 0$.

Example 1. Eigendata for a Heat-Flow Problem. Equation (3.1) is a steady-state description of a system wherein the heat generated within an insulated bar of length b diffuses toward heat sinks at the surfaces $t = 0$ and $t = b$. We now modify the second boundary condition. At $t = b$ we withdraw heat from the system by convection. The equation and modified boundary conditions for the temperature distribution \mathbf{f} are as follows:

$$\begin{aligned} (\mathbf{L}\mathbf{f})(t) &\triangleq -\frac{d^2\mathbf{f}(t)}{dt^2} = \mathbf{u}(t) \\ \beta_1(\mathbf{f}) &\triangleq \mathbf{f}(0) = \alpha_1, \quad \beta_2(\mathbf{f}) \triangleq \mathbf{f}'(b) + \mathbf{f}(b) = \alpha_2 \end{aligned} \tag{4.42}$$

The characteristic equation for $(\mathbf{L} - \lambda\mathbf{I})$ is

$$-\mu^2 - \lambda = 0$$

with roots $\mu = \pm i\sqrt{\lambda}$. We pick as a fundamental set of solutions (for $\lambda \neq 0$):

$$\mathbf{v}_1(t) = \cos \sqrt{\lambda} t, \quad \mathbf{v}_2(t) = \sin \sqrt{\lambda} t$$

The eigenvalue equation is

$$\begin{aligned} \det(\mathbf{B}(\lambda)) &= \begin{vmatrix} 1 & 0 \\ -\sqrt{\lambda} \sin(\sqrt{\lambda} b) + \cos(\sqrt{\lambda} b) & \sqrt{\lambda} \cos(\sqrt{\lambda} b) + \sin(\sqrt{\lambda} b) \end{vmatrix} \\ &= \sqrt{\lambda} \cos(\sqrt{\lambda} b) + \sin(\sqrt{\lambda} b) = 0 \end{aligned}$$

or

$$\tan \sqrt{\lambda} b = -\sqrt{\lambda} \tag{4.43}$$

Making the substitution $r \triangleq \sqrt{\lambda} b$, (4.43) becomes

$$\tan r = -\frac{r}{b} \quad (4.44)$$

Figure 4.3 shows the two halves of the eigenvalue equation plotted versus r for $b=2$. If $\{r_k, k=0, \pm 1, \pm 2, \dots\}$ are the roots of (4.44), then the eigenvalues for (4.42) are

$$\lambda_k = \frac{r_k^2}{b^2} \quad k = 1, 2, 3, \dots \quad (4.45)$$

The root r_0 has been eliminated. It corresponds to $\lambda=0$, for which the sinusoids are not a fundamental set of solutions. That $\lambda=0$ is not an eigenvalue is easily seen by repeating the above, using a fundamental set of solutions for $(\mathbf{L} - \mathbf{0I})$. Since

$$(r_{-k})^2 = (-r_k)^2 = r_k^2$$

the negative values of k are unnecessary. We find the eigenfunctions \mathbf{f}_k for λ_k by (4.41):

$$\mathbf{B}(\lambda_k) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{r_k}{b} \sin r_k + \cos r_k & \frac{r_k}{b} \cos r_k + \sin r_k \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

or $c_1 = 0$ and c_2 is arbitrary. Therefore, letting $c_2 = 1$, we obtain only one independent eigenvector,

$$\mathbf{f}_k(t) = \sin\left(\frac{r_k}{b} t\right) \quad (4.46)$$

for each eigenvalue $\lambda_k = r_k^2/b^2$, $k = 1, 2, 3, \dots$

In this example, the modes are not harmonic; the frequencies r_k^2/b^2 are not integral multiples of the lowest frequency. Although the operator of (4.42) is diagonalizable (the eigenvectors (4.46) form a basis for the domain of \mathbf{L}), we are not presently prepared to show it.

Eigendata for Integral Operators

We found in (4.20) that if an operator \mathbf{T} is invertible and $\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$, then $\mathbf{T}^{-1}\mathbf{x} = (1/\lambda)\mathbf{x}$. That is, the eigenvectors of \mathbf{T} and \mathbf{T}^{-1} are identical and correspond to reciprocal eigenvalues. From (4.40) we know that a differential system \mathbf{T} has the eigenvalue $\lambda=0$ if and only if $\det(\mathbf{B}(\lambda)) = \det(\mathbf{B}(\mathbf{0})) = 0$. But this is just the opposite of the condition (3.28) for invertibility of \mathbf{T} . Thus a differential system \mathbf{T} is invertible if and only if $\lambda=0$ is not an eigenvalue for \mathbf{T} . If we think in terms of a diagonalized $(\infty \times \infty)$ matrix representation of \mathbf{T} , it is clear that a zero eigenvalue is equivalent to singularity of the operator. Thus if $\lambda=0$ is an eigenvalue of \mathbf{T} , then the

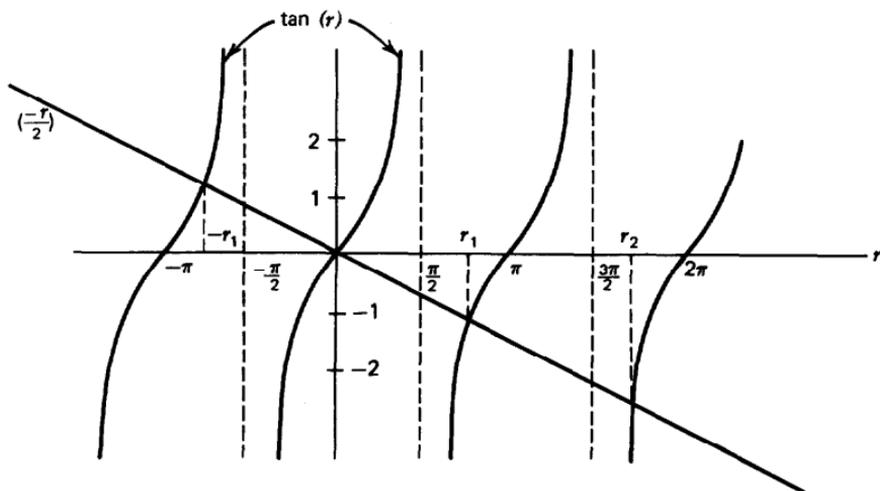


Figure 4.3. Roots of the eigenvalue equation (4.44) for $b = 2$.

Green's function for \mathbf{T} does not exist. Invertible differential and integral equations come in pairs, one the inverse of the other. Because the properties of integration are theoretically and computationally less troublesome than those of differentiation, we use the integral form to derive useful information about the eigenfunctions of operators and the solutions of equations (Sections 5.4 and 5.5). We also use the integral form for approximate numerical solution of equations. Yet because integral equations are difficult to solve, we often return to the differential form and standard differential equation techniques to determine the eigenfunctions of specific operators or the solutions of specific equations. In the following example, we obtain the eigendata for an integral operator from its differential inverse.

Example 2. Eigendata for an Integral Operator. The eigendata for the system \mathbf{T} represented by the differential operator $\mathbf{L} = \mathbf{D}^2 + \mathbf{D}$ with $\boldsymbol{\phi}(0) = 0$ and $\boldsymbol{\phi}(b) = 0$ are given in (4.37) and (4.38). They are

$$\lambda_k = -\frac{1}{4} - \left(\frac{k\pi}{b}\right)^2, \quad \boldsymbol{\phi}_k(t) = e^{-t/2} \sin\left(\frac{k\pi t}{b}\right), \quad k = 1, 2, \dots$$

Note that $\lambda = 0$ is not an eigenvalue. The Green's function for this operator is (3.42). Using this Green's function, we write the inverse of the differential system as

$$\begin{aligned} \boldsymbol{\phi}(t) &= \frac{1 - e^b e^{-t}}{e^b - 1} \int_0^t (e^s - 1) \mathbf{u}(s) ds + \frac{1 - e^{-t}}{e^b - 1} \int_t^b (e^s - e^b) \mathbf{u}(s) ds \\ &= (\mathbf{T}^{-1} \mathbf{u})(t) \end{aligned} \quad (4.47)$$

We expect the eigenfunctions of \mathbf{T}^{-1} to be the same as those of \mathbf{T} . Operating on ϕ_k with \mathbf{T}^{-1} , a complicated integration, we find

$$\begin{aligned} (\mathbf{T}^{-1}\phi_k)(t) &= \frac{1-e^b e^{-t}}{e^b-1} \int_0^t (e^s-1)e^{-s/2} \sin\left(\frac{k\pi s}{b}\right) ds \\ &\quad + \frac{1-e^{-t}}{e^b-1} \int_t^b (e^s-e^b)e^{-s/2} \sin\left(\frac{k\pi s}{b}\right) ds \\ &= \frac{1}{-1/4-(k\pi/b)^2} e^{-t/2} \sin\left(\frac{k\pi t}{b}\right) \\ &= \left(\frac{1}{\lambda_k}\right) \phi_k(t) \quad k=1,2,3,\dots \end{aligned} \tag{4.48}$$

The eigenvalues of the integral operator \mathbf{T}^{-1} are clearly $\{1/\lambda_k\}$.

Eigenvalue Problems in State Space

We introduced the state space model for dynamic systems in Section 3.4. We reproduce it here:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{4.49}$$

where \mathbf{A} is an $n \times n$ matrix multiplying the $n \times 1$ state vector $\mathbf{x}(t)$, and \mathbf{B} is an $n \times m$ matrix multiplying the $m \times 1$ input vector $\mathbf{u}(t)$. We know the differential system of (4.49) has no eigenvalues—it is an initial-value problem.* However, there is a meaningful and interesting eigenvalue problem associated with (4.49). It has to do with the system matrix \mathbf{A} . We introduce the relationship between the eigendata for the system matrix and the solutions of (4.49) by examining the system matrix for the n th-order constant-coefficient differential equation, the companion matrix of (3.36). The eigenvalues of \mathbf{A} are the roots of the equation $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$.

Exercise 2. Show that if \mathbf{A} is the companion matrix for the n th-order constant-coefficient differential equation

$$\mathbf{D}^n \mathbf{f} + a_1 \mathbf{D}^{n-1} \mathbf{f} + \dots + a_n \mathbf{f} = \mathbf{u} \tag{4.50}$$

then the characteristic equation for \mathbf{A} is

$$\det(\lambda\mathbf{I} - \mathbf{A}) = (\lambda^n + a_1 \lambda^{n-1} + \dots + a_n) = 0 \tag{4.51}$$

*If the initial condition vector is $\mathbf{x}(0) = \boldsymbol{\theta}$, then $\dot{\mathbf{x}}(t) - \mathbf{A}\mathbf{x}(t) - \lambda\mathbf{x}(t) = 0$ has *only* the zero solution, $\mathbf{x}(t) = \boldsymbol{\theta}$.

From (4.51), we see that if \mathbf{A} is the system matrix corresponding to an n th-order constant-coefficient differential equation, the characteristic equation for \mathbf{A} is the same as the characteristic equation (3.37) for the underlying n th-order differential equation. The eigenvalues of the system matrix are the exponents for a fundamental set of solutions to the differential equation. They are sometimes referred to as poles of the system. This relationship between the eigenvalues of the system matrix and the fundamental set of solutions to the underlying set of differential equations holds for any system matrix \mathbf{A} , not just for those in companion matrix form. [See the discussion below (4.94); refer also to P&C 4.16] Thus in the state-space equation (4.49) the concepts of matrix transformations and differential operators merge in an interesting way. The origin of the term “characteristic equation for the differential equation” is apparent. Fortunately, the state-space formulation is not convenient for boundary value problems. Thus eigenvalues of a system matrix and eigenvalues of a differential equation usually do not appear in the same problem.

Suppose we use the eigenvectors of the system matrix \mathbf{A} as a new basis for the state space, assuming, of course, that \mathbf{A} is diagonalizable. We change coordinates as in (4.16)-(4.18). (We can think of the state vector $\mathbf{x}(t)$ in $\mathcal{R}^{n \times 1}$ as representing itself relative to the standard basis for $\mathcal{R}^{n \times 1}$.) If $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a basis of eigenvectors for \mathbf{A} corresponding to the eigenvalues $\{\lambda_1, \dots, \lambda_n\}$, we transform $\mathbf{x}(t)$ into the new coordinates $\mathbf{y}(t)$ by the transformation

$$\mathbf{y}(t) = \mathbf{S}^{-1}\mathbf{x}(t) \quad (4.52)$$

where \mathbf{S} is the modal matrix for \mathbf{A} :

$$\mathbf{S} = (\mathbf{x}_1 \ ; \ \dots \ ; \ \mathbf{x}_n) \quad (4.53)$$

Then, by (4.18), (4.49) becomes

$$\begin{aligned} \mathbf{S}\dot{\mathbf{y}}(t) &= \mathbf{A}\mathbf{S}\mathbf{y}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{S}\mathbf{y}(0) &= \mathbf{x}_0 \\ \dot{\mathbf{y}}(t) &= \mathbf{S}^{-1}\mathbf{A}\mathbf{S}\mathbf{y}(t) + \mathbf{S}^{-1}\mathbf{B}\mathbf{u}(t) \\ &= \Lambda\mathbf{y}(t) + \mathbf{S}^{-1}\mathbf{B}\mathbf{u}(t), & \mathbf{y}(0) &= \mathbf{S}^{-1}\mathbf{x}_0 \end{aligned} \quad (4.54)$$

Equation (4.54) is a set of n uncoupled first-order differential equations which can be solved independently. The eigenvectors (or modes) of \mathbf{A} in a sense express natural relationships among the state variables [the elements of $\mathbf{x}(t)$] at each instant t . By using these eigenvectors as a basis, we eliminate the interactions—the new state variables [the elements of $\mathbf{y}(t)$] do not affect each other.

Example 3. Diagonalizing a State Equation The state equation for the armature controlled dc motor of (3.40) was obtained in Example 1 of Section 3.4.

It is

$$\dot{\mathbf{x}}(t) = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mathbf{u}(t), \quad \mathbf{x}(0) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \quad (4.55)$$

The eigendata for the system matrix are

$$\lambda_1 = 0, \quad \lambda_2 = -1 \quad \mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad (4.56)$$

The modal matrix is its own inverse

$$\mathbf{S}^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} = \mathbf{S} \quad (4.57)$$

The decoupled state equation is

$$\dot{\mathbf{y}}(t) = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{y}(t) + \begin{pmatrix} 1 \\ -1 \end{pmatrix} \mathbf{u}(t), \quad \mathbf{y}(0) = \begin{pmatrix} \alpha_1 + \alpha_2 \\ -\alpha_2 \end{pmatrix} \quad (4.58)$$

Denote the new state variables [elements of $\mathbf{y}(t)$] by $\mathbf{g}_1(t)$ and $\mathbf{g}_2(t)$. We can solve independently for \mathbf{g}_1 and \mathbf{g}_2 . On the other hand, we can use (3.79) with \mathbf{x} , \mathbf{A} , and \mathbf{B} replaced by \mathbf{y} , $\mathbf{\Lambda}$, and $\mathbf{S}^{-1}\mathbf{B}$, respectively. By either approach the result is

$$\mathbf{y}(t) \triangleq \begin{pmatrix} \mathbf{g}_1(t) \\ \mathbf{g}_2(t) \end{pmatrix} = \int_0^t \begin{pmatrix} 1 \\ -e^{-(t-s)} \end{pmatrix} \mathbf{u}(s) ds + \begin{pmatrix} 1 & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} \alpha_1 + \alpha_2 \\ -\alpha_2 \end{pmatrix} \quad (4.59)$$

Then

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{S}^{-1} \mathbf{y}(t) \\ &= \int_0^t \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -e^{-(t-s)} \end{pmatrix} \mathbf{u}(s) ds + \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} \alpha_1 + \alpha_2 \\ -\alpha_2 \end{pmatrix} \\ &= \int_0^t \begin{pmatrix} 1 - e^{-(t-s)} \\ e^{-(t-s)} \end{pmatrix} \mathbf{u}(s) ds + \begin{pmatrix} \alpha_1 + \alpha_2 - \alpha_2 e^{-t} \\ \alpha_2 e^{-t} \end{pmatrix} \end{aligned} \quad (4.60)$$

Compare this result with (3.80).

Note that the modal matrix in Example 3 is the Vandermond matrix for the system. Whenever the system matrix is in companion matrix form and the poles of the system are distinct, the Vandermond matrix is a modal matrix; then the eigenvectors of \mathbf{A} need not be calculated, but follow directly from the eigenvalues. See P&C 4.16.

Eigenvalue Problems and Partial Differential Equations

As we found in Example 10 of Section 4.1, not all differential operators have eigenvalues. This statement applies to both ordinary and partial differential operators. However, the most common analytical method for solving partial differential equations, separation of variables, generally introduces an eigenvalue problem even if the partial differential operator itself does not have eigenvalues. In point of fact, an analytical solution to a partial differential equation and its associated boundary conditions is usually obtainable only by summing eigenfunctions of a related differential operator. See Wylie [4.18]. On the other hand, some partial differential operators do have eigenvalues. One example is the Laplacian operator ∇^2 , defined by

$$\nabla^2 \mathbf{f}(s, t) \triangleq \frac{\partial^2 \mathbf{f}(s, t)}{\partial s^2} + \frac{\partial^2 \mathbf{f}(s, t)}{\partial t^2} \quad (4.61)$$

together with the “many-point” boundary conditions

$$\mathbf{f}(s, t) = 0 \quad \text{on} \quad \Gamma \quad (4.62)$$

where Γ is a closed curve in the (s, t) plane,

Exercise 3. Let Γ be the boundary of the rectangle with sides at $s = 0$, $s = a$, $t = 0$, and $t = b$. Show (by separation of variables) or verify that the eigenvalues and eigenfunctions for ∇^2 together with the boundary conditions (4.62) are:

$$\begin{aligned} \lambda_{mk} &= -\left(\frac{m\pi}{a}\right)^2 - \left(\frac{k\pi}{b}\right)^2 \\ \mathbf{f}_{mk}(s, t) &= \sin\left(\frac{m\pi s}{a}\right) \sin\left(\frac{k\pi t}{b}\right) \\ m &= 1, 2, \dots \quad k = 1, 2, \dots \end{aligned} \quad (4.63)$$

Notice that $\lambda = 0$ is not an eigenvalue of (4.61)-(4.62). Therefore the operator must be invertible, and we can expect to find a unique solution to Poisson's equation, $\nabla^2 \mathbf{f} = \mathbf{u}$, together with the boundary conditions of Example 3.

4.4 Nondiagonalizable Operators and Jordan Form

Most useful linear transformations are diagonalizable. However, there occasionally arises in practical analysis a system which is best modeled by a nondiagonalizable transformation. Probably the most familiar example is

a dynamic system with a pair of nearly equal poles. We use such an example to introduce the concept of nondiagonalizability.

Suppose we wish to solve the undriven differential equation $(\mathbf{D} + \mathbf{1})(\mathbf{D} + \mathbf{1} + \epsilon)\mathbf{f} = 0$ with the boundary conditions $\mathbf{f}(0) = \alpha_1$ and $\mathbf{f}'(0) = \alpha_2$, where ϵ is a small constant. The solution is of the form

$$\mathbf{f}(t) = c_1 e^{-t} + c_2 e^{-(1+\epsilon)t} \quad (4.64)$$

Applying the boundary conditions, we find

$$\begin{pmatrix} 1 & 1 \\ -1 & -(1+\epsilon) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

Since ϵ is small, this equation is ill-conditioned; it is difficult to compute accurately the multipliers c_1 and c_2 (see Section 1.5). The difficulty occurs because the poles of the system (or roots of the characteristic equation) are nearly equal; the functions e^{-t} and $e^{-(1+\epsilon)t}$ are nearly indistinguishable (see Figure 4.4). We resolve this computational difficulty by replacing e^{-t} and $e^{-(1+\epsilon)t}$ by a more easily distinguishable pair of functions; (4.64) becomes

$$\begin{aligned} \mathbf{f}(t) &= e^{-t}(c_1 + c_2 e^{-\epsilon t}) \\ &= e^{-t} \left[c_1 + c_2 \left(1 - \epsilon t + \frac{(\epsilon t)^2}{2!} - \dots \right) \right] \\ &\approx e^{-t} [(c_1 + c_2) - c_2 \epsilon t] \\ &= d_1 e^{-t} + d_2 t e^{-t} \end{aligned} \quad (4.65)$$

where $d_1 = c_1 + c_2$ and $d_2 = -\epsilon c_2$. Since ϵ is small, the functions e^{-t} and $t e^{-t}$ span essentially the same space as e^{-t} and $e^{-(1+\epsilon)t}$; yet this new pair of functions is clearly distinguishable (Figure 4.4b). The “new” function

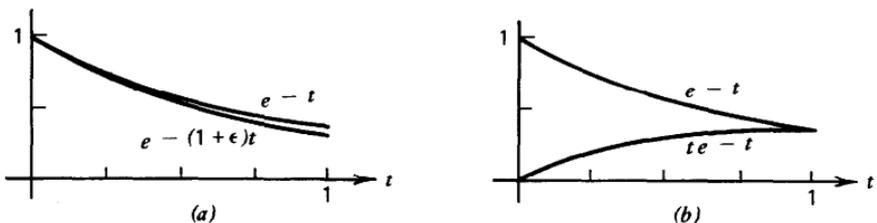


Figure 4.4. Alternative pairs of solutions to $(\mathbf{D} + \mathbf{1})(\mathbf{D} + \mathbf{1} + \epsilon)\mathbf{f} = \mathbf{0}$.

te^{-t} is essentially the difference between the two nearly equal exponentials. The boundary conditions now require

$$\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

or $\mathbf{f}(t) = \alpha_1 e^{-t} + (\alpha_1 + \alpha_2)te^{-t}$. We have eliminated the computational difficulty by equating the nearly equal poles of the system. When the roots of the characteristic equation are equal, (4.65) is the exact complementary function for the differential operator.

