Non-Parametric Least Squares

Based on Chapter 13 of *High Dimensional Statistics* by Martin Wainwright

Behrad Moniri

Department of Electrical Engineering
Sharif University of Technology

Statistics Reading Group
Tehran Institute for Advanced Studies (TeIAS)

First Session:

1. Introduction
2. Examples
3. Critical Radius: Bounding the Prediction Error
4. How to Compute the Critical Radius?

Second Session:

5. Review and Examples
6. Oracle Inequalities
7. Examples
8. Regularized Estimators
9. Kernel Ridge Regression

Part I

First Session

Section 1

Introduction

Non-Parametric Least Squares

- **MMSE Estimation**:

$$\bar{\mathcal{L}}_f = \mathbb{E}_{X,Y}[(Y - f(X))^2] \implies f^\star(x) = \mathbb{E}[Y|X = x]$$

- In practice we are given a collection of samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, which can be used to compute an empirical analog of the MSE:

$$\widehat{\mathcal{L}}_f = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2$$

- **Non-Parametric Least Squares**: Minimizing this criterion over some suitably controlled function class.

Different Measures of Quality

- **Excess Risk**:

$$||f - f^\star||_{L^2(\mathbb{P})} = \mathbb{E}_{X\sim\mathbb{P}}\left[\left(f(X) - f^\star(X)\right)^2\right]$$

where $\mathbb{P}$ denotes the distribution over covariates.

- **This Chapter**: Let $\{\mathbf{x}_i\}_{i=1}^n$ be the set of fixed covariates and $\mathbb{P}_n$ be their Empirical measure. Define:

$$||f - f^\star||_{L^2(\mathbb{P}_n)} = \left[\frac{1}{n}\sum_{i=1}^n \left(f(\mathbf{x}_i) - f^\star(\mathbf{x}_i)\right)^2\right]^{1/2}.$$

In this talks, we will use this measure and denote it as $||f - f^\star||_n$.

Fixed or Stochastic Covariates?

Note

- We will view the samples $\{\mathbf{x}_i\}_{i=1}^n$ as being fixed, a set-up known as *regression with a fixed design*.
- Results from Chapter 14 to follow can be used to translate these bounds into equivalent results in the population $L_2(\mathbb{P})$-norm.

Estimation via Least Squares

Given a fixed collection $\{\mathbf{x}_i\}_{i=1}^n$, model the responses as

$$y_i = f^\star(\mathbf{x}_i) + \nu_i, \quad \text{for } i = 1, 2, \ldots, n.$$

where $\nu_i = \sigma w_i$ in which $w_i \sim \mathcal{N}(0, 1)$. The least squares estimate is given by the function

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 \right\}.$$

- When $\nu_i \sim \mathcal{N}(0, \sigma^2)$, the LS estimate is equivalent to the constrained maximum likelihood.
- When $\mathcal{F}$ is an RKHS, it can also be convenient to use regularized estimators of the form:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 + \lambda_n ||f||_{\mathcal{F}}^2 \right\}.$$

Section 2

Examples

Parametric: Linear Regression

- For a given vector $\boldsymbol{\theta} \in \mathbb{R}^d$, define $f_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$ and consider the function class $\mathcal{F}_C = \{f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R} | \boldsymbol{\theta} \in C\}$ for a compact C.
- The least squares estimate:

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in C} \left\{ \frac{1}{n} ||y - \mathbf{X}\boldsymbol{\theta}||_2^2 \right\}$$

  where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix.

- The constrained form of Ridge Regression:

$$C = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; ||\boldsymbol{\theta}||_2^2 \leq R \right\}.$$

- The constrained form of LASSO:

$$C = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \;\middle|\; ||\boldsymbol{\theta}||_1 \leq R \right\}.$$

Non-Parametric: Cubic Smoothing Spline

Consider the class of twice cont. differentiable functions $f : [0, 1] \to \mathbb{R}$ and for a given radius R, define:

$$\mathcal{F}(R) = \left\{ f : [0,1] \to \mathbb{R} \;\middle|\; \int_0^1 \left(f''(x) \right)^2 dx \leq R \right\}.$$

The integral can be understood as a Hilbert norm bound in the second-order Sobolev space.

Section 3

Critical Radius: Bounding the Prediction Error

Critical Radius: Bounding the Prediction Error

- Intuitively, the difficulty of estimating the function $f^\star$ should depend on the complexity of the function class $\mathcal{F}$ in which it lies.
- **a localized form of Gaussian complexity:** it measures the complexity of the function class $\mathcal{F}$, locally in a neighborhood around the true regression function $f^\star$.

Local Gaussian Complexity

Define $f^\star = \mathcal{F} - \{f^\star\} = \{f - f^\star | f \in \mathcal{F}\}$. For a given radius $\delta > 0$, *the local Gaussian complexity* around $f^\star$ at scale δ is given by:

$$\mathcal{G}_n(\delta; f^\star) = \mathbb{E}_w \left[\sup_{g \in \mathcal{F}^\star, ||g||_n \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} w_i \, g(\mathbf{x}_i) \right| \right]$$

Critical Inequality

Critical Inequality

A central object in our analysis is the set of positive scalars δ that satisfy the *critical inequality*:

$$\frac{\delta}{2\sigma} \geq \frac{\mathcal{G}_n(\delta; \mathcal{F}^\star)}{\delta}$$

- We refer to any $\delta > 0$ satisfying the critical inequality as being valid.
- We use $\delta_n^\star$ to denote the smallest positive radius for which the critical inequality holds.

Some Intuition

- Note that $\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\widehat{f}(\mathbf{x}_i)\right)^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-f^\star(\mathbf{x}_i)\right)^2$.
- After some computations: $\frac{1}{2}||\widehat{f}-f^\star||_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^{n} w_i\left(\widehat{f}(\mathbf{x}_i)-f^\star(\mathbf{x}_i)\right)$.
- Note that $\widehat{f}-f^\star \in f^\star$. Reasoning hueristically, this observation suggests that

$$\frac{\delta^2}{2} \leq \sigma\mathcal{G}_n(\delta;\mathcal{F}^\star) \text{ or equivalently } \frac{\delta}{2\sigma} \leq \frac{\mathcal{G}_n(\delta;\mathcal{F}^\star)}{\delta}.$$

- We will present a rigorous result later in this talk.

The Main Result

Star-Shaped Classes

a function class $\mathcal{F}$ is star-shaped if for any $\alpha \in [0, 1]$ we have

$$f \in \mathcal{F} \implies \alpha f \in \mathcal{F}.$$

Theorem

Suppose that the shifted function class $\mathcal{F}^\star$ is star-shaped, and let δ be any solution to the critical inequality. Then for any $t \geq \delta$, the nonparametric least-squares estimate $\widehat{f}_n$ satisfies the bound

$$\mathbb{P}\left[||\widehat{f}_n - f^\star||_n^2 \geq 16t\delta\right] \leq \exp\left(\frac{-nt\delta}{2\sigma^2}\right) \tag{1}$$

Properties of Star-Shaped Classes

- The star shaped condition on $\mathcal{F}^\star$ is needed in various parts of the proof, including ensuring a valid radii δ exists.
- $\mathcal{F}^\star$ Star shaped $= \mathcal{F}$ is star shaped around $f^\star$.

Property

If $\mathcal{F}$ is convex, it is star-shaped around any $f^\star$.

Proof

Let $\widetilde{f} = f - f^\star \in \mathcal{F}^\star$ be an arbitrary point in $\mathcal{F}^\star$ with $f, f^\star \in \mathcal{F}$.
For any $\alpha \in [0,1]$, by convexity we have $g = \alpha f + (1-\alpha) f^\star \in \mathcal{F}$.
Hence $\alpha\widetilde{f} = \alpha(f - f^\star) = g - f^\star \in \mathcal{F}^\star$, proving $\mathcal{F}^\star$ is star-shaped.