It is enlightening to view the differential system in state-space form. By writing the differential equation in the form $(\mathbf{D}^2 + (2 + \epsilon)\mathbf{D} + (1 + \epsilon)\mathbf{I})\mathbf{f} = \boldsymbol{\theta}$, we recognize from (3.63) that the state equation is

$$\dot{\mathbf{x}}(t) = \begin{pmatrix} 0 & 1 \\ -(1 + \epsilon) & -(2 + \epsilon) \end{pmatrix} \mathbf{x}(t), \quad \mathbf{x}(0) = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

The nearly equal poles of the system appear now as nearly equal eigenvalues of the system matrix, $\lambda_1 = -1, \lambda_2 = -(1 + \epsilon)$. We know from P&C 4.16 that the modal matrix is the Vandermond matrix;

$$\mathbf{S} = \begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -1 & -(1 + \epsilon) \end{pmatrix}$$

Since this matrix is ill-conditioned, we would have computational difficulty in finding \mathbf{S}^{-1} in order to carry out a diagonalization of the system matrix \mathbf{A} . However, if we equate the eigenvalues (as we did above), the system matrix becomes

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix}$$

which is not diagonalizable. Moreover, the earlier computational difficulty arose because we tried to diagonalize a “nearly nondiagonalizable” matrix.

The above example has demonstrated the need for dealing with nondiagonalizable transformations. In this section we explore nondiagonalizable finite-dimensional operators in detail. We discover that they can be represented by simple, nearly diagonal matrices which have the eigenvalues on the diagonal. Thus the conceptual clarity associated with the decoupling of system equations extends, to a great extent, to general linear operators.

To avoid heavy use of the cumbersome coordinate matrix notation, we focus throughout this section on matrices. However, we should keep in mind that an $n \times n$ matrix \mathbf{A} which arises in a system model usually

represents an underlying linear operator \mathbf{T} . The eigenvectors of \mathbf{A} are the coordinates of the eigenvectors of \mathbf{T} . Thus when we use a similarity transformation, $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$, to convert \mathbf{A} to a new form, we are merely changing the coordinate system for the space on which \mathbf{T} operates.

Generalized Nullspace and Range

Unlike a scalar, a linear operator \mathbf{U} is generally neither invertible nor zero. It lies in a “gray region” in between; \mathbf{U} takes some vectors to zero (acting like the zero operator); others it does not take to zero (thereby acting invertible). Perhaps even more significant is the fact that the nullspace and range of \mathbf{U} may overlap. The second and higher operations by \mathbf{U} may annihilate additional vectors. In some ways, the subspace annihilated by higher powers of \mathbf{U} is more characteristic of the operator than is nullspace (\mathbf{U}).

Example 1. Overlapping Nullspace and Range. Define the operator \mathbf{U} on $\mathfrak{R}^{3 \times 1}$ by $\mathbf{U}\mathbf{x} \triangleq \mathbf{B}\mathbf{x}$, where

$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Then \mathbf{U} has the following effect on a general vector in $\mathfrak{R}^{3 \times 1}$:

$$\begin{array}{ccccc} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} & \xrightarrow{\mathbf{U}} & \begin{pmatrix} \xi_2 \\ 0 \\ \xi_3 \end{pmatrix} & \xrightarrow{\mathbf{U}} & \begin{pmatrix} 0 \\ 0 \\ \xi_3 \end{pmatrix} \\ \mathfrak{R}^{3 \times 1} & & \text{range}(\mathbf{U}) & & \text{range}(\mathbf{U}^2) \end{array}$$

The vectors annihilated by various powers of \mathbf{U} are described by

$$\text{nullspace}(\mathbf{U}) = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\}, \quad \text{nullspace}(\mathbf{U}^2) = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$$

The nullspace and range of \mathbf{U}^k for $k > 2$ are the same as the nullspace and range of \mathbf{U}^2 .

Definition. The **generalized nullspace** $\mathfrak{N}_g(\mathbf{U})$ of a linear operator \mathbf{U} acting on an n -dimensional space \mathfrak{V} is the largest subspace of \mathfrak{V} annihilated by powers of \mathbf{U} . Since \mathfrak{V} is finite dimensional, the annihilation must terminate. Let q be that power of \mathbf{U} required for maximum annihilation.

We call q the **index of annihilation** for \mathbf{U} . Then $\mathfrak{N}_g(\mathbf{U}) = \text{nullspace}(\mathbf{U}^q)$. The generalized range $\mathfrak{R}_g(\mathbf{U})$ of the operator \mathbf{U} is defined by $\mathfrak{R}_g(\mathbf{U}) = \text{range}(\mathbf{U}^q)$. Since multiplication by a square matrix is a linear operator, we speak also of the generalized nullspace and generalized range of square matrices.

In Example 1, the index of annihilation is $q = 2$. The generalized range and generalized nullspace are

$$\mathfrak{R}_g(\mathbf{U}) = \text{span} \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}, \quad \mathfrak{N}_g(\mathbf{U}) = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$$

Notice that \mathfrak{V} is the direct sum of the generalized range and the generalized nullspace of \mathbf{U} . It is proved in Theorem 1 of Appendix 3 that any linear operator on an n -dimensional space splits the space in this manner. It is further shown in that theorem that both $\mathfrak{N}_g(\mathbf{U})$ and $\mathfrak{R}_g(\mathbf{U})$ are invariant under \mathbf{U} , and that \mathbf{U} acts like a reduced invertible operator on the generalized range of \mathbf{U} . These facts are verified by Example 1. An operator (or a square matrix) some power of which is zero is said to be **nilpotent**; \mathbf{U} acts like a reduced nilpotent operator on the generalized nullspace of \mathbf{U} .

Exercise 1. Let \mathbf{U} be the operator of Example 1. Define $\mathbf{U}_1: \mathfrak{R}_g(\mathbf{U}) \rightarrow \mathfrak{R}_g(\mathbf{U})$ by $\mathbf{U}_1 \mathbf{x} \triangleq \mathbf{U} \mathbf{x}$ for all \mathbf{x} in $\mathfrak{R}_g(\mathbf{U})$; define $\mathbf{U}_2: \mathfrak{N}_g(\mathbf{U}) \rightarrow \mathfrak{N}_g(\mathbf{U})$ by $\mathbf{U}_2 \mathbf{x} \triangleq \mathbf{U} \mathbf{x}$ for all \mathbf{x} in $\mathfrak{N}_g(\mathbf{U})$. Pick as bases for $\mathfrak{R}_g(\mathbf{U})$, $\mathfrak{N}_g(\mathbf{U})$, and $\mathfrak{N}^{3 \times 1}$ the standard bases

$$\mathfrak{X}_1 = \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}, \quad \mathfrak{X}_2 = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}, \quad \text{and} \quad \mathfrak{X} = \{ \mathfrak{X}_1, \mathfrak{X}_2 \}$$

respectively. Show that

$$[\mathbf{U}]_{\mathfrak{X} \mathfrak{X}} = \begin{pmatrix} [\mathbf{U}_1]_{\mathfrak{X}_1 \mathfrak{X}_1} & \mathbf{0} \\ \mathbf{0} & [\mathbf{U}_2]_{\mathfrak{X}_2 \mathfrak{X}_2} \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 1 \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}$$

What are the characteristics of \mathbf{U}_1 and \mathbf{U}_2 ? Why is the matrix of \mathbf{U} in "block-diagonal" form? (See P&C 4.3.)

Generalized Eigendata

The characteristic polynomial of an $n \times n$ matrix \mathbf{A} can be expressed in the form

$$c(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_p)^{m_p} \quad (4.66)$$

where p is the number of distinct eigenvalues, and $m_1 + \cdots + m_p = n$. We call m_i the **algebraic multiplicity** of λ_i . The eigenspace for λ_i is nullspace $(\mathbf{A} - \lambda_i \mathbf{I})$. The dimension of this eigenspace, the nullity of $(\mathbf{A} - \lambda_i \mathbf{I})$, we denote by k_i . We call k_i the **geometric multiplicity** of λ_i ; it is the number of independent eigenvectors of \mathbf{A} for λ_i . If the geometric multiplicity equals the algebraic multiplicity for each eigenvalue, it is reasonable to believe that there is a basis for $\mathfrak{R}^{n \times 1}$ composed of eigenvectors for \mathbf{A} , and that \mathbf{A} is diagonalizable.

If λ_i is deficient in eigenvectors ($k_i < m_i$), we say \mathbf{A} is **defective** at λ_i . If \mathbf{A} has any defective eigenvalues, we must pick noneigenvectors to complete the basis. We seek $(m_i - k_i)$ additional independent vectors from the subspace associated with λ_i —from the generalized nullspace of $(\mathbf{A} - \lambda_i \mathbf{I})$. Define

$$\begin{aligned} \mathfrak{W}_i &\stackrel{\Delta}{=} \text{generalized nullspace of } (\mathbf{A} - \lambda_i \mathbf{I}) \\ &= \text{nullspace}(\mathbf{A} - \lambda_i \mathbf{I})^{q_i} \end{aligned} \quad (4.67)$$

where q_i is the **index of annihilation** for $(\mathbf{A} - \lambda_i \mathbf{I})$. It is shown in Theorem 2 of Appendix 3 that

$$\dim(\mathfrak{W}_i) = m_i \quad (4.68)$$

We will think of all vectors in the generalized nullspace of $(\mathbf{A} - \lambda_i \mathbf{I})$ as generalized eigenvectors of \mathbf{A} for λ_i . Specifically, we call \mathbf{x}_r a **generalized eigenvector of rank r** for λ_i if

$$\begin{aligned} (\mathbf{A} - \lambda_i \mathbf{I})^r \mathbf{x}_r &= \mathbf{0} \\ (\mathbf{A} - \lambda_i \mathbf{I})^{r-1} \mathbf{x}_r &\neq \mathbf{0} \end{aligned} \quad (4.69)$$

If \mathbf{x}_r is a generalized eigenvector of rank r for λ_i , then $(\mathbf{A} - \lambda_i \mathbf{I}) \mathbf{x}_r$ is a

generalized eigenvector of rank $r - 1$; for (4.69) can be rewritten

$$(\mathbf{A} - \lambda_i \mathbf{I})^{r-1}(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_r = \mathbf{0}$$

$$(\mathbf{A} - \lambda_i \mathbf{I})^{r-2}(\mathbf{A} - \lambda_i \mathbf{I})\mathbf{x}_r \neq \mathbf{0}$$

Thus each vector in \mathscr{W}_i is a member of some chain of generalized eigenvectors generated by repeated multiplication with $(\mathbf{A} - \lambda_i \mathbf{I})$; the last member of each chain is a true eigenvector (of rank 1). We think of \mathscr{W}_i as the **generalized eigenspace** for λ_i ; \mathscr{W}_i contains precisely the m_i independent vectors associated with λ_i that we intuitively expect in a basis for $\mathscr{N}^{n \times 1}$.

In Theorem 3 of Appendix 3 we show that

$$\mathscr{N}^{n \times 1} = \mathscr{W}_1 \oplus \dots \oplus \mathscr{W}_p \tag{4.70}$$

Therefore, any bases which we pick for $\{ \mathscr{W}_i \}$ combine to form a basis for $\mathscr{N}^{n \times 1}$. Any basis for \mathscr{W}_i consists in m_i generalized eigenvectors. Furthermore, k_i of these m_i generalized eigenvectors can be true eigenvectors for λ_i .

Jordan Canonical Form

If \mathbf{A} is diagonalizable, we can diagonalize it by the similarity transformation $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$, where the columns of \mathbf{S} are a basis for $\mathscr{N}^{n \times 1}$ composed of eigenvectors of \mathbf{A} . Suppose \mathbf{A} is not diagonalizable. What form can we expect for the matrix $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ if the columns of \mathbf{S} are a basis of generalized eigenvectors of \mathbf{A} ? It depends on the way we pick the bases for the subspaces $\{ \mathscr{W}_i \}$. We demonstrate, by example, a way to pick the bases which results in as simple a form for the matrix $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ as we can possibly get in the presence of multiple eigenvalues. In order that the form be as nearly diagonal as possible, we include, of course, the true eigenvectors for λ_i in the basis for \mathscr{W}_i .

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & 0 & \vdots & & \\ 0 & 2 & 4 & \vdots & & \circ \\ 0 & 0 & 2 & \vdots & & \\ \hline & & & 2 & -1 & \\ & \circ & & 0 & 2 & \\ & & & & & \ddots \\ & & & & & & 3 \end{pmatrix} \tag{4.71}$$

Then $c(\lambda) = (\lambda - 2)^5(\lambda - 3)$, or $p = 2$, $\lambda_1 = 2$, $m_1 = 5$, $\lambda_2 = 3$, and $m_2 = 1$. Also,

$$\begin{aligned}
 (\mathbf{A} - 2\mathbf{I}) &= \begin{pmatrix} 0 & 3 & 0 & \vdots & & \\ 0 & 0 & 4 & & \circ & \\ 0 & 0 & 0 & \vdots & & \\ \hline & \circ & & 0 & -1 & \\ & & & 0 & 0 & \\ & & & & & \vdots \\ & & & & & 1 \end{pmatrix} \\
 (\mathbf{A} - 2\mathbf{I})^2 &= \begin{pmatrix} 0 & 0 & 12 & \vdots & & \\ 0 & 0 & 0 & & \circ & \\ 0 & 0 & 0 & \vdots & & \\ \hline & \circ & & 0 & 0 & \\ & & & 0 & 0 & \\ & & & & & \vdots \\ & & & & & 1 \end{pmatrix} \\
 (\mathbf{A} - 2\mathbf{I})^3 &= \begin{pmatrix} 0 & 0 & 0 & \vdots & & \\ 0 & 0 & 0 & & \circ & \\ 0 & 0 & 0 & \vdots & & \\ \hline & \circ & & 0 & 0 & \\ & & & 0 & 0 & \\ & & & & & \vdots \\ & & & & & 1 \end{pmatrix}, \\
 (\mathbf{A} - 3\mathbf{I}) &= \begin{pmatrix} -1 & 3 & 0 & \vdots & & \\ 0 & -1 & 4 & & \circ & \\ 0 & 0 & -1 & \vdots & & \\ \hline & \circ & & -1 & -1 & \\ & & & 0 & -1 & \\ & & & & & \vdots \\ & & & & & 0 \end{pmatrix}
 \end{aligned}$$

It is apparent that

$$\begin{aligned}
 \text{nullity}(\mathbf{A} - 2\mathbf{I}) &= 2 = k_1 \\
 \text{nullity}(\mathbf{A} - 2\mathbf{I})^2 &= 4 \\
 \text{nullity}(\mathbf{A} - 2\mathbf{I})^3 &= 5 \\
 \text{nullity}(\mathbf{A} - 3\mathbf{I}) &= 1 = k_2
 \end{aligned} \tag{4.72}$$

The indices of annihilation for $(\mathbf{A} - \lambda_1\mathbf{I})$ and $(\mathbf{A} - \lambda_2\mathbf{I})$, respectively, are $q_1 = 3$ and $q_2 = 1$. The five-dimensional subspace \mathcal{W}_1 , the generalized

eigenspace for λ_1 , consists in vectors of the form $(\xi_1 \ \xi_2 \ \xi_3 \ \xi_4 \ \xi_5 \ 0)^T$; vectors in \mathcal{W}_2 , the generalized eigenspace for λ_2 , are of the form $(0 \ 0 \ 0 \ 0 \ 0 \ \xi_6)^T$. [Note that (4.68) and (4.70) are verified in this example.]

Any eigenvector for $\lambda = 3$ will form a basis \mathcal{Q}_2 for \mathcal{W}_2 . Clearly, a basis \mathcal{Q}_1 for \mathcal{W}_1 must contain five vectors. Since there are only two independent true eigenvectors (of rank 1), three of the vectors in the basis must be generalized eigenvectors of rank greater than 1.

Assume we pick a basis which reflects the nullity structure of (4.72); that is, we pick two generalized eigenvectors of rank 1 for $\lambda = 2$, two of rank 2 for $\lambda = 2$, one of rank 3 for $\lambda = 2$, and one of rank 1 for $\lambda = 3$. Also assume we pick the basis vectors in chains; that is, if \mathbf{x} is a vector of rank 3 for $\lambda = 2$, and \mathbf{x} is in the basis, $(\mathbf{A} - 2\mathbf{I})\mathbf{x}$ and $(\mathbf{A} - 2\mathbf{I})^2\mathbf{x}$ will also be in the basis. We express both the nullity structure and chain structure by the following subscript notation:

$$\begin{array}{rcccl}
 & \text{rank 1} & \text{rank 2} & \text{rank 3} & \\
 & \downarrow & \downarrow & \downarrow & \\
 \mathcal{Q}_1 = & \begin{cases} \mathbf{x}_1 \\ \mathbf{x}_2 \end{cases} & \begin{cases} \mathbf{x}_{12} \\ \mathbf{x}_{22} \end{cases} & \mathbf{x}_{13} & \begin{array}{l} \leftarrow \text{chain 1} \\ \leftarrow \text{chain 2} \end{array} \\
 \mathcal{Q}_2 = & \{\mathbf{x}_3\} & & & \leftarrow \text{chain 3}
 \end{array} \tag{4.73}$$

This nullity and chain structure is expressed mathematically by the following equations:

$$\begin{aligned}
 (\mathbf{A} - 2\mathbf{I})\mathbf{x}_{13} &= \mathbf{x}_{12} \\
 (\mathbf{A} - 2\mathbf{I})\mathbf{x}_{12} &= \mathbf{x}_1 \\
 (\mathbf{A} - 2\mathbf{I})\mathbf{x}_1 &= \boldsymbol{\theta} \\
 (\mathbf{A} - 2\mathbf{I})\mathbf{x}_{22} &= \mathbf{x}_2 \\
 (\mathbf{A} - 2\mathbf{I})\mathbf{x}_2 &= \boldsymbol{\theta} \\
 (\mathbf{A} - 3\mathbf{I})\mathbf{x}_3 &= \boldsymbol{\theta}
 \end{aligned} \tag{4.74}$$

We propose the union of the sets \mathcal{Q}_i as a basis, denoted \mathcal{Q} , for $\mathcal{N}^{6 \times 1}$. It can be shown that a set of vectors of this form can be constructed and is a basis for $\mathcal{N}^{6 \times 1}$ (see Friedman [4.7]). Using the basis \mathcal{Q} , we form the change of coordinates matrix as in (4.17):

$$\mathbf{S} = (\mathbf{x}_1 \ ; \ \mathbf{x}_{12} \ ; \ \mathbf{x}_{13} \ ; \ \mathbf{x}_2 \ ; \ \mathbf{x}_{22} \ ; \ \mathbf{x}_3) \tag{4.75}$$

As in (4.18), this change of coordinates transforms \mathbf{A} into the matrix $\boldsymbol{\Lambda} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$. Recasting this similarity relation into the form $\mathbf{A}\mathbf{S} = \mathbf{S}\boldsymbol{\Lambda}$, we

recognize that

$$\begin{aligned}
 \mathbf{AS} &= \mathbf{A}(\mathbf{x}_1 : \mathbf{x}_{12} : \mathbf{x}_{13} : \mathbf{x}_2 : \mathbf{x}_{22} : \mathbf{x}_3) \\
 &= (2\mathbf{x}_1 : 2\mathbf{x}_{12} + \mathbf{x}_1 : 2\mathbf{x}_{13} + \mathbf{x}_{12} : 2\mathbf{x}_2 : 2\mathbf{x}_{22} + \mathbf{x}_2 : 3\mathbf{x}_3) \\
 &= (\mathbf{x}_1 : \mathbf{x}_{12} : \mathbf{x}_{13} : \mathbf{x}_2 : \mathbf{x}_{22} : \mathbf{x}_3) \begin{pmatrix} 2 & 1 & 0 & \vdots & & \\ 0 & 2 & 1 & & \circ & \\ 0 & 0 & 2 & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & \circ & & 2 & 1 & \\ & & & 0 & 2 & \\ & & & & & \ddots \\ & & & & & & 3 \end{pmatrix} \\
 &= \mathbf{SA} \tag{4.76}
 \end{aligned}$$

The form of $\mathbf{\Lambda}$ is as simple and as nearly diagonal a representation of \mathbf{A} as we can expect to obtain. The eigenvalues are on the diagonal. The off-diagonal 1's specify in a simple manner the "rank structure" or "chain structure" inherent in \mathbf{A} .

It is apparent that whenever the columns of \mathbf{S} form a basis for $\mathfrak{R}^{n \times 1}$ composed of generalized eigenvectors of \mathbf{A} , and these basis vectors consist in chains of vectors which express the nullity structure of \mathbf{A} as in (4.73)-(4.74), then $\mathbf{S}^{-1}\mathbf{AS}$ will be of the simple form demonstrated in (4.76). It will consist in a series of blocks on the diagonal; each block will be of the form

$$\begin{pmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & \lambda_i & 1 \\ 0 & \cdots & & 0 & \lambda_i \end{pmatrix}$$

By analogy with (4.16)-(4.18) in our discussion of diagonalization, we call \mathbf{S} the **modal matrix for \mathbf{A}** . We also call the near-diagonal matrix $\mathbf{\Lambda}$ the **spectral matrix for \mathbf{A}** (or for the underlying transformation \mathbf{T}). The spectral matrix is also referred to as the **Jordan canonical form of \mathbf{A}** . Each square block consisting in a repeated eigenvalue on the diagonal and an unbroken string of 1's above the diagonal is called a **Jordan block**. There is one Jordan block in \mathbf{A} for each chain of generalized eigenvectors in the basis. The dimension of each block equals the length of the corresponding chain. Thus we can tell from the nullity structure (4.71) alone, the form of the basis (4.73) and the precise form of $\mathbf{\Lambda}$ (4.76). Observe that the Jordan form is not unique. We can choose arbitrarily the order of the Jordan blocks by choosing the order in which we place the generalized eigenvectors in the basis.

Bases of Generalized Eigenvectors

We now generate a specific basis for $\mathfrak{N}^{6 \times 1}$ which is composed of generalized eigenvectors of the matrix \mathbf{A} of (4.71). That is, we find a basis of the form (4.73) by satisfying (4.74). We use (4.69) to find the highest rank vector in each chain. We first seek the vector \mathbf{x}_{13} of (4.73). All five of the basis vectors in \mathcal{Q}_1 satisfy $(\mathbf{A} - 2\mathbf{I})^3 \mathbf{x} = \boldsymbol{\theta}$. But only \mathbf{x}_{13} satisfies, in addition, $(\mathbf{A} - 2\mathbf{I})^2 \mathbf{x} \neq \boldsymbol{\theta}$. Therefore, we let $\mathbf{x}_{13} = (c_1 \ c_2 \ c_3 \ c_4 \ c_5 \ 0)^T$, the general solution to $(\mathbf{A} - 2\mathbf{I})^3 \mathbf{x} = \boldsymbol{\theta}$. Then

$$(\mathbf{A} - 2\mathbf{I})^2 \mathbf{x}_{13} = \begin{pmatrix} 12c_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \neq \boldsymbol{\theta} \quad (4.77)$$

or $c_3 \neq 0$. Thus any vector in $\mathfrak{N}^{6 \times 1}$ which has a zero sixth element and a nonzero third element is a generalized eigenvector of rank 3 for $\lambda = 2$. We have a lot of freedom in picking \mathbf{x}_{13} . Arbitrarily, we let $c_3 = 1$, and $c_1 = c_2 = c_4 = c_5 = 0$. Then

$$\mathbf{x}_{13} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}_{12} = (\mathbf{A} - 2\mathbf{I})\mathbf{x}_{13} = \begin{pmatrix} 0 \\ 4 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}_1 = (\mathbf{A} - 2\mathbf{I})\mathbf{x}_{12} = \begin{pmatrix} 12 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (4.78)$$

Notice that in (4.77) we looked at the eigenvector, $\mathbf{x}_1 = (\mathbf{A} - 2\mathbf{I})^2 \mathbf{x}_{13}$, at the end of the chain in order to determine the vector \mathbf{x}_{13} at the head of the chain.

To find the remaining vectors of \mathcal{Q}_1 , we look for the vector \mathbf{x}_{22} at the head of the second chain. By (4.69), all vectors $(d_1 \ d_2 \ d_3 \ d_4 \ d_5 \ d_6)^T$ of rank 2 or less satisfy

$$(\mathbf{A} - 2\mathbf{I})^2 \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \end{pmatrix} = \begin{pmatrix} 12d_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ d_6 \end{pmatrix} = \boldsymbol{\theta}$$

or $d_6 = d_3 = 0$. The vectors which are precisely of rank 2 also satisfy

$$(\mathbf{A} - 2\mathbf{I}) \begin{pmatrix} d_1 \\ d_2 \\ 0 \\ d_4 \\ d_5 \\ 0 \end{pmatrix} = \begin{pmatrix} 3d_2 \\ 0 \\ 0 \\ -d_5 \\ 0 \\ 0 \end{pmatrix} \neq \mathbf{0} \quad (4.79)$$

Again we are looking at the eigenvector at the end of the chain as we pick the constants. We must pick d_2 and d_5 , not both zero, such that \mathbf{x}_2 is independent of the eigenvector \mathbf{x}_1 selected above (i.e., $d_2 = 1, d_5 = 0$ will not do). Arbitrarily, we let $d_5 = 1, d_1 = d_2 = d_4 = 0; d_3$ is already zero. Thus

$$\mathbf{x}_{22} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = (\mathbf{A} - 2\mathbf{I})\mathbf{x}_{22} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \end{pmatrix} \quad (4.80)$$

The five vectors of (4.78) and (4.80) satisfy (4.73), and they are a basis for \mathcal{W}_1 . The equation $(\mathbf{A} - 3\mathbf{I})\mathbf{x} = 0$ determines the form of eigenvectors for $\lambda = 3$: $\mathbf{x} = (0 \ 0 \ 0 \ 0 \ 0 \ b_6)^T$. We arbitrarily let $b_6 = 1$ to get

$$\mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

a basis for \mathcal{W}_2 . By (4.76), this basis of generalized eigenvectors generates the modal matrix \mathbf{S} :

$$\mathbf{S} = \begin{pmatrix} 12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{12} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

where l_j is the length of the j th chain for λ_i and q_i is the index of annihilation for $(\mathbf{A} - \lambda_i \mathbf{I})$; thus q_i is the length of the longest chain. The nullities of various powers of $(\mathbf{A} - \lambda_i \mathbf{I})$ determine the structure of (4.82) just as (4.73) is determined by (4.72). The procedure for construction of the basis \mathcal{C}_i is as follows:

1. Determine the form of vectors of rank q_i or less by solving $(\mathbf{A} - \lambda_i \mathbf{I})^{q_i} \mathbf{x} = \mathbf{0}$.
2. Observe the true eigenvectors $(\mathbf{A} - \lambda_i \mathbf{I})^{q_i - 1} \mathbf{x}$; choose from the vectors found in (1) a total of $(\text{nullity}(\mathbf{A} - \lambda_i \mathbf{I})^{q_i} - \text{nullity}(\mathbf{A} - \lambda_i \mathbf{I})^{q_i - 1})$ vectors which lead to independent eigenvectors. These vectors are of rank q_i , and are the highest rank generalized eigenvectors in their respective chains.
3. Multiply each vector chosen in (2) by $(\mathbf{A} - \lambda_i \mathbf{I})$, thereby obtaining a set of generalized eigenvectors of rank $(q_i - 1)$, which is part of the set of basis vectors of rank $(q_i - 1)$.
4. Complete the set of basis vectors of rank $(q_i - 1)$ by adding enough vectors of rank $(q_i - 1)$ to obtain a total of $(\text{nullity}(\mathbf{A} - \lambda_i \mathbf{I})^{q_i - 1} - \text{nullity}(\mathbf{A} - \lambda_i \mathbf{I})^{q_i - 2})$ vectors which lead to independent eigenvectors. This step requires work equivalent to steps 1 and 2 with q_i replaced by $(q_i - 1)$. The vectors which are added are highest rank vectors in new chains.
5. Repeat steps 3 and 4 for lower ranks until a set of k_i eigenvectors is obtained.

Because $\mathfrak{N}^{n \times 1} = \mathfrak{W}_1 \oplus \dots \oplus \mathfrak{W}_p$, we can obtain a basis \mathcal{C} for $\mathfrak{N}^{n \times 1}$ consisting of generalized eigenvectors of \mathbf{A} by merely combining the bases for the subspaces \mathfrak{W}_i :

$$\mathcal{C} = \{ \mathcal{C}_1, \dots, \mathcal{C}_p \}$$

Proceeding as in the example of (4.71), we can use the basis \mathcal{C} to convert \mathbf{A} to its nearly diagonal Jordan canonical form $\mathbf{\Lambda}$.

Example 3. A Basis of Generalized Eigenvectors. Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 & 0 & 0 \\ -1 & 1 & 2 & 1 & 0 & 0 \\ -1 & 1 & -1 & 4 & 0 & 0 \\ -1 & 1 & -1 & 1 & 3 & 0 \\ -1 & 1 & -1 & 1 & 1 & 2 \end{pmatrix}$$

The process of finding and factoring the characteristic polynomial is complicated. We merely state it in factored form:

$$c(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - 3)^5 (\lambda - 2)$$

Therefore, $\lambda_1 = 3$, $m_1 = 5$, $\lambda_2 = 2$, and $m_2 = 1$. Furthermore,

$$(\mathbf{A} - 3\mathbf{I}) = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}, \quad (\mathbf{A} - 3\mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Clearly, $\text{nullity}(\mathbf{A} - 3\mathbf{I}) = 3$ and $\text{nullity}(\mathbf{A} - 3\mathbf{I})^2 = 5 = m_1$. It is also apparent that $\text{nullity}(\mathbf{A} - 3\mathbf{I})^3 = 5$. Thus $k_1 = 3$, $q_1 = 2$, and $\dim(\mathcal{W}_1) = 5$. Moreover,

$$(\mathbf{A} - 2\mathbf{I}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & -1 & 2 & 0 & 0 \\ -1 & 1 & -1 & 1 & 1 & 0 \\ -1 & 1 & -1 & 1 & 1 & 0 \end{pmatrix}$$

and $\text{nullity}(\mathbf{A} - 2\mathbf{I}) = 1$. As a result, $k_2 = 1$, $q_2 = 1$, and $\dim(\mathcal{W}_2) = 1$. [Note that $\dim(\mathcal{W}_1) + \dim(\mathcal{W}_2) = \dim(\mathcal{N}^{6 \times 1})$.] From the nullity information above, we know that the Jordan form of \mathbf{A} is

$$\Lambda = \begin{pmatrix} 3 & 1 & \vdots & & & \\ 0 & 3 & \vdots & & & \\ \vdots & \vdots & \vdots & \circ & & \\ & & 3 & 1 & & \\ & & 0 & 3 & & \\ \circ & & & & 3 & \\ & & & & & \vdots \\ & & & & & 2 \end{pmatrix}$$

We find a basis \mathcal{Q} for $\mathcal{N}^{6 \times 1}$ consisting in chains of generalized eigenvectors with the following structure:

$$\mathcal{Q} = \begin{cases} \mathcal{Q}_1 = \begin{cases} \mathbf{x}_1 & \mathbf{x}_{12} \\ \mathbf{x}_2 & \mathbf{x}_{22} \\ \mathbf{x}_3 \end{cases} \\ \mathcal{Q}_2 = \{\mathbf{x}_4\} \end{cases}$$

We first seek \mathbf{x}_{12} and \mathbf{x}_{22} , the vectors at the heads of the two longest chains. All generalized eigenvectors for $\lambda = 3$ satisfy $(\mathbf{A} - 3\mathbf{I})^2 \mathbf{x} = \mathbf{0}$. The solutions to this equation are of the form $\mathbf{x} = (c_1 \ c_2 \ c_3 \ c_4 \ c_5 \ c_5)^T$. The vectors of rank 2 also satisfy

$$(\mathbf{A} - 3\mathbf{I}) \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_5 \end{pmatrix} = (c_2 - c_1) \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + (c_4 - c_3) \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

We are looking at the true eigenvector at the end of the most general chain of

length 2. We must select two different sets of constants in order to specify both \mathbf{x}_{12} and \mathbf{x}_{22} . Furthermore, we must specify these constants in such a way that the eigenvectors \mathbf{x}_1 and \mathbf{x}_2 (which are derived from \mathbf{x}_{12} and \mathbf{x}_{22} , respectively) are independent. It is clear by inspection of the above equation that precisely two independent eigenvectors are available. By choosing $c_2 = 1$ and $c_1 = c_3 = c_4 = c_5 = 0$, we make

$$\mathbf{x}_{12} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

By selecting $c_4 = 1$ and $c_1 = c_2 = c_3 = c_5 = 0$ we get

$$\mathbf{x}_{22} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Of course, many other choices of \mathbf{x}_{12} and \mathbf{x}_{22} would yield the same \mathbf{x}_1 and \mathbf{x}_2 . Furthermore, other choices of \mathbf{x}_1 and \mathbf{x}_2 would also have been appropriate. We now seek \mathbf{x}_3 , a third true eigenvector for $\lambda = 3$ which is independent of \mathbf{x}_1 and \mathbf{x}_2 . The eigenvectors for $\lambda = 3$ satisfy $(\mathbf{A} - 3\mathbf{I}) = \mathbf{0}$. From the matrix $\mathbf{A} - 3\mathbf{I}$ we recognize that $c_1 = c_2$ and $c_3 = c_4$, as well as $c_5 = c_6$ for all eigenvectors for $\lambda = 3$. Letting $c_1 = c_2 = c_3 = c_4 = 0$ and $c_5 = c_6 = 1$, we obtain

$$\mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

an eigenvector independent of the other two. It is a simple matter to determine \mathbf{x}_4 , an eigenvector for $\lambda = 2$; we choose

$$\mathbf{x}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Exercise 3. Continuing Example 3, let

$$\mathbf{S} = (\mathbf{x}_1 : \mathbf{x}_{12} : \mathbf{x}_2 : \mathbf{x}_{22} : \mathbf{x}_3 : \mathbf{x}_4)$$

Show that $\mathbf{A} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$.

Generalized Eigenvectors in Function Spaces

Our discussion of generalized eigenvectors has been directed primarily toward matrices and, through matrices of transformations, toward any linear operator on an n -dimensional vector space. However, the concepts apply also to transformations on infinite-dimensional spaces. We have already noted that for the operator \mathbf{D} acting on the space $\mathcal{C}^1(0, 1)$, any scalar λ is an eigenvalue, and that $e^{\lambda t}$ is a corresponding eigenfunction. Furthermore, there is no other eigenfunction for λ which is independent from $e^{\lambda t}$ —the geometric multiplicity of λ is one.

We have not to this point explored the generalized nullspace for λ . In point of fact, powers of $(\mathbf{D} - \lambda \mathbf{I})$ do annihilate additional functions. Specifically, $(\mathbf{D} - \lambda \mathbf{I})^r$ annihilates the r -dimensional subspace of functions of the form $c_1 e^{\lambda t} + c_2 t e^{\lambda t} + c_3 t^2 e^{\lambda t} + \dots + c_r t^{r-1} e^{\lambda t}$. The annihilation does not terminate as r increases; the index of annihilation is infinite. It is apparent that the following functions constitute an infinite chain of generalized eigenfunctions of \mathbf{D} for the eigenvalue λ :

$$e^{\lambda t}, t e^{\lambda t}, \frac{1}{2!} t^2 e^{\lambda t}, \frac{1}{3!} t^3 e^{\lambda t}, \dots \quad (4.83)$$

Generally, differential operators are accompanied by boundary conditions. The eigenvalues of a differential operator \mathbf{L} (with its boundary conditions) are the roots of the eigenvalue equation (4.40), $\det(\mathbf{B}(\lambda)) = 0$. As in (4.41), the eigenfunctions corresponding to the eigenvalue λ_i are linear combinations of a set of fundamental solutions for \mathbf{L} , where the multipliers in the linear combination satisfy

$$\mathbf{B}(\lambda_i) \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

The **algebraic multiplicity** of the eigenvalue λ_i is the multiplicity of λ_i as a root of the eigenvalue equation. The nullity of $\mathbf{B}(\lambda_i)$ equals the number of independent eigenfunctions of \mathbf{L} for the single eigenvalue λ_i ; we call this number the **geometric multiplicity** of λ_i . It can be shown that $k_i \leq m_i$, just as we found for matrices (see Ince [4.10]). In the above example, where no boundary conditions were applied to the operator \mathbf{D} , these definitions do not apply. However, it seems appropriate in that case to assume that $m_i = \infty$ and $k_i = 1$ for each scalar λ_i , since there is an infinite string of generalized eigenfunctions associated with each λ_i . See P&C 4.12d for a differential operator (with boundary conditions) which possesses multiple eigenvalues.

The Minimal Polynomial

We showed in (4.15) that if an $n \times n$ matrix \mathbf{A} has distinct roots, its characteristic polynomial in \mathbf{A} is Θ ; that is, $c(\mathbf{A}) = (\mathbf{A} - \lambda_1 \mathbf{I}) \cdots (\mathbf{A} - \lambda_n \mathbf{I}) = \Theta$. We are now in a position to extend this result to all square matrices. The fact that $\mathfrak{N}^{n \times 1} = \mathfrak{W}_1 \oplus \cdots \oplus \mathfrak{W}_p$ is proved in Theorem 3 of Appendix 3. By definition (4.67), $(\mathbf{A} - \lambda_i \mathbf{I})^{q_i}$ annihilates \mathfrak{W}_i . Furthermore, \mathfrak{W}_j is invariant under $(\mathbf{A} - \lambda_i \mathbf{I})^{q_i}$ if $j \neq i$. Therefore, the matrix

$$(\mathbf{A} - \lambda_1 \mathbf{I})^{q_1} \cdots (\mathbf{A} - \lambda_p \mathbf{I})^{q_p}$$

annihilates the whole space $\mathfrak{N}^{n \times 1}$. We call

$$m(\lambda) \triangleq (\lambda - \lambda_1)^{q_1} \cdots (\lambda - \lambda_p)^{q_p}$$

the **minimal polynomial** for \mathbf{A} . The minimal polynomial in \mathbf{A} satisfies

$$m(\mathbf{A}) \triangleq (\mathbf{A} - \lambda_1 \mathbf{I})^{q_1} \cdots (\mathbf{A} - \lambda_p \mathbf{I})^{q_p} = \Theta \tag{4.84}$$

If $r \triangleq q_1 + \cdots + q_p$, then $m(\mathbf{A}) = \mathbf{A}^r + a_1 \mathbf{A}^{r-1} + \cdots + a_r \mathbf{I}$, an r th-order polynomial in \mathbf{A} . In fact, $m(\mathbf{A})$ is the lowest-order polynomial in \mathbf{A} which annihilates the whole space. It is apparent that polynomials in \mathbf{A} which include higher powers of $(\mathbf{A} - \lambda_i \mathbf{I})$ also annihilate the space. For instance, recalling that $m_i \geq q_i$, the characteristic polynomial in \mathbf{A} satisfies

$$c(\mathbf{A}) = (\mathbf{A} - \lambda_1 \mathbf{I})^{m_1} \cdots (\mathbf{A} - \lambda_p \mathbf{I})^{m_p} = \Theta \tag{4.85}$$

for any square matrix \mathbf{A} . Equation (4.85) is the Cayley-Hamilton theorem. Equations (4.84) and (4.85) find considerable use in computing. See, for example, Krylov's method (4.23) for finding the characteristic equation; see also the computation of functions of matrices via (4.108).