Conversely

Similarly if $\mathcal{F}$ is not convex, there must exists a $f^\star$ such that $\mathcal{F}$ is not star shaped around $f^\star$.

Properties of Star-Shaped Classes

- If the star-shaped condition fails to hold, then the main Theorem can instead by applied with δ defined in terms of the star hull (we will see next session.)

$$\text{star}(\mathcal{F}^\star) = \left\{\alpha(f - f^\star) | f \in \mathcal{F}, \alpha \in [0,1]\right\}.$$

Existence of Critical Radius

For any star shaped class $\mathcal{F}^\star$, the function $\delta \to \frac{\mathcal{G}_n(\delta, \mathcal{F}^\star)}{\delta}$ is non-increasing on $(0, \infty)$. Consequently, for any $c > 0$, the inequality

$$\frac{\mathcal{G}_n(\delta, \mathcal{F}^\star)}{\delta} \leq c\delta$$

has a smallest positive solution.

Proof of the Existence of Critical Radius

Given $0 < \delta \leq t$, we should show that $\frac{\delta}{t}\mathcal{G}_n(t) \leq \mathcal{G}_n(\delta)$.
Given any $h \in \mathcal{F}^\star$ with $||h||_n \leq t$, define $\widetilde{h} = \frac{\delta}{t}h \in \mathcal{F}^\star$ (by star-shaped assumption) and write

$$\frac{1}{n}\Big\{\frac{\delta}{t}\sum_{i=1}^{n} w_i h(\mathbf{x}_i)\Big\} = \frac{1}{n}\Big\{\sum_{i=1}^{n} w_i \widetilde{h}(\mathbf{x}_i)\Big\}.$$

By construction, $||\widetilde{h}||_n \leq \delta$. Consequently:

- The RHS is at most $\mathcal{G}_n(\delta)$ in expectation.
- Taking supremum over the set $\mathcal{F} \cap \{||h||_n \leq t\}$ following by expectation, yields $\frac{\delta}{t}\mathcal{G}_n(t)$ on the left hand side.