Example 4. A Minimal Polynomial Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Then $p = 1$, $\lambda_1 = 1$, and $c(\lambda) = (\lambda - 1)^3$. Since

$$(\mathbf{A} - \mathbf{I}) = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and $(\mathbf{A} - \mathbf{I})^2 = \Theta$, $q_i = 2$, and $m(\lambda) = (\lambda - 1)^2$. It is apparent that $c(\mathbf{A}) = m(\mathbf{A}) = \Theta$.

4.5 Applications of Generalized Eigendata

The concept of the Jordan form of a matrix is useful partly because it is mnemonic—it helps us remember and categorize the fundamental properties of the matrix (or the linear transformation which the matrix represents). The diagonal form of a diagonalizable matrix is merely a special case of the Jordan form. Whether an operator is diagonalizable or not, a complete eigenvalue analysis—obtaining eigenvalues and eigenvectors—is a computationally expensive process. Thus computational efficiency alone does not ordinarily justify the use of spectral decomposition (decomposition by means of eigenvectors) as a technique for solving an operator equation. However, our reason for analyzing an operator is usually to gain insight into the input-output relation which it describes. Spectral analysis of a model does develop intuitive insight concerning this input-output relation. In some instances a basis of eigenvectors is known a priori, and it need not be computed (e.g., the symmetrical components of (4.28), the Vandermond matrix of P&C 4.16, and the complex exponential functions of Fourier series expansions). In these instances, we gain the insight of spectral decomposition with little more effort than that involved in solution of the operator equation.

Nearly Equal Eigenvalues

True multiple eigenvalues rarely appear in physical systems. But nearly equal eigenvalues are often accompanied by near singularity of the linear operator and, therefore, by computational difficulty. This difficulty can sometimes be avoided by equating the nearly equal eigenvalues and computing generalized eigenvectors in the manner described earlier.

Example 1. Nearly Equal Eigenvalues. In the introduction to Section 4.4 we described a dynamic system with nearly equal poles: $(\mathbf{D} + \mathbf{1})(\mathbf{D} + \mathbf{1} + \epsilon)\mathbf{f} = \boldsymbol{\theta}$ with $\mathbf{f}(0) = \boldsymbol{\alpha}_1$ and $\mathbf{f}'(0) = \boldsymbol{\alpha}_2$. As we found in our earlier discussion, the near equality of the poles causes computational difficulty which we remove by equating the poles. But equating the nearly equal poles is equivalent to replacing the nearly dependent set of solutions $\{e^{-t}, e^{-(1+\epsilon)t}\}$ by the easily distinguishable pair of functions $\{e^{-t}, te^{-t}\}$. Since the poles are made identical ($\epsilon = 0$), the state-space representation of the system becomes $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, where

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix} \quad (4.86)$$

This system matrix is not diagonalizable. The pair of vectors $\mathbf{x}_1 = (1 \ -1)^T$ and $\mathbf{x}_{12} = (\frac{1}{2} \ \frac{1}{2})^T$ is a two-vector chain of generalized eigenvectors of \mathbf{A} for the single

eigenvalue $\lambda = -1$. This pair of vectors is a basis for the state space. Therefore, the matrix

$$\mathbf{S} = \begin{pmatrix} 1 & \frac{1}{2} \\ -1 & \frac{1}{2} \end{pmatrix} \quad (4.87)$$

is a modal matrix for the system. Note that \mathbf{S} is well conditioned. There will be no computational difficulty in inverting \mathbf{S} . The nondiagonal spectral matrix for the system is

$$\mathbf{\Lambda} = \mathbf{S}^{-1} \mathbf{A} \mathbf{S} = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \quad (4.88)$$

Example 1 demonstrates the practical value of the concepts of generalized eigenvectors and Jordan form. Even though these concepts are important, the full generality of the Jordan form is seldom, if ever, needed. We are unlikely to encounter, in practice, a generalized eigenspace more complex than that characterized by the single two-vector chain of generalized eigenvectors of Example 1. In Example 1, the system matrix \mathbf{A} is nondiagonalizable only for $\epsilon = 0$. We focused on this nondiagonalizable case because it characterizes the situation for small ϵ better than does the true barely diagonalizable case.* It seems that diagonalizability is the rule in models which represent nature, except at the boundary between certain regions or at the limit of certain approximations. In Example 1, diagonalizability broke down completely only at the boundary between the two regions defined by $\epsilon > 0$ and $\epsilon < 0$. Yet from a practical point of view the boundary is a fuzzy, "small ϵ " transition region.

Pease [4.12, p. 81] presents a spectral analysis of the transmission of electrical signals through a 2-port system. His analysis illustrates the way that nondiagonalizability characterizes the boundary between different regions. The 2×2 system matrix which describes the transmission of signals through the 2-port network is diagonalizable for all sinusoidal signals except signals at the upper or lower cutoff frequencies. At these two frequencies the spectral analysis breaks down because of nondiagonalizability of the matrix of 2-port parameters. However, the analysis can be salvaged by using generalized eigenvectors. Even for frequencies *near* the cutoff frequencies, the spectral analysis is aided by the use of generalized eigenvectors because of the *near* nondiagonalizability of the system matrix.

* Forsythe [4.6] explores other problems in which accuracy is improved by treating near singularity as true singularity.

Application of Jordan Form—Feedback Control

The most common model for a linear time-invariant dynamic system is the state equation (3.67):

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) \text{ given} \quad (4.89)$$

where $\mathbf{x}(t)$ is the state (or condition) of the system at time t , and $\mathbf{u}(t)$ is the control (or input) at time t ; \mathbf{A} and \mathbf{B} are arbitrary $n \times n$ and $n \times m$ matrices, respectively. In (3.79) we inverted the state equation, obtaining

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-s)}\mathbf{B}\mathbf{u}(s)ds \quad (4.90)$$

where the state transition matrix (or matrix exponential) $e^{\mathbf{A}t}$ is defined as the sum of an infinite series of matrices (3.72).

Equations (4.89) and (4.90) are generalizations of the simple first-order linear constant-coefficient differential equation

$$\dot{\mathbf{f}}(t) = a\mathbf{f}(t) + b\mathbf{u}(t), \quad \mathbf{f}(0) \text{ given} \quad (4.91)$$

which has the solution

$$\mathbf{f}(t) = e^{at}\mathbf{f}(0) + \int_0^t e^{a(t-s)}b\mathbf{u}(s)ds$$

Another approach to the solution of (4.91) is through frequency domain analysis.* Taking the Laplace transform of (4.91), we obtain

$$s\mathbf{F}(s) - \mathbf{f}(0) = a\mathbf{F}(s) + b\mathbf{U}(s)$$

or

$$\mathbf{F}(s) = \left(\frac{1}{s-a}\right)\mathbf{f}(0) + \left(\frac{b}{s-a}\right)\mathbf{U}(s) \quad (4.92)$$

where the symbols \mathbf{F} and \mathbf{U} are the Laplace transforms of \mathbf{f} and \mathbf{u} , respectively. The function $(s-a)^{-1}$ is known as the transfer function of the system (4.91). The pole of the transfer function ($s=a$) characterizes the time response of the system. In fact, the transfer function is the Laplace transform of the impulse response of the system, e^{at} .

The relationships among the variables in a linear equation can be represented pictorially by means of a signal flow graph. A signal flow

*For an introduction to frequency domain analysis, see Appendix 2. Refer also to Schwartz and Friedland [4.16] or DeRusso, Roy, and Close [4.3].

graph for (4.91) is shown in Figure 4.5. The variables in the system are associated with nodes in the graph. The arrows indicate the flow of information (or the relationships among the variables). The encircled symbols contained in each arrow are multipliers. Thus the variable $\mathbf{f}(t)$ is multiplied by \mathbf{a} as it flows to the node labeled $\dot{\mathbf{f}}(t)$. The symbol $1/s$ represents an integration operation on the variable $\dot{\mathbf{f}}$ (multiplication of the Laplace transform of $\dot{\mathbf{f}}$ by $1/s$ yields the Laplace transform of \mathbf{f}). Nodes are treated as summing points for all incoming signals. Thus the node labeled $\dot{\mathbf{f}}(t)$ is a graphic representation of the differential equation (4.91). The primary information about the system, the position of the pole, is contained in the feedback path. The signal flow graph focuses attention on the feedback nature of the system represented by the differential equation.

We can also obtain a transformed equation and a signal flow graph corresponding to the vector state equation (4.89). Suppose the state variables [or elements of $\mathbf{x}(t)$] are denoted by $\mathbf{f}_i(t)$, $i = 1, \dots, n$. Then we define the Laplace transform of the vector \mathbf{x} of (4.89) by

$$\mathbf{X} \triangleq \mathcal{L}(\mathbf{x}) \triangleq \begin{pmatrix} \mathcal{L}(\mathbf{f}_1) \\ \vdots \\ \mathcal{L}(\mathbf{f}_n) \end{pmatrix} \quad (4.93)$$

Exercise 1. Show that $\mathcal{L}(\mathbf{A}\mathbf{x}) = \mathbf{A} \mathcal{L}(\mathbf{x})$ for any $n \times n$ matrix \mathbf{A} .

Using definition (4.93) and Exercise 1, we take the Laplace transform of (4.89):

$$s\mathbf{X}(s) - \mathbf{x}(0) = \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s)$$

Solving for $\mathbf{X}(s)$, we obtain the following generalization of (4.92):

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}(0) + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s) \quad (4.94)$$

The matrix $(s\mathbf{I} - \mathbf{A})^{-1}$ is called the **matrix transfer function** for the system represented by (4.89). The poles of the transfer function are those values of

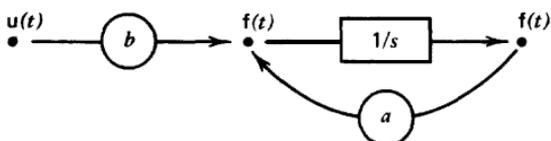


Figure 4.5. Signal flow graph for (4.91).

s for which $(s\mathbf{I}-\mathbf{A})$ is singular. Therefore, the poles of the system are the eigenvalues of the system matrix \mathbf{A} , a fact which we discovered for a restricted class of system matrices in (4.51). Because of the formal similarity between the results for the first-order system equation and for the n -dimensional state equation, we suspect that

$$\mathcal{L}(e^{\mathbf{A}t}) = (s\mathbf{I}-\mathbf{A})^{-1} \quad (4.95)$$

Equation (4.95) is easily verified by comparing (4.90) and (4.94). We can think of the state transition matrix $e^{\mathbf{A}t}$ as a **matrix impulse response** [see (3.77)-(3.78)]. The vector signal flow graph is formally the same as that for the scalar equation (Figure 4.6). However, individual nodes now represent vector variables. Again, the feedback nature of the system is emphasized by the flow graph model. The feedback path in Figure 4.6 contains all the information peculiar to the particular system, although the poles of the system are stated only implicitly as the eigenvalues of \mathbf{A} . The graph would be more specific if we were to use a separate node for each element of each vector variable; however, the diagram would be much more complicated. We draw such a detailed flow graph for a special case in Figure 4.8.

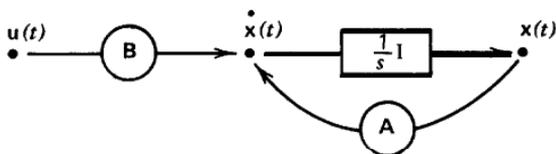


Figure 4.6. Vector signal flow graph for (4.89).

In order to obtain as much insight concerning the feedback nature of the state equation as we did for the scalar case, we change to a coordinate system which emphasizes the poles of the system. Let $\mathbf{x} = \mathbf{S}\mathbf{z}$, where \mathbf{S} is an invertible $n \times n$ matrix. Then $\mathbf{z}(t)$ describes the state of the system relative to a new set of coordinates, and (4.89) becomes

$$\dot{\mathbf{z}}(t) = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}\mathbf{z}(t) + \mathbf{S}^{-1}\mathbf{B}\mathbf{u}(t), \quad \mathbf{z}(0) = \mathbf{S}^{-1}\mathbf{x}(0) \text{ given} \quad (4.96)$$

We choose \mathbf{S} so that $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, the spectral matrix (or Jordan form) of \mathbf{A} . Thus \mathbf{S} consists in a basis for the state space composed of generalized eigenvectors of \mathbf{A} as in (4.76). The new signal flow graph is Figure 4.7.

In order to see that this new signal flow graph is particularly informative, we must examine the interconnections between the individual elements of $\mathbf{z}(t)$. We do so for a particular example.

In the flow graph of Figure 4.8 the vector system is viewed as a set of nearly uncoupled scalar systems. The poles of the system (the eigenvalues of \mathbf{A}) appear in the main feedback paths in the graph. The only other feedback paths are those corresponding to the off-diagonal 1's in \mathbf{A} . It is these off-diagonal 1's that give rise to nonexponential terms (te^{2t}) in the response of the system. Specifically, if the input function \mathbf{u} is zero,

$$\begin{aligned}\mathbf{v}_3(t) &= \mathbf{v}_3(0)e^{2t} \\ \mathbf{v}_2(t) &= \mathbf{v}_2(0)e^{2t} \\ \mathbf{v}_1(t) &= \mathbf{v}_1(0)e^{2t} + \mathbf{v}_2(0)te^{2t}\end{aligned}$$

The extra term in \mathbf{v}_1 arises because the scalar system which determines \mathbf{v}_1 is driven by \mathbf{v}_2 .

It is evident that the Jordan form of a system matrix is a convenient catalog of the information available concerning the system. The modal matrix \mathbf{S} describes the interconnections between the canonical variables and the state variables. Suppose the above system is undriven [$\mathbf{u}(t) = 0$] and the initial values of the canonical variables are $\mathbf{v}_1(0) = \mathbf{v}_2(0) = 0$ and $\mathbf{v}_3(0) = 1$. Then $\mathbf{v}_1(t) = \mathbf{v}_2(t) = 0$ and $\mathbf{v}_3(t) = e^{2t}$. The corresponding output vector $\mathbf{x}(t)$ is

$$\mathbf{x}(t) = \begin{pmatrix} \mathbf{f}_1(t) \\ \mathbf{f}_2(t) \\ \mathbf{f}_3(t) \end{pmatrix} = \begin{pmatrix} e^{2t} \\ e^{2t} \\ 0 \end{pmatrix} = e^{2t} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

At each instant, the output vector is proportional to the third column of \mathbf{S} , one of the eigenvectors of \mathbf{A} . Under these circumstances, we say only one "mode of response" of the system has been excited. There is one mode of response corresponding to each canonical variable; corresponding to the variable $\mathbf{v}_i(t)$ is the mode where $\mathbf{x}(t)$ is proportional to the i column of \mathbf{S} .

We call the system represented by (4.89) **controllable** if there is some input $\mathbf{u}(t)$ that will drive the system [$\mathbf{z}(t)$ or $\mathbf{x}(t)$] from one arbitrary state to another arbitrary state in a finite amount of time. It should be apparent from Example 2 that in order to be able to control all the canonical state variables in the system, the input variables must be coupled to the inputs of each chain in the flow graph, namely, $\dot{\mathbf{v}}_2(t)$ and $\dot{\mathbf{v}}_3(t)$ in Figure 4.8. If in the above example $\mathbf{B} = (0 \ 1 \ 0)^T$, $\mathbf{u}(t)$ is not coupled to (and has no influence on) $\mathbf{v}_3(t)$. On the other hand, if $\mathbf{B} = (1 \ 0 \ 0)^T$, the input is coupled to all the canonical state variables; the system appears to be controllable. However, the variables $\mathbf{v}_2(t)$ and $\mathbf{v}_3(t)$ respond identically to \mathbf{u} —they are associated with identical poles. As a result, $\mathbf{v}_2(t)$ and $\mathbf{v}_3(t)$ cannot be

controlled independently. In point of fact, we cannot consider the single input system of Example 2 fully controllable regardless of which input matrix \mathbf{B} we use. A system can be fully controlled only if we can influence identical subsystems independently. In Example 2, the use of a *pair* of inputs with the input matrix

$$\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

yields a controllable system.

In physical systems we may not be able to measure the state variables directly. Perhaps we can only measure variables $\{\mathbf{g}_i(t)\}$ which are related to the state variables by

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)$$

where $\mathbf{y}(t) = (\mathbf{g}_1(t) \cdots \mathbf{g}_p(t))^T$ and \mathbf{C} is $p \times n$. The matrix \mathbf{C} would appear in the flow graph of Figure 4.8 as a set of connections between the state variables $\{\mathbf{f}_i(t)\}$ and the output (or measurable) variables $\{\mathbf{g}_i(t)\}$. Clearly, we cannot fully determine the state of the system from the measurements unless the output variables are coupled to the output of each chain; namely, $\mathbf{v}_1(t)$ and $\mathbf{v}_3(t)$. Furthermore, in this specific example, measurement of a single output variable $\mathbf{g}_i(t)$ is not sufficient to distinguish between the variables $\mathbf{v}_2(t)$ and $\mathbf{v}_3(t)$, because their behavior is identical. In general, we call a system **observable** if by observing the output $\mathbf{y}(t)$ of the undriven system for a finite interval of time, we get enough information to determine the initial state $\mathbf{x}(0)$. See Brown [4.2] or Zadeh and Desoer [4.20] for convenient tests for controllability and observability.

4.6 Functions of Matrices and Linear Operators

In previous examples we have encountered several functions of square matrices; namely, \mathbf{A}^k , $e^{\mathbf{A}t}$, and $(s\mathbf{I} - \mathbf{A})^{-1}$. In later sections we encounter additional matrix functions. The actual computation of such functions of matrices is a problem of practical importance, especially in the analysis of dynamic systems. In this section we develop a definition for functions of matrices which applies in essentially all situations where we might expect such functions to be meaningful. The definition applies to diagonalizable and nondiagonalizable matrices, and also to the linear operators that these matrices represent. (Functions of diagonalizable linear operators on infinite-dimensional spaces are considered in Section 5.5.) Much of this section

is devoted to the development of techniques for analyzing and evaluating functions of matrices.

Two of the matrix functions mentioned above, \mathbf{A}^k and $(s\mathbf{I} - \mathbf{A})^{-1}$, are defined in terms of ordinary matrix operations—addition, scalar multiplication, and inversion. The third matrix function, $e^{\mathbf{A}t}$, represents the sum of an infinite polynomial series in \mathbf{A} , as defined in (3.72). This latter function suggests an approach to the definition of general functions of the square matrix \mathbf{A} . Polynomial functions of matrices are clearly defined; they can be evaluated by matrix multiplications and additions. Suppose the non-polynomial function f can be expanded in the power series*

$$f(\lambda) = \sum_{k=0}^{\infty} a_k \lambda^k$$

One reasonable way to define $f(\mathbf{A})$ is by using the same power series in \mathbf{A} ,

$$f(\mathbf{A}) \triangleq \sum_{k=0}^{\infty} a_k \mathbf{A}^k \quad (4.97)$$

Each term of the series can be evaluated using ordinary matrix operations. Of course, the definition (4.97) is useful only if the series converges and we can evaluate the sum of the series. We explore the question of convergence of (4.97) shortly. The essential properties of \mathbf{A} are displayed in its spectral matrix $\mathbf{\Lambda}$ and its modal matrix \mathbf{S} . Substituting $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$ into (4.97) we find

$$\begin{aligned} f(\mathbf{A}) &= f(\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}) \\ &= \sum_k a_k (\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1})^k \\ &= \sum_k a_k \mathbf{S}\mathbf{\Lambda}^k \mathbf{S}^{-1} \\ &= \mathbf{S} \left(\sum_k a_k \mathbf{\Lambda}^k \right) \mathbf{S}^{-1} \\ &= \mathbf{S} f(\mathbf{\Lambda}) \mathbf{S}^{-1} \end{aligned} \quad (4.98)$$

*The power series used in (4.97) is a Taylor series expansion about the origin. The matrix function could have been defined in terms of a Taylor series or Laurent series expansion about some other point in the complex plane. See Wylie [4.18] a discussion of such power series expansions.