This concludes the proof. □

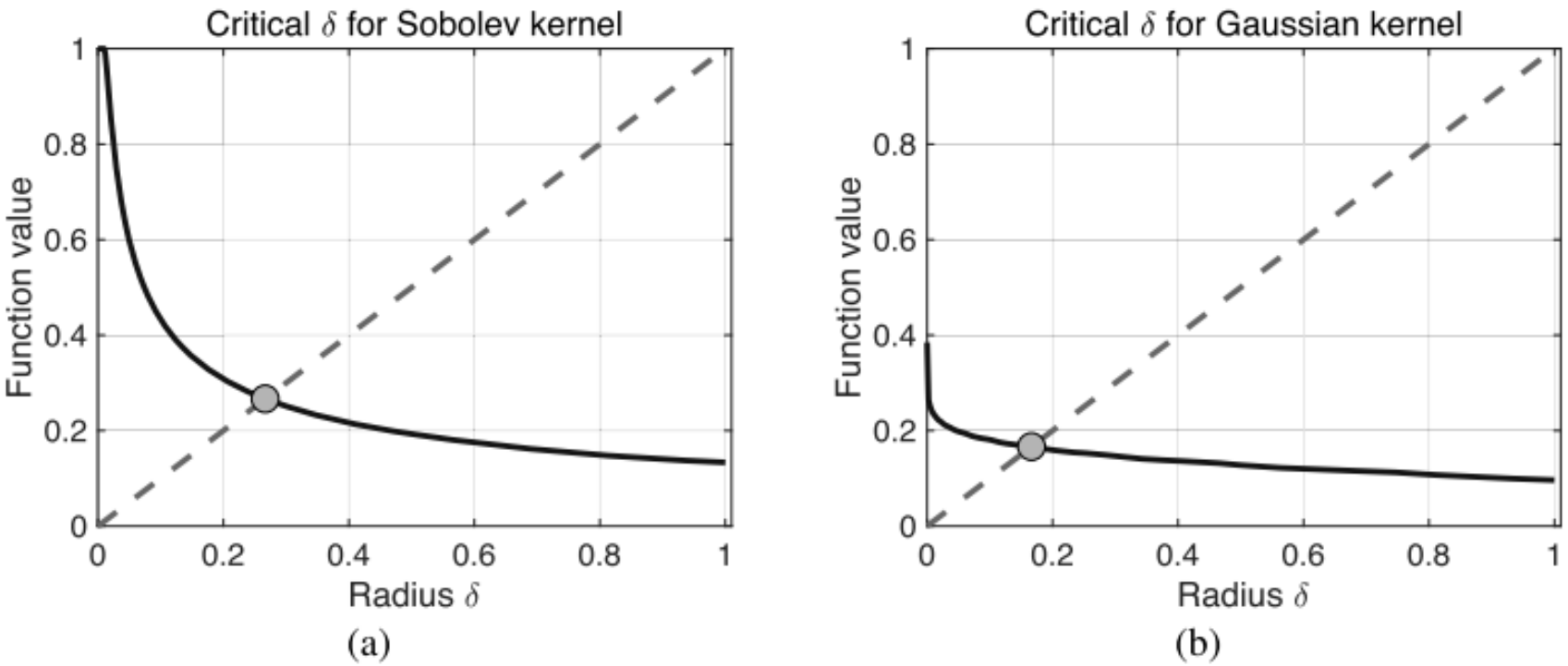

Figure 13.2 Illustration of the critical radius for sample size $n = 100$ and two different function classes. (a) A first-order Sobolev space. (b) A Gaussian kernel class. In both cases, the function $\delta \mapsto \frac{\mathcal{G}_n(\delta; \mathcal{F})}{\delta}$, plotted as a solid line, is non-increasing, as guaranteed by Lemma 13.6. The critical radius δ_n^*, marked by a gray dot, is determined by finding its intersection with the line of slope $1/(2\sigma)$ with $\sigma = 1$, plotted as the dashed line. The set of all valid δ_n consists of the interval $[\delta_n^*, \infty)$.

Proof of the main theorem

Theorem

Suppose that the shifted function class $\mathcal{F}^\star$ is star-shaped, and let δ be any solution to the critical inequality. Then for any $t \geq \delta$, the nonparametric least-squares estimate $\widehat{f}_n$ satisfies the bound

$$\mathbb{P}\left[||\widehat{f}_n - f^\star||_n^2 \geq 16t\delta\right] \leq \exp\left(\frac{-nt\delta}{2\sigma^2}\right)$$

- Denote $\widehat{\Delta} = \widehat{f} - f^\star \in \mathcal{F}^\star$, we have $\frac{1}{2}||\widehat{\Delta}||_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i \widehat{\Delta}(\mathbf{x}_i)$. We need to control the stochastic component of the right-hand side.

Auxiliary Lemma

Let $\mathcal{F}^\star$ be a star-shaped class and δ satisfy the critical inequality. Define the bad event $A(u)$ as

$$A(u) = \left\{ \exists g \in \mathcal{F}^\star, ||g||_n \geq u \ : \ \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(\mathbf{x}_i) \right| \geq 2u||g||_n \right\}.$$

For a given $u \geq \delta$, we have $\mathbb{P}[A(u)] \leq \exp\left(\frac{-nu^2}{2\sigma^2}\right)$.