(We are able to take the similarity transformation outside the infinite sum because matrix multiplication is a continuous operator; see Section 5.4.) Thus if $f(\mathbf{A})$ as given in (4.97) is well-defined, then evaluation of $f(\mathbf{A})$ reduces to evaluation of $f(\mathbf{\Lambda})$. We again apply the power series definition to determine $f(\mathbf{\Lambda})$. If \mathbf{A} is diagonalizable, then $\mathbf{\Lambda}$ is diagonal, and

$$\begin{aligned}
 f(\mathbf{\Lambda}) &= \sum_k a_k \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}^k \\
 &= \sum_k a_k \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{pmatrix} \\
 &= \begin{pmatrix} \sum_k a_k \lambda_1^k & & \\ & \ddots & \\ & & \sum_k a_k \lambda_n^k \end{pmatrix} \\
 &= \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_n) \end{pmatrix} \tag{4.99}
 \end{aligned}$$

On the other hand, if \mathbf{A} is not diagonalizable, $f(\mathbf{\Lambda})$ differs from (4.99) only as a result of the off-diagonal 1's in $\mathbf{\Lambda}$. By the same logic, we can express $f(\mathbf{\Lambda})$ as

$$f(\mathbf{\Lambda}) = \sum_k a_k \begin{pmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_r \end{pmatrix}^k = \begin{pmatrix} f(\mathbf{J}_1) & & \\ & \ddots & \\ & & f(\mathbf{J}_r) \end{pmatrix} \tag{4.100}$$

where \mathbf{J}_i is the i th Jordan block in $\mathbf{\Lambda}$. Thus calculation of $f(\mathbf{A})$ reduces to the determination of $f(\mathbf{J}_i)$.

We explore $f(\mathbf{J}_i)$ by means of an example. For a 4×4 Jordan block we have

$$\mathbf{J} = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} \quad \mathbf{J}^2 = \begin{pmatrix} \lambda^2 & 2\lambda & 1 & 0 \\ 0 & \lambda^2 & 2\lambda & 1 \\ 0 & 0 & \lambda^2 & 2\lambda \\ 0 & 0 & 0 & \lambda^2 \end{pmatrix}$$

$$\mathbf{J}^3 = \begin{pmatrix} \lambda^3 & 3\lambda^2 & 3\lambda & 1 \\ 0 & \lambda^3 & 3\lambda^2 & 3\lambda \\ 0 & 0 & \lambda^3 & 3\lambda^2 \\ 0 & 0 & 0 & \lambda^3 \end{pmatrix}, \quad \mathbf{J}^4 = \begin{pmatrix} \lambda^4 & 4\lambda^3 & 6\lambda^2 & 4\lambda \\ 0 & \lambda^4 & 4\lambda^3 & 6\lambda^2 \\ 0 & 0 & \lambda^4 & 4\lambda^3 \\ 0 & 0 & 0 & \lambda^4 \end{pmatrix}$$

Observe that in each matrix the element which appears on the j th “superdiagonal” is $(1/j!)$ times the j th derivative (with respect to λ) of the element on the main diagonal. Thus, continuing the example,

$$f(\mathbf{J}) = \sum_k a_k \mathbf{J}^k$$

$$= \begin{pmatrix} \sum_k a_k \lambda^k & \sum_k \frac{a_k}{1!} \frac{d\lambda^k}{d\lambda} & \sum_k \frac{a_k}{2!} \frac{d^2\lambda^k}{d\lambda^2} & \sum_k \frac{a_k}{3!} \frac{d^3\lambda^k}{d\lambda^3} \\ 0 & \sum_k a_k \lambda^k & \sum_k \frac{a_k}{1!} \frac{d\lambda^k}{d\lambda} & \sum_k \frac{a_k}{2!} \frac{d^2\lambda^k}{d\lambda^2} \\ 0 & 0 & \sum_k a_k \lambda^k & \sum_k \frac{a_k}{1!} \frac{d\lambda^k}{d\lambda} \\ 0 & 0 & 0 & \sum_k a_k \lambda^k \end{pmatrix}$$

Relying on the term-by-term differentiability of power series (Kaplan [4.11, p. 353]), we take all derivatives outside the summations to obtain

$$f(\mathbf{J}) = \begin{pmatrix} f(\lambda) & \frac{f'(\lambda)}{1!} & \frac{f''(\lambda)}{2!} & \frac{f^{(3)}(\lambda)}{3!} \\ 0 & f(\lambda) & \frac{f'(\lambda)}{1!} & \frac{f''(\lambda)}{2!} \\ 0 & 0 & f(\lambda) & \frac{f'(\lambda)}{1!} \\ 0 & 0 & 0 & f(\lambda) \end{pmatrix} \quad (4.101)$$

Corresponding to each Jordan block \mathbf{J}_i of \mathbf{A} (with eigenvalue λ_i), $f(\mathbf{A})$ contains a block which has $f(\lambda_i)$ on the main diagonal. The upper elements in the block are filled with appropriately scaled derivatives of f (evaluated at λ_i). The elements on the j th super-diagonal are

$$\frac{1}{j!} \frac{d^j f(\lambda_i)}{d\lambda^j}$$

Surprisingly, $f(\mathbf{A})$ is not in Jordan form.

Example 1. Matrix Inversion as a Matrix Function. Suppose $f(\lambda) = 1/\lambda$. If \mathbf{A} is an invertible $n \times n$ matrix, we use (4.98) and (4.99) to find

$$\begin{aligned} \mathbf{A}^{-1} &= \mathbf{S}\mathbf{\Lambda}^{-1}\mathbf{S}^{-1} \\ &= \mathbf{S} \begin{pmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_n \end{pmatrix} \mathbf{S}^{-1} \end{aligned}$$

Suppose

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Then $\mathbf{S} = \mathbf{S}^{-1} = \mathbf{I}$, and

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \infty \end{pmatrix}$$

It is clear that \mathbf{A}^{-1} does not exist if zero is an eigenvalue of \mathbf{A} . The function $1/\lambda$ is not defined at $\lambda = 0$, and (4.99) cannot be evaluated.

Example 2. A Function of a Nondiagonalizable Matrix. As in Example 1, if $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$, $\mathbf{A}^{-1} = \mathbf{S}\mathbf{\Lambda}^{-1}\mathbf{S}^{-1}$. Suppose

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 1 & 0 & \vdots \\ 0 & \lambda_1 & 1 & \vdots \\ 0 & 0 & \lambda_1 & \vdots \\ \vdots & \vdots & \vdots & \ddots \\ & & & \lambda_1 & \vdots \\ & & & & \lambda_2 \end{pmatrix}$$

Letting $f(\lambda) = 1/\lambda$ we find that $f'(\lambda) = -1/\lambda^2$ and $f''(\lambda)/2! = 1/\lambda^3$. Thus, using

(4.101) for each Jordan block,

$$\mathbf{A}^{-1} = \begin{pmatrix} 1/\lambda_1 & -1/\lambda_1^2 & 1/\lambda_1^3 & \vdots & & \\ 0 & 1/\lambda_1 & -1/\lambda_1^2 & \vdots & & \\ 0 & 0 & 1/\lambda_1 & \vdots & & \\ \dots & \dots & \dots & \dots & \dots & \\ & & & & 1/\lambda_1 & \vdots \\ & & & & \vdots & 1/\lambda_2 \end{pmatrix}$$

An Alternative Definition

Although we have used (4.97) to define $f(\mathbf{A})$, we have used (4.98) and (4.100) to perform the actual evaluation of $f(\mathbf{A})$. [Note that (4.99) is a special case of (4.100).] It can be shown that our original definition of $f(\mathbf{A})$, (4.97), converges if and only if f is analytic in a circle of the complex plane which contains all the eigenvalues of \mathbf{A} .^{*} Yet (4.98) and (4.100), which we derived from (4.97), provide a correct evaluation of $f(\mathbf{A})$ in cases which do not satisfy this criterion. For example,

$$\text{if } \mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \quad \text{then} \quad \mathbf{A}^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}$$

The function $f(\lambda) = \lambda^{-1}$ is not analytic at $\lambda = 0$. No circle encloses the points 2 and -2 while excluding the point 0, yet (4.98) and (4.99) provide the correct inverse. It is apparent that (4.98)-(4.101) provide a *more general definition* of $f(\mathbf{A})$ than does (4.97).

The definition (4.98)-(4.101) applies to all functions f and matrices \mathbf{A} for which $f(\mathbf{J}_i)$ can be evaluated for each Jordan block \mathbf{J}_i . If \mathbf{A} is diagonalizable, this evaluation requires only that f be *defined on the spectrum*; that is, that f be defined at all the eigenvalues of \mathbf{A} . If \mathbf{A} is not diagonalizable, the evaluation of $f(\mathbf{A})$ requires the existence of derivatives of f at some of the eigenvalues of \mathbf{A} . Thus the definition of $f(\mathbf{A})$ given in (4.98)-(4.101) certainly applies to all f and \mathbf{A} for which f is not only defined on the spectrum of \mathbf{A} but also analytic at those eigenvalues of \mathbf{A} for which \mathbf{A} is defective (i.e., for which the corresponding Jordan blocks \mathbf{J}_i are larger than 1×1). In every case where the definition (4.97) applies, the evaluation of $f(\mathbf{A})$ which results is identical to the evaluation provided by (4.98)-(4.101).

As illustrated in (4.101), the actual evaluation of $f(\mathbf{A})$ leads to evaluation of

$$f(\lambda_i), f'(\lambda_i), \dots, f^{(q_i-1)}(\lambda_i), \quad i = 1, \dots, p \quad (4.102)$$

^{*}Rinehart [4.14]. A function $f(\lambda)$ is said to be analytic at λ_1 if it is differentiable (as a function of a complex variable λ) in a neighborhood of λ_1 (see Wylie [4.18]).

We refer to this set of evaluations as **evaluation on the spectrum** of \mathbf{A} . It is apparent that any two functions that have the same evaluation on the spectrum lead to the same function of \mathbf{A} .

Exercise 1. Compare $f(\mathbf{A})$ and $g(\mathbf{A})$ for $f(\lambda) \triangleq 4\lambda - 8$, $g(\lambda) \triangleq \lambda^2 - 4$, and

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$$

Equations (4.98)-(4.101) provide a suitable definition of $f(\mathbf{A})$ for most choices of f and \mathbf{A} . Rinehart [4.14] shows that with this definition of $f(\mathbf{A})$ and with single-valued functions g and h for which $g(\mathbf{A})$ and $h(\mathbf{A})$ exist,

1. If $f(\lambda) = c$ then $f(\mathbf{A}) = c\mathbf{I}$
2. If $f(\lambda) = \lambda$ then $f(\mathbf{A}) = \mathbf{A}$
3. If $f(\lambda) = g(\lambda) + h(\lambda)$ then $f(\mathbf{A}) = g(\mathbf{A}) + h(\mathbf{A})$
4. If $f(\lambda) = g(\lambda) \cdot h(\lambda)$ then $f(\mathbf{A}) = g(\mathbf{A}) \cdot h(\mathbf{A})$
5. If $f(\lambda) = g(h(\lambda))$ then $f(\mathbf{A}) = g(h(\mathbf{A}))$

If g or h is not single valued, then the matrix $f(\mathbf{A})$ depends upon which branches of g and h are used in the evaluation on the spectrum of \mathbf{A} . From these properties it follows that scalar functional identities extend to matrices. For example, $\sin^2(\mathbf{A}) + \cos^2(\mathbf{A}) = \mathbf{I}$ and $e^{\ln \mathbf{A}} = \mathbf{A}$.

The Fundamental Formula for Matrices

Let \mathbf{A} be a 3×3 diagonalizable matrix with only two distinct eigenvalues; that is, $c(\lambda) = (\lambda - \lambda_1)^2(\lambda - \lambda_2)$, and the eigenspace for λ_1 is two-dimensional. Suppose also that the function f is defined at λ_1 and λ_2 . Then we can express $f(\mathbf{A})$ in the manner of Example 1:

$$f(\mathbf{A}) = \begin{pmatrix} f(\lambda_1) & 0 & 0 \\ 0 & f(\lambda_1) & 0 \\ 0 & 0 & f(\lambda_2) \end{pmatrix}$$

In order to express $f(\mathbf{A})$ in a manner that clearly separates the essential properties of \mathbf{A} from those off, we introduce the following notation. Let $\mathbf{E}_{i0}^{\mathbf{A}}$ be a matrix which has a one wherever $f(\mathbf{A})$ has $f(\lambda_i)$, and zeros elsewhere. (The second subscript, "0," is used only to provide consistency with the nondiagonalizable case introduced later.) Specifically,

$$\mathbf{E}_{10}^{\mathbf{A}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{E}_{20}^{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Then we can express $f(\mathbf{\Lambda})$ by

$$f(\mathbf{\Lambda}) = f(\lambda_1)\mathbf{E}_{10}^{\mathbf{\Lambda}} + f(\lambda_2)\mathbf{E}_{20}^{\mathbf{\Lambda}}$$

Since $f(\mathbf{A}) = \mathbf{S}f(\mathbf{\Lambda})\mathbf{S}^{-1}$ to obtain $f(\mathbf{A})$ we simply perform the similarity transformations $\mathbf{E}_{i0}^{\mathbf{A}} = \mathbf{S}\mathbf{E}_{i0}^{\mathbf{\Lambda}}\mathbf{S}^{-1}$ to obtain

$$f(\mathbf{A}) = f(\lambda_1)\mathbf{E}_{10}^{\mathbf{A}} + f(\lambda_2)\mathbf{E}_{20}^{\mathbf{A}}$$

It is evident that we can express any well-defined function of the specific matrix \mathbf{A} by means of this formula. Once we have the matrices $\mathbf{E}_{i0}^{\mathbf{A}}$, evaluation of $f(\mathbf{A})$ requires only evaluation of f on the spectrum of \mathbf{A} . By a derivation similar to that above, we can show that for any $n \times n$ diagonalizable matrix \mathbf{A} and any f defined on the spectrum of \mathbf{A} , $f(\mathbf{A})$ can be expressed as

$$f(\mathbf{A}) = \sum_{i=1}^p f(\lambda_i)\mathbf{E}_{i0}^{\mathbf{A}} \quad (4.103)$$

where p is the number of distinct eigenvalues of \mathbf{A} . We call (4.103) the **fundamental formula for $f(\mathbf{A})$** . The matrices $\mathbf{E}_{i0}^{\mathbf{A}}$ are called the **constituent matrices** (or **components**) of \mathbf{A} . (We drop the superscript \mathbf{A} when confusion seems unlikely.) Notice that (4.103) separates the contributions of f and \mathbf{A} . In fact, (4.103) is a satisfactory definition of $f(\mathbf{A})$, equivalent to (4.98-4.99).

The definition of the fundamental formula (4.103) can be extended to nondiagonalizable matrices as well. Suppose f is analytic at λ_1 and defined at λ_2 . Then we can write $f(\mathbf{\Lambda})$ for the matrix $\mathbf{\Lambda}$ of Example 2 as

$$f(\mathbf{\Lambda}) = \begin{pmatrix} f(\lambda_1) & f'(\lambda_1) & \frac{f''(\lambda_1)}{2} & \vdots \\ 0 & f(\lambda_1) & f'(\lambda_1) & \vdots \\ 0 & 0 & f(\lambda_1) & \vdots \\ \vdots & \vdots & \vdots & \ddots \\ & & & f(\lambda_1) & \vdots \\ & & & \vdots & f(\lambda_2) \end{pmatrix}$$

In order to separate the essential properties of $\mathbf{\Lambda}$ from those off, we define $\mathbf{E}_{ik}^{\mathbf{\Lambda}}$ to be a matrix which has a one wherever $f(\mathbf{\Lambda})$ has $(1/k!)f^{(k)}(\lambda_i)$, $k=0$,

1, and 2, and zeros elsewhere. Thus

$$\mathbf{E}_{10}^\Lambda = \begin{pmatrix} 1 & 0 & 0 & \vdots \\ 0 & 1 & 0 & \vdots \\ 0 & 0 & 1 & \vdots \\ \dots & \dots & \dots & \dots \\ & & & 1 \\ & & & \vdots \\ & & & 0 \end{pmatrix} \quad \mathbf{E}_{11}^\Lambda = \begin{pmatrix} 0 & 1 & 0 & \vdots \\ 0 & 0 & 1 & \vdots \\ 0 & 0 & 0 & \vdots \\ \dots & \dots & \dots & \dots \\ & & & 0 \\ & & & \vdots \\ & & & 0 \end{pmatrix}$$

$$\mathbf{E}_{12}^\Lambda = \begin{pmatrix} 0 & 0 & 1 & \vdots \\ 0 & 0 & 0 & \vdots \\ 0 & 0 & 0 & \vdots \\ \dots & \dots & \dots & \dots \\ & & & 0 \\ & & & \vdots \\ & & & 0 \end{pmatrix} \quad \mathbf{E}_{20}^\Lambda = \begin{pmatrix} 0 & 0 & 0 & \vdots \\ 0 & 0 & 0 & \vdots \\ 0 & 0 & 0 & \vdots \\ \dots & \dots & \dots & \dots \\ & & & 0 \\ & & & \vdots \\ & & & 1 \end{pmatrix}$$

Then we can express $f(\mathbf{A})$ by

$$f(\mathbf{A}) = f(\lambda_1)\mathbf{E}_{10}^\Lambda + f'(\lambda_1)\mathbf{E}_{11}^\Lambda + \frac{f''(\lambda_1)}{2!}\mathbf{E}_{12}^\Lambda + f(\lambda_2)\mathbf{E}_{20}^\Lambda$$

As in the diagonalizable case, we perform the similarity transformations $\mathbf{E}_{ij}^\Lambda = \mathbf{S}\mathbf{E}_{ij}^\Lambda\mathbf{S}^{-1}$ to obtain

$$f(\mathbf{A}) = f(\lambda_1)\mathbf{E}_{10}^\Lambda + f'(\lambda_1)\mathbf{E}_{11}^\Lambda + \frac{f''(\lambda_1)}{2!}\mathbf{E}_{12}^\Lambda + f(\lambda_2)\mathbf{E}_{20}^\Lambda$$

We can compute any well-defined function of the matrix \mathbf{A} of Example 2 by means of this formula. By a derivation similar to that above, we can show that for any $n \times n$ matrix \mathbf{A} and any f which is defined on the spectrum of \mathbf{A} and analytic at eigenvalues where \mathbf{A} is defective, $f(\mathbf{A})$ can be expressed as

$$f(\mathbf{A}) = \sum_{i=1}^p \left[f(\lambda_i)\mathbf{E}_{i0}^\Lambda + \frac{f'(\lambda_i)}{1!}\mathbf{E}_{i1}^\Lambda + \dots + \frac{f^{(q_i-1)}(\lambda_i)}{(q_i-1)!}\mathbf{E}_{i(q_i-1)}^\Lambda \right] \quad (4.104)$$

where p is the number of distinct eigenvalues of \mathbf{A} , and q_i is the index of annihilation for λ_i [see (4.66) and (4.67)]. Equation (4.104) is the general form of the **fundamental formula** for $f(\mathbf{A})$. Again, we refer to the matrices \mathbf{E}_{ij}^Λ as constituent matrices (or components) of \mathbf{A} .*

*The constituent matrices are sometimes defined as $\mathbf{E}_{ij}^\Lambda/j!$.

The fundamental formula can be used to generate a **spectral decomposition of \mathbf{A}** . If we let $f(\lambda) = \lambda$ in (4.104), we obtain

$$\mathbf{A} = \sum_{i=1}^p (\lambda_i \mathbf{E}_{i0}^{\mathbf{A}} + \mathbf{E}_{i1}^{\mathbf{A}}) \quad (4.105)$$

If \mathbf{A} is diagonalizable, $q_i = 1$ for each i , and (4.105) becomes

$$\mathbf{A} = \sum_{i=1}^p \lambda_i \mathbf{E}_{i0}^{\mathbf{A}}$$

It is apparent that in the diagonalizable case $\mathbf{E}_{i0}^{\mathbf{A}}$ describes the projection onto the eigenspace associated with λ_i . That is, if $\mathbf{x} = \mathbf{x}_1 + \cdots + \mathbf{x}_p$, where \mathbf{x}_i is the component of \mathbf{x} in the eigenspace for λ_i , then $\mathbf{x}_i = \mathbf{E}_{i0}^{\mathbf{A}} \mathbf{x}$ and \mathbf{A} acts like λ_i on \mathbf{x}_i . In the nondiagonalizable case, $\mathbf{E}_{i0}^{\mathbf{A}}$ describes the projection onto the *generalized* eigenspace for λ_i . Furthermore, $\mathbf{E}_{ik}^{\mathbf{A}}$ acts like the nilpotent operator $(\mathbf{A} - \lambda_i \mathbf{I})^k$ on the generalized eigenspace for λ_i ; that is, $\mathbf{E}_{ik}^{\mathbf{A}} = (\mathbf{A} - \lambda_i \mathbf{I})^k \mathbf{E}_{i0}^{\mathbf{A}}$.