Now consider two cases:

- **Case 1:** $||\widehat{\Delta}||_n < \sqrt{t\delta}$ which implies $||\widehat{\Delta}||_n^2 \leq t\delta$ trivially.
- **Case 2:** $||\widehat{\Delta}||_n \geq \sqrt{t\delta}$. We condition on $A^c(\sqrt{t\delta})$. Set $u = \sqrt{t\,\delta}$, so that we have $\mathbb{P}[A^c(\sqrt{t\,\delta})] \geq 1 - \exp(\frac{-nt\delta}{2\sigma^2})$. Hence:

$$||\widehat{\Delta}||_n^2 \leq 2\left| \frac{\sigma}{n} \sum_{i=1}^n w_i \widehat{\Delta}(\mathbf{x}_i) \right| \leq 4||\widehat{\Delta}||_n \sqrt{t\delta} \implies ||\Delta||_n^2 \leq 16t\delta. \quad \square$$

Proof of the Auxiliary Lemma

Recall $A(u) = \left\{ \exists g \in \mathcal{F}^\star, ||g||_n \geq u \ : \ \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(\mathbf{x}_i) \right| \geq 2u||g||_n \right\}$.

Step 1: reduce to controlling a sup over a subset with $||\widetilde{g}||_n \leq u$.

- Suppose that $A(u)$ is true, hence there exists $g \in \mathcal{F}^\star$ with $||g||_n \geq u$ such that

$$\left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(\mathbf{x}_i) \right| \geq 2||g||_n u.$$

- Define $\widetilde{g} = \frac{u}{||g||_n} g \in \mathcal{F}^\star$, we observe that $||\widetilde{g}||_n = u$. We have

$$\left| \frac{\sigma}{n} \sum_{i=1}^n w_i \widetilde{g}(\mathbf{x}_i) \right| = \frac{u}{||g||_n} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(\mathbf{x}_i) \right| \geq 2u^2.$$

- We thus conclude that:

$$\mathbb{P}[A(u)] \leq \mathbb{P}[Z_n(u) \geq 2u^2], \ Z_n(u) := \sup_{\widetilde{g} \in \mathcal{F}^\star, \ ||\widetilde{g}||_n \leq u} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i \widetilde{g}(\mathbf{x}_i) \right|.$$

Proof of the Auxiliary Lemma

$\mathbb{P}[A(u)] \leq \mathbb{P}[Z_n(u) \geq 2u^2]$, $Z_n(u) := \sup_{\widetilde{g} \in \mathcal{F}^\star, \ ||\widetilde{g}||_n \leq u} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i \widetilde{g}(\mathbf{x}_i) \right|$.

Step 2: Concentration of suprema.

- Recall that $w_i \sim \mathcal{N}(0,1)$, hence $\frac{\sigma}{n} \sum_{i=1}^n w_i \widetilde{g}(\mathbf{x}_i) \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} ||\widetilde{g}||_n^2\right)$.
- $Z_n(u) = h(w_1, \ldots, w_n) := \sup_{\widetilde{g} \in \mathcal{F}^\star, \ ||\widetilde{g}||_n \leq u} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i \widetilde{g}(\mathbf{x}_i) \right|$.
- The Lipschitz constant of h is at most $\frac{\sigma u}{\sqrt{n}}$.
- We know that for $Z_i \sim \mathcal{N}(0, \sigma^2)$ and h a L-Lipschitz function:

$$\mathbb{P}[h(Z) - \mathbb{E}h(Z) \geq t] \leq e^{-\frac{s^2}{2L^2\sigma^2}}.$$

- Let $s = u^2$, we obtain $\mathbb{P}(Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2) \leq \exp\left(\frac{-nu^2}{2\sigma^2}\right)$.

Proof of the Auxiliary Lemma

Step 3: Bounding the expectation.

- Note that $\mathbb{E}[Z_n(u)] = \sigma\, \mathcal{G}_n(u)$.
- We know that $\nu \to \frac{\mathcal{G}_n(\nu)}{\nu}$ is non-increasing. We have also assumed that $u \geq \delta$.

$$\sigma \frac{\mathcal{G}_n(u)}{u} \leq \sigma \frac{\mathcal{G}_n(\delta)}{\delta} \leq \frac{\delta}{2} \leq \delta \implies \mathbb{E}[Z_n(u)] \leq u\delta \leq u^2.$$

Comibining our results, we obtain

$$\mathbb{P}[A(u)] \leq \mathbb{P}[Z_n(u) \geq 2u^2] \leq \mathbb{P}(Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2) \leq \exp\left(\frac{-nu^2}{2\sigma^2}\right),$$

Which concludes the proof. □

Section 4

How to Compute the Critical Radius?