Exercise 2. Verify that the matrices $\mathbf{E}_{10}^{\mathbf{A}}$ and $\mathbf{E}_{20}^{\mathbf{A}}$ of Example 2 satisfy the properties (4.3) for projectors. Show also that $\mathbf{E}_{ik}^{\mathbf{A}} = (\mathbf{A} - \lambda_i \mathbf{I})^k \mathbf{E}_{i0}^{\mathbf{A}}$.

Functions of Linear Operators

The fundamental formula also serves to define functions of the underlying operator represented by \mathbf{A} . If \mathbf{T} operates on an n -dimensional vector space \mathfrak{V} , if \mathbf{P}_{i0} is the operator which projects onto \mathfrak{W}_i (the generalized eigenspace for λ_i) along $\sum_{j \neq i} \mathfrak{W}_j$, and if $\mathbf{P}_{ik} \triangleq (\mathbf{T} - \lambda_i \mathbf{I})^k \mathbf{P}_{i0}$, then the **fundamental formula for $f(\mathbf{T})$** is

$$f(\mathbf{T}) \triangleq \sum_{i=1}^p \left[f(\lambda_i) \mathbf{P}_{i0} + \frac{1}{1!} f'(\lambda_i) \mathbf{P}_{i1} + \cdots + \frac{1}{(q_i - 1)!} f^{(q_i - 1)}(\lambda_i) \mathbf{P}_{i(q_i - 1)} \right] \quad (4.106)$$

If \mathfrak{X} is a basis for \mathfrak{V} and we define $\mathbf{A} \triangleq [\mathbf{T}]_{\mathfrak{X}\mathfrak{X}}$, then $\mathbf{E}_{ij}^{\mathbf{A}} = [\mathbf{P}_{ij}]_{\mathfrak{X}\mathfrak{X}}$. As a result, (4.104) and (4.106) require that $[f(\mathbf{T})]_{\mathfrak{X}\mathfrak{X}} = f([\mathbf{T}]_{\mathfrak{X}\mathfrak{X}})$. For diagonalizable \mathbf{T} (\mathbf{T} for which there exists a basis for \mathfrak{V} composed of eigenvectors for \mathbf{T}), (4.106) simplifies to $f(\mathbf{T}) = \sum_{i=1}^p f(\lambda_i) \mathbf{P}_{i0}$. This simple result is extended to certain infinite-dimensional operators in (5.90).

Example 3. A Function of a Linear Operator. Consider $\mathbf{D}: \mathscr{P}^3 \rightarrow \mathscr{P}^3$. We first find the eigendata for \mathbf{D} (as an operator on \mathscr{P}^3). The set $\mathscr{R} \triangleq \{\mathbf{f}_i(t) = t^{i-1}, i = 1, 2, 3\}$ is a basis for \mathscr{P}^3 . In Example 2 of Section 2.5 we found that

$$\mathbf{A} = [\mathbf{D}]_{\mathscr{R}\mathscr{R}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

This matrix has only one eigenvalue, $\lambda_1 = 0$; a basis of generalized eigenvectors for $[\mathbf{D}]_{\mathscr{R}\mathscr{R}}$ is

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{x}_{12} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_{13} = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \end{pmatrix}$$

Thus

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}, \quad \text{and} \quad \mathbf{S}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The generalized eigenfunctions of \mathbf{D} corresponding to \mathbf{x}_1 , \mathbf{x}_{12} , and \mathbf{x}_{13} are

$$\mathbf{g}_1(t) = 1, \quad \mathbf{g}_{12}(t) = t, \quad \mathbf{g}_{13}(t) = \frac{t^2}{2}$$

Because the chain of generalized eigenvectors is of length 3, $q_1 = 3$. Therefore, in order to evaluate $\mathbf{f}(\mathbf{D})$, we must determine three operators: \mathbf{P}_{10} , \mathbf{P}_{11} , and \mathbf{P}_{12} . Since the generalized eigenspace of \mathbf{D} for $\lambda_1 = 0$ is the whole space \mathscr{P}^3 , the projector \mathbf{P}_{10} onto the generalized eigenspace for λ_1 is $\mathbf{P}_{10} = \mathbf{I}$. We find the other two operators by

$$\mathbf{P}_{11} = (\mathbf{D} - \lambda_1 \mathbf{I})\mathbf{P}_{10} = \mathbf{D}\mathbf{I} = \mathbf{D}$$

$$\mathbf{P}_{12} = (\mathbf{D} - \lambda_1 \mathbf{I})^2 \mathbf{P}_{10} = \mathbf{D}^2 \mathbf{I} = \mathbf{D}^2$$

By (4.106), if f is analytic at $\lambda = 0$,

$$f(\mathbf{D}) = f(0)\mathbf{I} + f'(0)\mathbf{D} + \frac{f''(0)}{2}\mathbf{D}^2$$

Let $f(\lambda) = \lambda$. Then $f(\mathbf{D})$ reduces to

$$\mathbf{D} = (0)\mathbf{I} + (1)\mathbf{D} + (0)\mathbf{D}^2$$

which verifies the formula for $f(\mathbf{D})$. Let $f(\lambda) = e^\lambda$. Then

$$\begin{aligned} e^{\mathbf{D}} &= e^0 \mathbf{I} + e^0 \mathbf{D} + \frac{1}{2} e^0 \mathbf{D}^2 \\ &= \mathbf{I} + \mathbf{D} + \frac{1}{2} \mathbf{D}^2 \end{aligned}$$

Returning to $\mathbf{A} = [\mathbf{D}]_{\mathcal{R}\mathcal{R}}$, we generate those functions of \mathbf{A} which correspond to the functions $f(\mathbf{D})$, \mathbf{D} , and $e^{\mathbf{D}}$ above. By inspection of $\mathbf{\Lambda}$ we find that

$$\mathbf{E}_{10}^{\mathbf{A}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{E}_{11}^{\mathbf{A}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_{12}^{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Using the similarity transformation $\mathbf{E}_{ij}^{\mathbf{A}} = \mathbf{S}\mathbf{E}_{ij}^{\mathbf{\Lambda}}\mathbf{S}^{-1}$, we obtain

$$\mathbf{E}_{10}^{\mathbf{A}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{I}, \quad \mathbf{E}_{11}^{\mathbf{A}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{A}, \quad \mathbf{E}_{12}^{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{A}^2$$

These constituents of \mathbf{A} are $[\mathbf{P}_{10}]_{\mathcal{R}\mathcal{R}}$, $[\mathbf{P}_{11}]_{\mathcal{R}\mathcal{R}}$, and $[\mathbf{P}_{12}]_{\mathcal{R}\mathcal{R}}$, respectively. By (4.104),

$$f(\mathbf{A}) = f(0)\mathbf{I} + f'(0)\mathbf{A} + \frac{f''(0)}{2}\mathbf{A}^2$$

If $f(\lambda) = \lambda$, we find

$$\mathbf{A} = (0)\mathbf{I} + (1)\mathbf{A} + (0)\mathbf{A}^2$$

Let $f(\lambda) = e^{\lambda}$. Then

$$\begin{aligned} e^{\mathbf{A}} &= e^0\mathbf{I} + e^0\mathbf{A} + \frac{1}{2}e^0\mathbf{A}^2 \\ &= \mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 \\ &= \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

We easily verify that $e^{\mathbf{A}} = [e^{\mathbf{A}}]_{\mathcal{R}\mathcal{R}}$. These results are consistent with the definition (3.72) of $e^{\mathbf{A}^t}$, because $\mathbf{A}^k = \mathbf{0}$ for $k > 2$.

Computation of Functions of Matrices

We have already derived a method for computing $f(\mathbf{A})$ which relies on a complete eigenvalue analysis of \mathbf{A} . We summarize the method.

Computation of $f(\mathbf{A})$ by eigenvalue analysis of \mathbf{A} (4.107)

1. Determine the Jordan form $\mathbf{\Lambda}$, the modal matrix \mathbf{S} , and \mathbf{S}^{-1} such that $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$.

2. Determine $\mathbf{E}_{ij}^{\mathbf{\Lambda}}$ by inspection of $\mathbf{\Lambda}$.

3. Determine $\mathbf{E}_{ij}^{\mathbf{A}}$ by the similarity transformation $\mathbf{E}_{ij}^{\mathbf{A}} = \mathbf{S}\mathbf{E}_{ij}^{\mathbf{\Lambda}}\mathbf{S}^{-1}$.

4. Evaluate f on the spectrum of \mathbf{A} .

5. Determine $f(\mathbf{A})$ from the fundamental formula, (4.103) or (4.104).

Example 4. Computing $e^{\mathbf{A}t}$ Using Complete Eigenvalue Analysis Let $f(\lambda) = e^{\lambda t}$. Let \mathbf{A} be the matrix of Example 2, Section 4.5:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ -1 & 1 & 2 \end{pmatrix}$$

Then $f(\mathbf{A}) = e^{\mathbf{A}t}$ is the state transition matrix for that example. We found in that example that

$$(1) \quad \mathbf{\Lambda} = \begin{pmatrix} 2 & 1 & \vdots & 0 \\ 0 & \dots & 2 & \vdots & 0 \\ 0 & 0 & \vdots & 2 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{S}^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

Following the other steps outlined above,

$$(2) \quad \mathbf{E}_{10}^{\Lambda} = \begin{pmatrix} 1 & 0 & \vdots & 0 \\ 0 & \dots & 1 & \vdots & 0 \\ 0 & 0 & \vdots & 1 \end{pmatrix}, \quad \mathbf{E}_{11}^{\Lambda} = \begin{pmatrix} 0 & 1 & \vdots & 0 \\ 0 & \dots & 0 & \vdots & 0 \\ 0 & 0 & \vdots & 0 \end{pmatrix}$$

$$(3) \quad \mathbf{E}_{10}^{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{E}_{11}^{\Lambda} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \end{pmatrix}$$

$$(4) \quad f(2) = e^{2t}, \quad f'(2) = te^{2t}$$

$$(5) \quad e^{\mathbf{A}t} = e^{2t} \mathbf{E}_{10}^{\Lambda} + te^{2t} \mathbf{E}_{11}^{\Lambda} \\ = \begin{pmatrix} e^{2t} - te^{2t} & te^{2t} & 0 \\ -te^{2t} & e^{2t} + te^{2t} & 0 \\ -te^{2t} & te^{2t} & e^{2t} \end{pmatrix}$$

Determination of $f(\mathbf{A})$ using complete eigenvalue analysis is lengthy and computationally expensive. The eigenvalue analysis serves only to determine constituents of \mathbf{A} . [Of course, it provides considerable insight into the structure of the matrix \mathbf{A} in addition to producing $f(\mathbf{A})$]. We can eliminate most of this computation by employing the fundamental formula in evaluating the constituents. If we substitute several different functions into (4.103)-(4.104), we obtain several equations involving the constituents as unknowns. By a judicious choice of functions, we can obtain equations that allow us to determine each constituent independently. If the minimal polynomial $m(\lambda)$ is evaluated on the spectrum, the evaluations are all zero. If one factor is cancelled from $m(\lambda)$ and the resulting polynomial evaluated on the spectrum, precisely one evaluation is nonzero; if we evaluate this same polynomial in \mathbf{A} , precisely one constituent will remain in the fundamental formula. By successively cancelling factors from $m(\lambda)$, and evaluat-

ing the resulting polynomials in \mathbf{A} , we obtain the constituents in an efficient manner.*

Computation of $f(\mathbf{A})$ by evaluating factors of $m(\lambda)$ (4.108)

1. Find and factor $m(\lambda)$, the minimal polynomial for \mathbf{A} .
2. Cancel one factor from $m(\lambda)$. Denote the resulting polynomial $g_1(\lambda)$. Evaluating $g_1(\mathbf{A})$ will determine precisely one constituent matrix.
3. Cancel an additional factor from $m(\lambda)$. Let $g_i(\lambda)$ denote the polynomial which results from cancelling i factors from $m(\lambda)$. Evaluation of $g_i(\mathbf{A})$ determines precisely one constituent matrix in terms of previously determined constituents. This step is repeated until all the constituents E_{ij}^\wedge are known.
4. Evaluate f on the spectrum of \mathbf{A} .
5. Compute $f(\mathbf{A})$ from the fundamental formula, (4.103) or (4.104).

Example 5. Computing $e^{\mathbf{A}t}$ by Evaluating Factors of the Minimal Polynomial. Let $f(\lambda) = e^{\lambda t}$. Assume \mathbf{A} is the matrix given in Example 4. We compute the state transition matrix $e^{\mathbf{A}t}$ by the steps outlined above:

1. The characteristic polynomial for \mathbf{A} is $c(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - 2)^3$. The only eigenvalue is $\lambda_1 = 2$. By investigating the nullities of $(\mathbf{A} - 2\mathbf{I})$ and $(\mathbf{A} - 2\mathbf{I})^2$, we find that $q_1 = 2$ and $m(\lambda) = (\lambda - 2)^2$. Thus

$$f(\mathbf{A}) = f(2)\mathbf{E}_{10}^\wedge + f'(2)\mathbf{E}_{11}^\wedge$$

2. $g_1(\lambda) = (\lambda - 2)$ and $g_1'(\lambda) = 1$. Therefore,

$$\begin{aligned} g_1(\mathbf{A}) &\triangleq (\mathbf{A} - 2\mathbf{I}) = g_1(2)\mathbf{E}_{10}^\wedge + g_1'(2)\mathbf{E}_{11}^\wedge \\ &= (0)\mathbf{E}_{10}^\wedge + (1)\mathbf{E}_{11}^\wedge \end{aligned}$$

and $\mathbf{E}_{11}^\wedge = \mathbf{A} - 2\mathbf{I}$.

3. $g_2(\lambda) = 1$ and $g_2'(\lambda) = 0$. Then,

$$\begin{aligned} g_2(\mathbf{A}) &\triangleq \mathbf{I} = g_2(2)\mathbf{E}_{10}^\wedge + g_2'(2)\mathbf{E}_{11}^\wedge \\ &= (1)\mathbf{E}_{10}^\wedge + (0)\mathbf{E}_{11}^\wedge \end{aligned}$$

and $\mathbf{E}_{10}^\wedge = \mathbf{I}$.

4. $f(2) = e^{2t}$, and $f'(2) = te^{2t}$.

5. $e^{\mathbf{A}t} = e^{2t}\mathbf{I} + te^{2t}(\mathbf{A} - 2\mathbf{I})$

$$= \begin{pmatrix} e^{2t} - te^{2t} & te^{2t} & 0 \\ -te^{2t} & e^{2t} + te^{2t} & 0 \\ -te^{2t} & te^{2t} & e^{2t} \end{pmatrix}$$

*From Zadeh and Desoer [4.20].

Evaluating factors of $m(\lambda)$ is probably the most efficient known method for computing $f(\mathbf{A})$. A suitable sequence of functions can also be obtained by successively cancelling factors from the characteristic polynomial $c(\lambda)$, thereby avoiding determination of the nullities of powers of $(\mathbf{A} - \lambda_i \mathbf{I})$. If $c(\lambda)$ had been used in Example 5, we would have found that $\mathbf{E}_{12}^{\mathbf{A}} = \mathbf{0}$.

From our computation of $f(\mathbf{A})$ by evaluating factors of the minimal polynomial, we recognize that each of the constituents $\mathbf{E}_{ij}^{\mathbf{A}}$ equals a polynomial in \mathbf{A} ; the order of the polynomial is, in each case, less than that of the minimal polynomial. Therefore, by the fundamental formula, $f(\mathbf{A})$ is also equal to a polynomial in \mathbf{A} . Since powers of \mathbf{A} , and thus polynomials in \mathbf{A} , commute with each other, functions of \mathbf{A} commute with each other also. See P&C 4.29 for properties of commuting matrices. Additional techniques for computing $f(\mathbf{A})$ are given in P&C 4.25-4.27.

Application of Functions of Matrices—Modes of Oscillation

Figure 4.9 is an idealized one-dimensional representation of a piece of spring-mounted equipment. The variables \mathbf{v}_1 , \mathbf{v}_2 , and \mathbf{u} represent the positions, relative to their respective references, of the two identical masses (labeled m) and the frame which holds the equipment. The three springs have identical spring constants k . We treat the position (or vibration) of the frame as an independent variable; we seek the motions, $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$, of the spring-mounted objects. The dynamic equations which describe these motions are

$$\begin{aligned} m\ddot{\mathbf{v}}_1(t) &= -2k\mathbf{v}_1(t) + k\mathbf{v}_2(t) + k\mathbf{u}(t) \\ m\ddot{\mathbf{v}}_2(t) &= k\mathbf{v}_1(t) - 2k\mathbf{v}_2(t) + k\mathbf{u}(t) \end{aligned} \tag{4.109}$$

We could convert (4.109) to a four-dimensional first-order state equation. However, emboldened by the formal analogy which we found between the solution to the state equation and its scalar counterpart, we develop a second-order vector equation which is equivalent to (4.109) and which keeps explicit the second-order nature of the individual equations.

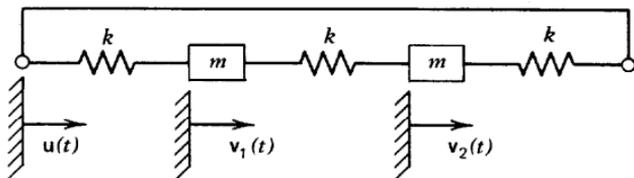


Figure 4.9. A model for spring-mounted equipment.

Let $\mathbf{x} = (\mathbf{v}_1 \ \mathbf{v}_2)^T$. Then (4.109) becomes

$$\ddot{\mathbf{x}}(t) + \begin{pmatrix} 2k/m & -k/m \\ -k/m & 2k/m \end{pmatrix} \mathbf{x}(t) = \begin{pmatrix} k/m \\ k/m \end{pmatrix} \mathbf{u}(t) \quad (4.110)$$

The 2×2 matrix in (4.110) is known as the **stiffness matrix** for the system. Equation (4.110) is a special case of the general vector equation

$$\ddot{\mathbf{x}}(t) + \mathbf{A}\mathbf{x}(t) = \mathbf{B}\mathbf{u}(t) \quad (4.111)$$

where $\mathbf{x}(t)$ is $n \times 1$, $\mathbf{u}(t)$ is $m \times 1$, \mathbf{B} is $n \times m$, and \mathbf{A} is an $n \times n$ diagonalizable matrix with positive eigenvalues.* Equation (4.111) is a convenient way to express many conservative systems; for example, a frictionless mechanical system which contains n masses coupled by springs; or a lossless electrical network containing interconnected inductors and capacitors. We solve (4.110) and (4.111) by analogy with the scalar case.

The scalar counterpart of (4.111) is

$$\ddot{\mathbf{f}}(t) + \omega^2 \mathbf{f}(t) = \mathbf{u}(t) \quad (4.112)$$

We found in P&C 3.6 that the inverse of (4.112), in terms of the initial conditions $\mathbf{f}(0)$ and $\dot{\mathbf{f}}(0)$, is

$$\mathbf{f}(t) = \mathbf{f}(0) \cos \omega t + \frac{\dot{\mathbf{f}}(0)}{\omega} \sin \omega t + \int_0^t \frac{\sin \omega(t-s)}{\omega} \mathbf{u}(s) ds \quad (4.113)$$

The solution consists in an undamped oscillation of frequency ω plus a term affected by the input vibration \mathbf{u} .

Comparing (4.111) and (4.112), we recognize that \mathbf{x} is the vector analog of \mathbf{f} , and \mathbf{A} plays the same role as ω^2 . Therefore, we expect the solution to (4.111) to be

$$\begin{aligned} \mathbf{x}(t) = & \cos(\sqrt{\mathbf{A}} t) \mathbf{x}(0) + (\sqrt{\mathbf{A}})^{-1} \sin(\sqrt{\mathbf{A}} t) \dot{\mathbf{x}}(0) \\ & + \int_0^t (\sqrt{\mathbf{A}})^{-1} \sin[\sqrt{\mathbf{A}}(t-s)] \mathbf{B}\mathbf{u}(s) ds \end{aligned} \quad (4.114)$$

By $\sqrt{\mathbf{A}}$ we mean any matrix whose square equals \mathbf{A} . As with the scalar square root, $\sqrt{\mathbf{A}}$ is not unique. The fundamental formula (4.103) indicates that $\sqrt{\mathbf{A}}$ depends on the square roots of the eigenvalues of \mathbf{A} . We use in

*The matrix \mathbf{A} is symmetric and positive definite. Such a matrix necessarily has positive real eigenvalues. See P&C 5.9 and 5.28.

(4.114) the principal square root of \mathbf{A} —the one involving positive square roots of the eigenvalues (P&C 4.28). Recall from the discussion following Example 5 that functions of \mathbf{A} commute with each other; the order of multiplication of $(\sqrt{\mathbf{A}})^{-1}$ and $\sin(\sqrt{\mathbf{A}} t)$ is arbitrary.

Equation (4.114) can be derived by finding a matrix Green's function and matrix boundary kernel for (4.111) (P&C 4.32). Or it can be verified by showing that it is a solution to the differential equation (4.111).

Exercise 3. Verify (4.114) by substituting $\mathbf{x}(t)$ into (4.111). Hint:

$$\frac{d}{dt} f(\mathbf{A}t) = \mathbf{A}f(\mathbf{A}t) \quad (\text{P\&C 4.30})$$

$$\frac{d}{dt} \int_a^t g(t,s) ds = \int_a^t \frac{\partial}{\partial t} g(t,s) ds + g(t,t)$$

We now evaluate the solution (4.114) for the specific case (4.110) using the techniques derived for determining functions of matrices.

Exercise 4. Show that the eigendata for the 2×2 stiffness matrix \mathbf{A} of (4.110) are

$$\lambda_1 = \frac{k}{m}, \quad \lambda_2 = \frac{3k}{m}, \quad \mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Exercise 5. Show that for \mathbf{A} of (4.110),

$$f(\mathbf{A}) = f\left(\frac{k}{m}\right) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} + f\left(\frac{3k}{m}\right) \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

It follows from Exercise 5 that

$$\cos \sqrt{\mathbf{A}} t = \begin{pmatrix} \frac{\cos \sqrt{k/m} t + \cos \sqrt{3k/m} t}{2} & \frac{\cos \sqrt{k/m} t - \cos \sqrt{3k/m} t}{2} \\ \frac{\cos \sqrt{k/m} t - \cos \sqrt{3k/m} t}{2} & \frac{\cos \sqrt{k/m} t + \cos \sqrt{3k/m} t}{2} \end{pmatrix}$$

$$(\sqrt{\mathbf{A}})^{-1} \sin \sqrt{\mathbf{A}} t = \begin{pmatrix} \frac{\sin \sqrt{k/m} t}{2\sqrt{k/m}} + \frac{\sin \sqrt{3k/m} t}{2\sqrt{3k/m}} & \frac{\sin \sqrt{k/m} t}{2\sqrt{k/m}} - \frac{\sin \sqrt{3k/m} t}{2\sqrt{3k/m}} \\ \frac{\sin \sqrt{k/m} t}{2\sqrt{k/m}} - \frac{\sin \sqrt{3k/m} t}{2\sqrt{3k/m}} & \frac{\sin \sqrt{k/m} t}{2\sqrt{k/m}} + \frac{\sin \sqrt{3k/m} t}{2\sqrt{3k/m}} \end{pmatrix}$$

$$(\sqrt{\mathbf{A}})^{-1} \sin[\sqrt{\mathbf{A}}(t-s)] \mathbf{B} = \sqrt{k/m} \sin \sqrt{k/m} (t-s) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

These three matrices can be substituted into (4.114) to obtain $\mathbf{x}(t)$ explicitly as a complicated function of the input data $\mathbf{u}(t)$, $\mathbf{x}(0)$, and $\dot{\mathbf{x}}(0)$.

Even though the general form of $\mathbf{x}(t)$ is complicated, we can provide a simple physical interpretation of the eigendata of the stiffness matrix of (4.110). Let $\mathbf{x}(0) = \mathbf{x}_1$, $\dot{\mathbf{x}}(0) = \mathbf{0}$, and $\mathbf{u}(t) = \mathbf{0}$. Then recalling that \mathbf{A} and $f(\mathbf{A})$ have the same eigenvectors,

$$\mathbf{x}(t) \triangleq \begin{pmatrix} \mathbf{v}_1(t) \\ \mathbf{v}_2(t) \end{pmatrix} = \cos \sqrt{\mathbf{A}} t \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \cos \sqrt{k/m} t \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The first eigenvector initial condition excites a sinusoidal oscillation of frequency $\sqrt{k/m} = \sqrt{\lambda_1}$. In this first mode of oscillation, both masses move together—the center spring is not stressed. The system acts like a single mass with a spring-mass ratio of $2k/2m = k/m = \lambda_1$. A second mode of oscillation can be excited by the conditions $\mathbf{x}(0) = \mathbf{x}_2$, $\dot{\mathbf{x}}(0) = \mathbf{0}$, $\mathbf{u}(t) = \mathbf{0}$;

$$\mathbf{x}(t) \triangleq \begin{pmatrix} \mathbf{v}_1(t) \\ \mathbf{v}_2(t) \end{pmatrix} = \cos \sqrt{\mathbf{A}} t \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \cos \sqrt{3k/m} t \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The second eigenvector initial condition excites a sinusoidal oscillation of frequency $\sqrt{3k/m} = \sqrt{\lambda_2}$. In this mode of oscillation, the masses move in opposite directions—the midpoint of the center spring does not move. The system acts like a pair of mirror images, each with a spring-mass ratio of $(k + 2k)/m = 3k/m = \lambda_2$. Thus the eigenvectors and eigenvalues of \mathbf{A} are natural modes of oscillation and squares of natural frequencies of oscillation, respectively.

The initial conditions $\dot{\mathbf{x}}(0) = \mathbf{x}_1$ or $\dot{\mathbf{x}}(0) = \mathbf{x}_2$ also excite the above two natural modes of oscillation. We note that for this particular example $\mathbf{B}\mathbf{u}(t)$ is of the form of \mathbf{x}_1 . The motion excited by the input vibration $\mathbf{u}(t)$ can only be proportional to \mathbf{x}_1 . Whether or not the motion is a sinusoidal oscillation is determined by the form of $\mathbf{u}(t)$.

4.7 Problems and Comments

- 4.1 Let $\mathscr{W}_1 = \text{span}\{(1, 0, 1)\}$ and $\mathscr{W}_2 = \text{span}\{(1, 0, 0), (0, 1, 0)\}$ in \mathscr{R}^3 .
- Show that an arbitrary vector \mathbf{x} in \mathscr{R}^3 can be decomposed into a unique pair of components \mathbf{x}_1 and \mathbf{x}_2 from \mathscr{W}_1 and \mathscr{W}_2 , respectively.
 - Let \mathbf{P}_1 be the projector onto \mathscr{W}_1 along \mathscr{W}_2 , and \mathbf{P}_2 the

projector onto \mathcal{W}_2 along \mathcal{W}_1 . Let \mathcal{E} be the standard basis for \mathcal{R}^3 . Find $[\mathbf{P}_1]_{\mathcal{X}\mathcal{X}}$ and $[\mathbf{P}_2]_{\mathcal{X}\mathcal{X}}$.

- 4.2 Let the linear operator \mathbf{T} defined by $\mathbf{T}\mathbf{x} \triangleq \mathbf{A}\mathbf{x}$ operate on the space $\mathcal{N}^{n \times 1}$. Let the subspaces \mathcal{W}_1 and \mathcal{W}_2 of $\mathcal{N}^{n \times 1}$ be composed of vectors of the form

$$\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_m \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \eta_{m+1} \\ \vdots \\ \eta_n \end{pmatrix}$$

respectively. Determine the form of \mathbf{A} if

- (a) \mathcal{W}_1 is invariant under \mathbf{T} .
- (b) \mathcal{W}_2 is invariant under \mathbf{T} .
- (c) Both \mathcal{W}_1 and \mathcal{W}_2 are invariant under \mathbf{T} .

Hint: investigate an example where $m = 1, n = 3$.

- 4.3 The *Cartesian product* is useful for *building up* complicated vector spaces from simple ones. The *direct sum*, on the other hand, is useful for *subdividing* complicated vector spaces into smaller subspaces.

- (a) Define $\mathbf{T}_a : \mathcal{R}^2 \rightarrow \mathcal{R}^2$ by $\mathbf{T}_a(\xi_1, \xi_2) \triangleq (\xi_1 - \xi_2, \xi_1)$.

Let $\mathcal{X}_a = \{(1,0), (0,1)\}$. Find $[\mathbf{T}_a]_{\mathcal{X}_a\mathcal{X}_a}$.

Define $\mathbf{T}_b : \mathcal{R}^1 \rightarrow \mathcal{R}^1$ by $\mathbf{T}_b(\xi_3) \triangleq (-\xi_3)$.

Let $\mathcal{X}_b = \{(1)\}$. Find $[\mathbf{T}_b]_{\mathcal{X}_b\mathcal{X}_b}$.

- (b) If we do not distinguish between $((\xi_1, \xi_2), (\xi_3))$ and (ξ_1, ξ_2, ξ_3) , then $\mathcal{R}^3 = \mathcal{R}^2 \times \mathcal{R}^1$. Define $\mathbf{T} : \mathcal{R}^3 \rightarrow \mathcal{R}^3$ by $\mathbf{T}((\xi_1, \xi_2),$

$(\xi_3)) \triangleq (\mathbf{T}_a(\xi_1, \xi_2), \mathbf{T}_b(\xi_3))$. Let $\mathcal{X} = \{((1, 0), (0)), ((0, 1), (0)), ((0, 0), (1))\}$. Find $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}$. What is the relationship between $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}, [\mathbf{T}_a]_{\mathcal{X}_a\mathcal{X}_a}, [\mathbf{T}_b]_{\mathcal{X}_b\mathcal{X}_b}$?

- (c) Let $\mathcal{W}_1 = \mathcal{R}^2 \times \{(0)\}$ and $\mathcal{W}_2 = \{(0,0)\} \times \mathcal{R}^1$. Then $\mathcal{R}^3 = \mathcal{W}_1 \oplus \mathcal{W}_2$. Appropriate bases for \mathcal{W}_1 and \mathcal{W}_2 are $\mathcal{X}_1 = \{((1,0), (0)), ((0,1), (0))\}$ and $\mathcal{X}_2 = \{((0,0), (1))\}$. Define $\mathbf{T}_1 : \mathcal{W}_1 \rightarrow \mathcal{W}_1$ by $\mathbf{T}_1(\xi_1, \xi_2, 0) \triangleq (\xi_1 - \xi_2, \xi_1, 0)$. Define $\mathbf{T}_2 : \mathcal{W}_2 \rightarrow \mathcal{W}_2$ by $\mathbf{T}_2(0, 0, \xi_3) \triangleq (0, 0, -\xi_3)$. Find $[\mathbf{T}_1]_{\mathcal{X}_1\mathcal{X}_1}$ and $[\mathbf{T}_2]_{\mathcal{X}_2\mathcal{X}_2}$. What is the relationship between $[\mathbf{T}]_{\mathcal{X}\mathcal{X}}, [\mathbf{T}_1]_{\mathcal{X}_1\mathcal{X}_1}$, and $[\mathbf{T}_2]_{\mathcal{X}_2\mathcal{X}_2}$?

- (d) In general, if $\mathcal{V} = \mathcal{W}_1 \oplus \dots \oplus \mathcal{W}_p$, with each subspace \mathcal{W}_i invariant under \mathbf{T} , then $\{\mathcal{W}_i\}$ decomposes \mathbf{T} into $\{\mathbf{T}_i: \mathcal{W}_i \rightarrow \mathcal{W}_i\}$. Let \mathcal{X}_i be a basis for \mathcal{W}_i . Then $\mathcal{X} = \{\mathcal{X}_1, \dots, \mathcal{X}_p\}$ is a basis for \mathcal{V} . If \mathcal{V} is finite-dimensional, then

$$[\mathbf{T}]_{\mathcal{X}\mathcal{X}} = \begin{pmatrix} [\mathbf{T}_1]_{\mathcal{X}_1\mathcal{X}_1} & & \\ & \ddots & \\ & & [\mathbf{T}_p]_{\mathcal{X}_p\mathcal{X}_p} \end{pmatrix}$$

with zeros everywhere except in the blocks on the diagonal. Show that the transformation $\mathbf{T}: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ defined by $\mathbf{T}(\xi_1, \xi_2, \xi_3) \triangleq (\xi_1 + \xi_2, 2\xi_1 + \xi_2 - \xi_3, \xi_1 + \xi_3)$ is decomposed by \mathcal{W}_1 and \mathcal{W}_2 , where \mathcal{W}_1 consists in vectors of the form $(\xi_1, \xi_2, \xi_1 + \xi_2)$ and \mathcal{W}_2 consists in vectors of the form (ξ_1, ξ_1, ξ_1) . Note that there is no Cartesian product which corresponds to this invariant direct-sum decomposition in the same manner as (b) corresponds to (c).

- 4.4 Find the eigenvalues and eigenvectors of the following matrices:

$$(a) \begin{pmatrix} -3 & 0 & 0 \\ -5 & 2 & 0 \\ -5 & 1 & 1 \end{pmatrix} \quad (b) \begin{pmatrix} -2 & 0 & 0 \\ -3 & 1 & 3 \\ 0 & 0 & -2 \end{pmatrix}$$

- 4.5 Let \mathbf{A} be an $n \times n$ matrix. Denote the characteristic polynomial for \mathbf{A} by $c(\lambda) = \lambda^n + b_1\lambda^{n-1} + \dots + b_n$. The trace of a matrix is defined as the sum of its diagonal elements, an easily computed quantity. An iterative method based on the trace function has been proposed for computing the coefficients $\{b_i\}$ in the characteristic polynomial [4.3, p. 296]. The iteration is:

$$b_1 = -\text{Trace}(\mathbf{A})$$

$$b_2 = -\frac{1}{2} [b_1 \text{Trace}(\mathbf{A}) + \text{Trace}(\mathbf{A}^2)]$$

$$b_3 = -\frac{1}{3} [b_2 \text{Trace}(\mathbf{A}) + b_1 \text{Trace}(\mathbf{A}^2) + \text{Trace}(\mathbf{A}^3)]$$

$$\vdots$$

$$b_n = -\frac{1}{n} [b_{n-1} \text{Trace}(\mathbf{A}) + \dots + b_1 \text{Trace}(\mathbf{A}^{n-1}) + \text{Trace}(\mathbf{A}^n)]$$

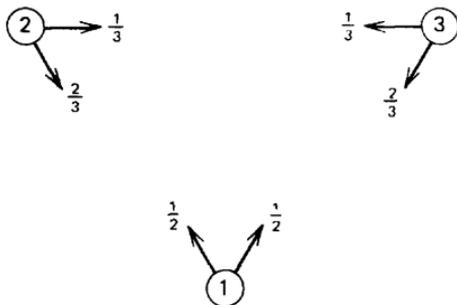
- (a) How many multiplications are required to compute the characteristic polynomial by means of this trace iteration? Compare the iteration with Krylov's method.

- (b) Compute the characteristic polynomial by Krylov's method and by the trace iteration for the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

- *4.6 Let \mathbf{A} be an $n \times n$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$. Then
- $\text{Det}(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$
 - $\text{Trace}(\mathbf{A}) \triangleq a_{11} + a_{22} + \cdots + a_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n$
 - If \mathbf{A} is triangular (i.e. if all elements to one side of the main diagonal are zero), then the diagonal elements of \mathbf{A} are $\mathbf{A}_{ii} = \lambda_i$.
- 4.7 Three men are playing ball. Every two seconds the one who has the ball tosses it to one of the others, with the probabilities shown in the diagram. Let $p_n(i)$ be the probability that the ball is held by the i th player (or is in the i th state) after the n th toss. Let p_{ij} be the probability with which player j throws the ball to player i . The theory of conditional probability requires that

$$p_n(i) = \sum_{j=1}^3 p_{ij} p_{n-1}(j) \text{ for } i=1,2,3$$



Let $\mathbf{x}_n \triangleq (p_n(1) \ p_n(2) \ p_n(3))^T$. We call \mathbf{x}_n a state probability vector. Let Ω denote the set of all possible 3×1 state probability vectors. The elements of each vector in Ω are non-negative and sum to one. Note that Ω is a *subset* of $\mathfrak{M}^{n \times 1}$, rather than a subspace. The game is an example of a Markov process. The future state probability vectors depend only on the present state, and not on the past history.

- (a) A matrix whose columns are members of Ω is called a *transition probability matrix*. Find the transition probability matrix \mathbf{A} such that $\mathbf{x}_n = \mathbf{A}\mathbf{x}_{n-1}$. Note that $\mathbf{x}_n = \mathbf{A}^n\mathbf{x}_0$; we refer to \mathbf{A}^n as the n -step transition probability matrix.
- (b) Determine the eigenvalues and eigenvectors of \mathbf{A} . What do they tell us about the game? (Hint: $\lambda = 1$ is an eigenvalue.)
- (c) Find the spectral matrix $\mathbf{\Lambda}$ and the modal matrix \mathbf{S} such that $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$. Show that every transition probability matrix has $\lambda = 1$ as an eigenvalue.
- (d) In the game described previously, the state probability vector \mathbf{x}_n becomes independent of the initial state as n becomes large. Find the form of the limiting state probability vector. (Hint: find $\lim_{n \rightarrow \infty} \mathbf{A}^n$ using the substitution $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$.) We note that the eigenvalues of every transition probability matrix satisfy $\lambda_i \leq 1$ [4.4, p. 429].
- (e) A transition probability matrix wherein the elements of each row also sum to one is called a *stochastic matrix*. What is the limiting state probability vector, $\lim_{n \rightarrow \infty} \mathbf{x}_n$, if the transition probabilities in the above game are modified to yield a stochastic matrix?

4.8 Let

$$\mathbf{A} = \begin{pmatrix} 3 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Find a matrix \mathbf{S} for which $\mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ is a diagonal matrix.

4.9 Find a nondiagonal matrix \mathbf{A} which has as its diagonal form the matrix

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

What are eigenvectors of \mathbf{A} ?

4.10 We wish to compute the eigendata of the matrix

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

Assume that numerical computations have produced the following

approximations to the eigenvalues: $\lambda_1 \approx 0.99$ and $\lambda_2 \approx -1.01$. Use the inverse iteration method to compute more accurate eigenvalues and corresponding eigenvectors. Start the iterations with the initial vector $\mathbf{z}_0 = (1 \ -1)^T$.

- 4.11 The **Jacobi method** for determining the eigenvalues and eigenvectors of a symmetric matrix \mathbf{A} consists in performing a sequence of similarity transformations which reduce the off-diagonal elements of \mathbf{A} to zero. In order to avoid a sequence of matrix inversions, we perform the similarity transformations with orthogonal matrices (matrices for which $\mathbf{S}^{-1} = \mathbf{S}^T$). Thus we let $\mathbf{A}_1 = \mathbf{S}_1^T \mathbf{A} \mathbf{S}_1$ and $\mathbf{A}_k = \mathbf{S}_k^T \mathbf{A}_{k-1} \mathbf{S}_k$ for $k=2, 3, \dots$. The eigenvalues of a matrix are not changed by similarity transformations. Consequently, the resulting diagonal matrix must be the spectral matrix (with the eigenvalues of \mathbf{A} on its diagonal); that is,

$$\lim_{k \rightarrow \infty} \mathbf{A}_k = \lim_{k \rightarrow \infty} (\mathbf{S}_1 \mathbf{S}_2 \cdots \mathbf{S}_k)^T \mathbf{A} (\mathbf{S}_1 \mathbf{S}_2 \cdots \mathbf{S}_k) = \mathbf{\Lambda}$$

Furthermore, the matrix $\mathbf{S} = \lim_{k \rightarrow \infty} (\mathbf{S}_1 \mathbf{S}_2 \cdots \mathbf{S}_k)$ must be a modal matrix for \mathbf{A} (with the eigenvectors of \mathbf{A} as its columns). Let $a_{ij} = (\mathbf{A}_{k-1})_{ij}$. It is shown in [4.13] that a_{ij} and a_{ji} can be driven to zero simultaneously by a similarity transformation which uses the orthogonal matrix \mathbf{S}_k which differs from the identity matrix only in the following elements:

$$(\mathbf{S}_k)_{ii} = (\mathbf{S}_k)_{jj} = \sqrt{(\gamma + |\beta|)/2\gamma} = \cos \phi$$

$$(\mathbf{S}_k)_{ij} = -(\mathbf{S}_k)_{ji} = \alpha \operatorname{sign}(\beta) / (2\gamma \cos \phi) = \sin \phi$$

where $\alpha = -a_{ij}$, $\beta = (a_{ii} - a_{jj})/2$, and $\gamma = (\alpha^2 + \beta^2)^{1/2}$. (Multiplication by the matrix \mathbf{S}_k can be interpreted as a rotation of the axes of the i and j coordinates through an angle ϕ .) In the Jacobi method we pick an \mathbf{S}_k of the above form which drives the largest pair of off-diagonal elements of \mathbf{A}_{k-1} to zero. Although later transformations will usually make these elements nonzero again, the sum of the squares of the off-diagonal elements is reduced at each iteration.

- (a) Use the Jacobi method to compute (to slide rule accuracy) the eigenvalues and eigenvectors of the matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

(b) Calculate the eigenvalues of \mathbf{A} by solving the characteristic polynomial. Determine the corresponding eigenvectors. Compare the results with (a).

4.12 Let \mathbf{L} be the differential operator defined by $\mathbf{L}\mathbf{f} \triangleq \mathbf{f}''$. Assume \mathbf{L} acts on the subspace of functions in $\mathcal{C}^2(0, \pi)$ which satisfy the boundary conditions $\beta_1(\mathbf{f}) = \beta_2(\mathbf{f}) = 0$. Find all the eigenvalues and corresponding eigenfunctions of \mathbf{L} for each of the following definitions of the boundary conditions:

(a) $\beta_1(\mathbf{f}) = \mathbf{f}(0), \quad \beta_2(\mathbf{f}) = \mathbf{f}(\pi)$

(b) $\beta_1(\mathbf{f}) = \mathbf{f}(0) + \mathbf{f}(\pi), \quad \beta_2(\mathbf{f}) = \mathbf{f}'(0) - \mathbf{f}'(\pi)$

(c) $\beta_1(\mathbf{f}) = \mathbf{f}(0) + 2\mathbf{f}(\pi), \quad \beta_2(\mathbf{f}) = \mathbf{f}'(0) - 2\mathbf{f}'(\pi)$

(d) $\beta_1(\mathbf{f}) = \mathbf{f}(0) - \mathbf{f}(\pi), \quad \beta_2(\mathbf{f}) = \mathbf{f}'(0) - \mathbf{f}'(\pi)$

4.13 Find the eigenvalues and eigenfunctions associated with the differential system $\mathbf{f}'' - c\mathbf{f} = \mathbf{u}, \mathbf{f}(0) = \mathbf{f}'(1) = 0$. Hint: $\ln(-1) = i(\pi + 2k\pi), k = 0, \pm 1, \pm 2, \dots$. For what values of the constant c is the system invertible?

4.14 Let \mathcal{V} be a space of functions \mathbf{f} whose values $\mathbf{f}(n)$ are defined only for integer values of n . Define the forward difference operator Δ on \mathcal{V} by

$$(\Delta\mathbf{f})(n) \triangleq \mathbf{f}(n+1) - \mathbf{f}(n)$$

(This operator can be used to approximate the differential operator \mathbf{D} .) Find the eigenvalues and eigenfunctions of Δ .

4.15 Define $\nabla^2\mathbf{f}(s, t) \triangleq (\partial^2\mathbf{f}/\partial s^2) + (\partial^2\mathbf{f}/\partial t^2)$ in the rectangular region $0 \leq s \leq a$ and $0 \leq t \leq b$. Let \mathbf{f} satisfy the boundary conditions

$$\frac{\partial\mathbf{f}}{\partial s}(0, t) = \frac{\partial\mathbf{f}}{\partial s}(a, t) = \frac{\partial\mathbf{f}}{\partial t}(s, 0) = \frac{\partial\mathbf{f}}{\partial t}(s, b) = 0$$

Show that the partial differential operator ∇^2 and the given boundary conditions have the eigendata

$$\lambda_{km} = -\left(\frac{m\pi}{a}\right)^2 - \left(\frac{k\pi}{b}\right)^2$$

$$\mathbf{f}_{km}(s, t) = \cos\left(\frac{m\pi s}{a}\right) \cos\left(\frac{k\pi t}{b}\right)$$

for $k, m = 0, 1, 2, \dots$

*4.16 Let \mathbf{A} be the companion matrix for an n th order constant-coefficient differential operator. Denote the eigenvalues of \mathbf{A} by $\lambda_1, \dots, \lambda_n$.

- (a) Show that the vector $\mathbf{z}_i = (1 \ \lambda_i \ \lambda_i^2 \ \cdots \ \lambda_i^{n-1})^T$ is an eigenvector of \mathbf{A} for the eigenvalue λ_i . Show further that there is only one independent eigenvector for each distinct eigenvalue.
- (b) Show that the Vandermond matrix

$$\begin{pmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & \vdots & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix}$$

is a modal matrix for \mathbf{A} if and only if the eigenvalues of \mathbf{A} are all distinct.

- *4.17 *The power method:* the inverse of the differential operator $\mathbf{L} = \mathbf{D}^2$ with the boundary conditions $\mathbf{f}(0) = \mathbf{f}(1) = \mathbf{0}$ is the integral operator \mathbf{T} defined by

$$(\mathbf{T}\mathbf{u})(t) = \int_0^t (t-1)s\mathbf{u}(s) ds + \int_t^1 t(s-1)\mathbf{u}(s) ds$$

The functions $\mathbf{f}_n(t) = \sin n\pi t$, $n = 1, 2, \dots$, are eigenfunctions for both the differential and integral operators. We can find the dominant eigenvalue and the corresponding eigenfunction of \mathbf{T} by the power method. We just compute the sequence of functions $\mathbf{u}_k = \mathbf{T}^k \mathbf{u}_0$, for some initial function \mathbf{u}_0 , until \mathbf{u}_k is a sufficiently good approximation to the dominant eigenfunction.

- (a) Let $\mathbf{u}_0(t) = 1$, and compute \mathbf{u}_1 and \mathbf{u}_2 .
- (b) Compare \mathbf{u}_1 and \mathbf{u}_2 with the true dominant eigenfunction. Use the iterates $\{\mathbf{u}_k\}$ to determine an approximation to the dominant eigenvalue.
- 4.18 (a) Determine an ordered basis of generalized eigenvectors for the matrix

$$\mathbf{A} = \begin{pmatrix} 5 & -1 & 1 & 1 & 0 & 0 \\ 1 & 3 & -1 & -1 & 0 & 0 \\ 0 & 0 & 4 & 0 & 1 & 1 \\ 0 & 0 & 0 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 1 & 3 \end{pmatrix}$$

Hint: $\det(\mathbf{A} - \lambda\mathbf{I}) = (4 - \lambda)^5 (2 - \lambda)$.

- (b) Determine the Jordan canonical form of \mathbf{A} [relative to the basis found in (a)].

(c) Determine the “change of coordinates” matrix \mathbf{S} which would be used in a similarity transformation on \mathbf{A} in order to obtain the Jordan form found in (b). (Obtain only the obvious matrix, not its inverse.)

4.19 Find a matrix \mathbf{S} such that $\mathbf{S}^{-1}\mathbf{B}\mathbf{S}$ is in Jordan form, for

$$\mathbf{B} = \begin{pmatrix} 3 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 1 & 1 & 3 & 1 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

Hint: $c(\lambda) = (2 - \lambda)^3(3 - \lambda)$.

4.20 The minimal polynomial $m(\lambda)$ and the characteristic polynomial $c(\lambda)$ are useful for reducing effort in matrix computations. Assume $f(\mathbf{A})$ is a polynomial in the $n \times n$ matrix \mathbf{A} , and $f(\mathbf{A})$ includes powers of \mathbf{A} higher than n . We divide $f(\lambda)$ by $m(\lambda)$ to determine a quotient $g(\lambda)$ and a remainder $r(\lambda)$; that is, $f(\lambda) = g(\lambda)m(\lambda) + r(\lambda)$. If we replace λ by \mathbf{A} , and use the fact that $m(\mathbf{A}) = \mathbf{0}$, we observe that $f(\mathbf{A}) = r(\mathbf{A})$. The remainder $r(\mathbf{A})$ is of lower degree (in \mathbf{A}) than $m(\mathbf{A})$, regardless of the degree of $f(\mathbf{A})$. Consequently, $r(\mathbf{A})$ is easier to compute than is $f(\mathbf{A})$. The same procedure can be carried out using the more easily determined characteristic polynomial rather than the minimal polynomial. Use this “remainder” method to compute the matrix \mathbf{A}^5 for

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

4.21 Assume f is analytic at the eigenvalues of the matrix \mathbf{A} . Find the component matrices of \mathbf{A} and express $f(\mathbf{A})$ as a linear combination of these components for:

$$(a) \quad \mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \quad (b) \quad \mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

4.22 The gamma function $\Gamma(p)$ is defined for all positive values of the scalar p . If p is a positive integer, $\Gamma(p) = (p - 1)!$ Find $\Gamma(\mathbf{A})$, where

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & -2 \\ 0 & 3 & -1 \\ 0 & 0 & 2 \end{pmatrix}$$

4.23 Let

$$f(\lambda) \triangleq 0, \quad \lambda \leq c$$

$$\triangleq (\lambda - c)^2, \quad \lambda \geq c$$

and

$$\mathbf{\Lambda} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

(a) Find $f(\mathbf{\Lambda})$.

(b) Consider various values of c . Is the resulting matrix what you would expect?

4.24 If \mathbf{A} is invertible, the inverse can be computed by evaluating $f(\mathbf{A})$ for $f(\lambda) \triangleq 1/\lambda$. By modifying f , we can compute a “pseudoinverse” for a matrix which has zero eigenvalues. We merely change the definition of f to

$$\hat{f}(\lambda) \triangleq \frac{1}{\lambda}, \quad \lambda \neq 0$$

$$\triangleq 0, \quad \lambda = 0$$

(See P&C 6.22 for an interpretation of this “pseudoinverse.”)

(a) Find the inverse of the matrix \mathbf{A} of P&C 4.21 *a* by evaluating $f(\mathbf{A})$.

(b) Find the “pseudoinverse” of the following matrix by evaluating $\hat{f}(\mathbf{B})$:

$$\mathbf{B} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 2 \end{pmatrix}$$

4.25 The constituent matrices of a square matrix \mathbf{A} can be determined by partial fraction expansion of the *resolvent matrix*, $(s\mathbf{I} - \mathbf{A})^{-1}$ (the resolvent matrix is the Laplace transform of $e^{\mathbf{A}t}$). Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ -1 & 1 & 2 \end{pmatrix}$$

- (a) Determine the resolvent matrix $(s\mathbf{I} - \mathbf{A})^{-1}$ by inverting $(s\mathbf{I} - \mathbf{A})$.
- (b) Perform a partial fraction expansion of $(s\mathbf{I} - \mathbf{A})^{-1}$; that is, perform a partial fraction expansion of each term of $(s\mathbf{I} - \mathbf{A})^{-1}$, and arrange the expansion into a sum of terms with multipliers which are constant 3×3 matrices.
- (c) Let $f(\lambda) \triangleq 1/(s - \lambda)$; then $f(\mathbf{A}) = (s\mathbf{I} - \mathbf{A})^{-1}$. Express the fundamental formula for $f(\mathbf{A})$ in terms of $\{\mathbf{E}_{ij}^{\lambda}\}$, the constituent matrices for \mathbf{A} . (The form of the fundamental formula is determined by the minimal polynomial for \mathbf{A} .) Determine the constituent matrices by comparing the fundamental formula for $f(\mathbf{A})$ with the partial fraction expansion obtained in (b).
- (d) Use the fundamental formula and the constituent matrices to evaluate \mathbf{A}^5 .

4.26 Let f be a scalar-valued function of a scalar variable. Assume f is defined on the spectrum of the $n \times n$ matrix \mathbf{A} . Then $f(\mathbf{A})$ can be expressed as a polynomial in \mathbf{A} of lower degree than the minimal polynomial for \mathbf{A} . That is, if r is the degree of the minimal polynomial, then $f(\mathbf{A}) = a_0\mathbf{I} + a_1\mathbf{A} + \cdots + a_{r-1}\mathbf{A}^{r-1}$. The coefficients $\{a_i\}$ can be determined by evaluating the corresponding scalar equation, $f(\lambda) = a_0 + a_1\lambda + \cdots + a_{r-1}\lambda^{r-1}$, on the spectrum of \mathbf{A} ; the resulting equations are always solvable.

- (a) Find the minimal polynomial for the matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ -1 & 1 & 2 \end{pmatrix}$$

- (b) For the matrix \mathbf{A} introduced in (a), evaluate the matrix function $f(\mathbf{A}) \triangleq \mathbf{A}^5$ by the technique described above.

4.27 Let the $n \times n$ matrix \mathbf{A} be diagonalizable. Then, the fundamental formula is $f(\mathbf{A}) = \sum_{i=1}^p f(\lambda_i)\mathbf{E}_{i0}^{\lambda}$, where p is the number of distinct eigenvalues. The constituent matrix $\mathbf{E}_{i0}^{\lambda}$ is the projector on the eigenspace for λ_i along the sum of the other eigenspaces. It can be expressed as

$$\mathbf{E}_{i0}^{\lambda} = \prod_{j \neq i} \left(\frac{\mathbf{A} - \lambda_j \mathbf{I}}{\lambda_i - \lambda_j} \right)$$

($\mathbf{E}_{i0}^{\lambda}$ acts like \mathbf{I} on the eigenspace for λ_i and like $\mathbf{0}$ on the eigenspace

for λ_j .) The scalar equivalent of the fundamental formula,

$$f(\lambda) = \sum_{i=1}^p f(\lambda_i) \prod_{j \neq i} \left(\frac{\lambda - \lambda_j}{\lambda_i - \lambda_j} \right)$$

is known as the *Lagrange interpolation formula* for the data points $\lambda_1, \dots, \lambda_p$.

(a) Let

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{pmatrix}$$

Find the constituent matrices $\mathbf{E}_{i0}^{\mathbf{A}}$ by evaluating the polynomial expressions given above.

(b) Use the fundamental formula to evaluate the matrix exponential, $e^{\mathbf{A}t}$, for the matrix \mathbf{A} given in (a).

4.28 Use the fundamental formula to find four square roots of the matrix

$$\mathbf{A} = \begin{pmatrix} 20 & -8 \\ 48 & -20 \end{pmatrix}$$

*4.29 (a) *Commuting matrices:* if \mathbf{A} and \mathbf{B} commute (i.e., $\mathbf{AB} = \mathbf{BA}$), then

$$(\mathbf{A} + \mathbf{B})^n = \sum_{k=0}^n \binom{n}{k} \mathbf{A}^{n-k} \mathbf{B}^k, \quad n=0, 1, 2, \dots$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

That is, the binomial theorem is satisfied.

(b) The algebra of matrices is essentially the same as the algebra of scalars if the matrices commute with each other. Therefore, a functional relation which holds for scalars also holds for commuting matrices if the required matrix functions are defined. For example, $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$, $\cos(\mathbf{A} + \mathbf{B}) = \cos \mathbf{A} \cos \mathbf{B} - \sin \mathbf{A} \sin \mathbf{B}$; the binomial theorem is satisfied; etc.

(c) If \mathbf{A} and \mathbf{B} are diagonalizable, then they are commutable if and only if they are diagonalizable by the same similarity

transformation (i.e., if and only if they have the same eigenvectors).

- 4.30 Use the fundamental formula to show that $(d/dt)f(\mathbf{A}t) = \mathbf{A}\dot{f}(\mathbf{A}t)$ for any square matrix \mathbf{A} and any function f which is analytic on the spectrum of \mathbf{A} .
- 4.31 Let $\mathbf{f}'' + 6\mathbf{f}' + 5\mathbf{f} = \mathbf{u}$, $\mathbf{f}(0) = \mathbf{f}'(0) = \mathbf{0}$.
- Express the differential system in state-space form.
 - Diagonalize the state equation found in (a).
 - Draw a signal flow diagram which relates the original state variables, the canonical state variables, and the input.
 - Find the state transition matrix and invert the state equation.
- 4.32 Let $\ddot{\mathbf{x}} + \mathbf{A}\dot{\mathbf{x}} = \mathbf{B}\mathbf{u}$, where $\mathbf{x}(t)$ is $n \times 1$, $\mathbf{u}(t)$ is $m \times 1$, \mathbf{B} is $n \times m$, and \mathbf{A} is $n \times n$ with positive eigenvalues. Assume $\mathbf{x}(0)$ and $\dot{\mathbf{x}}(0)$ are known.
- Use the power series method of Frobenius to show that the complementary function for this vector differential equation is

$$\mathbf{F}_c(t) = \cos(\sqrt{\mathbf{A}} t)\mathbf{C}_0 + (\sqrt{\mathbf{A}})^{-1} \sin(\sqrt{\mathbf{A}} t)\mathbf{C}_1$$

where \mathbf{C}_0 and \mathbf{C}_1 are arbitrary $n \times n$ matrices.

- The inverse of the differential equation is of the form

$$\mathbf{x}(t) = \int_0^\infty \mathbf{K}(t,s)\mathbf{B}\mathbf{u}(s) ds + \mathbf{R}_1(t)\mathbf{x}(0) + \mathbf{R}_2(t)\dot{\mathbf{x}}(0)$$

Show that the Green's function $\mathbf{K}(t,s)$ and boundary kernel $\mathbf{R}_j(t)$ satisfy:

$$\frac{d^2}{dt^2} \mathbf{K}(t,s) + \mathbf{A}\mathbf{K}(t,s) = \delta(t-s)\mathbf{I}$$

$$\mathbf{K}(0,s) = \frac{d}{dt} \mathbf{K}(0,s) = \mathbf{\Theta}$$

$$\frac{d^2}{dt^2} \mathbf{R}_j(t) + \mathbf{A}\mathbf{R}_j(t) = \mathbf{\Theta}, \quad j=1,2$$

$$\mathbf{R}_1(0) = \mathbf{I}, \quad \dot{\mathbf{R}}_1(0) = \mathbf{\Theta}$$

$$\mathbf{R}_2(0) = \mathbf{\Theta}, \quad \dot{\mathbf{R}}_2(0) = \mathbf{I}$$

(c) Show that

$$\mathbf{K}(t,s) = \mathbf{\Theta}, \quad t \leq s$$

$$= (\sqrt{\mathbf{A}})^{-1} \sin(\sqrt{\mathbf{A}}(t-s)), \quad t \geq s$$

$$\mathbf{R}_1(t) = \cos \sqrt{\mathbf{A}} t$$

$$\mathbf{R}_2(t) = (\sqrt{\mathbf{A}})^{-1} \sin(\sqrt{\mathbf{A}} t)$$

4.33 In optimal control problems we often need to solve a pair of simultaneous state equations. Suppose the equations are $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{B}^T\boldsymbol{\lambda}$ and $\dot{\boldsymbol{\lambda}} = -\mathbf{A}^T\boldsymbol{\lambda}$, where

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- (a) Write the pair of equations as a single state equation $\dot{\mathbf{y}} = \mathbf{Q}\mathbf{y}$, where $\mathbf{y} \triangleq \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix}$.
- (b) Find the eigenvalues and constituent matrices of \mathbf{Q} .
- (c) Find the solution \mathbf{y} to the state equation as a function of $\mathbf{y}(0)$.

4.8 References

- [4.1] Bother, M., *Introduction to Higher Algebra*, Macmillan, New York, 1921.
- [4.2] Brown, R. G., "Not Just Observable, But How Observable," *Proc. Natl. Elec. Conf.*, 22 (1966), 709-714.
- [4.3] DeRusso, Paul M., Rob J. Roy, and Charles M. Close, *State Variables for Engineers*, Wiley, New York, 1966.
- [4.4] Feller, W., *An Introduction to Probability Theory and its Applications*, 3rd Ed., Vol. I, Wiley, New York, 1968.
- [4.5] Forsythe, George E., "Singularity and Near Singularity in Numerical Analysis," *Am. Math. Mon.*, 65 (1958), 229-240.
- [4.6] Forsythe, George E., "Today's Computational Methods of Linear Algebra," *SIAM Rev.* 9 (July 1967), 489-515.
- [4.7] Friedman, Bernard, *Principles and Techniques of Applied Mathematics*, Wiley, New York, 1966.
- [4.8] Halmos, P. R., *Finite-Dimensional Vector Spaces*, Van Nostrand, Englewood Cliffs, N. J., 1958.
- [4.9] Hoffman, Kenneth and Roy Kunze, *Linear Algebra*, Prentice-Hall, Englewood Cliffs, N. J., 1961.