Bounds via Metric Entropy

For any star-shaped function class $\mathcal{F}^\star$, define:

- $B(\delta\,;\mathcal{F}^\star) = \{h \in \mathcal{F}^\star \;\big|\; ||h||_n \leq \delta\}$
- $\mathcal{N}\left(t; B_n(\delta; \mathcal{F}^\star)\right)$ be the t-convering number of $B_n(\delta; \mathcal{F})$ in norm $||.||_n$.

Critical Inequality via Metric Entropy

Any $\delta \in (0, \infty]$ satisfying

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}\left(t\,; B(\delta; \mathcal{F}^\star)\right)}\; dt \leq \frac{\delta^2}{4\sigma},$$

satisfies the critical inequality.

Bounds via Metric Entropy: Proof

- Construct a $\frac{\delta^2}{4\sigma}$-covering of the set $B(\delta; \mathcal{F})$, say $\{g^1, \ldots, g^M\}$. For any function $g \in B(\delta; \mathcal{F})$, there is a $j \in [M]$ such that $||g^j - g||_n \leq \frac{\delta^2}{4\sigma}$.
- We have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n w_i\, g(\mathbf{x}_i)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n w_i\, g^j(\mathbf{x}_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n w_i \left(g(\mathbf{x}_i) - g^j(\mathbf{x}_i)\right)\right| \\
&\leq \max_{j=1,\ldots,M}\left|\frac{1}{n}\sum_{i=1}^n w_i\, g^j(\mathbf{x}_i)\right| + \sqrt{\frac{\sum_{i=1}^n {w_i}^2}{n}}\sqrt{\frac{\sum_{i=1}^n (g(\mathbf{x}_i) - g^k(\mathbf{x}_i))^2}{n}} \\
&\leq \max_{j=1,\ldots,M}\left|\frac{1}{n}\sum_{i=1}^n w_i\, g^j(\mathbf{x}_i)\right| + \frac{\delta^2}{4\sigma}\sqrt{\frac{\sum_{i=1}^n {w_i}^2}{n}}.
\end{aligned}$$

Hence

$$\mathcal{G}_n(\delta) \leq \mathbb{E}\left[\max_{j=1,\ldots,M}\left|\frac{1}{n}\sum_{i=1}^n w_i\, g^j(\mathbf{x}_i)\right|\right] + \frac{\delta^2}{4\sigma}.$$

Bounds via Metric Entropy: Proof

- To upper bound the first term, we use Dudley's Integral. Define $Z(g^j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i g^j(\mathbf{x}_i)$ for $j = 1, \ldots, M$.
- The induced metric: $\rho_Z^2(g^j, g^k) := \mathrm{Var}\big(Z(g^j) - Z(g^k)\big) = ||g^j - gk||_n^2$
- Since $||g||_n \leq \delta$ for all $g \in B_n(\delta, \mathcal{F}^*)$, the coarsest resolution of the chaining can be set to δ. We can terminate it at $\frac{\delta^2}{4\sigma}$ since any member of our finite set can be reconstructed at this resolution.

$$\mathbb{E}\left[\max_{j=1,\ldots,M} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g^j(\mathbf{x}_i)\right|\right] = \mathbb{E}\left[\max_{j=1,\ldots,M} \frac{\left|Z(g^j)\right|}{\sqrt{n}}\right]$$

$$\leq \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t; B_n(\delta; \mathcal{F}))}\, dt$$

Hence $\mathcal{G}_n(\delta) \leq \frac{\delta^2}{4\sigma} + \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t; B_n(\delta; \mathcal{F}))}\, dt.$ □

Part II

Second Session

Section 5

Review and Examples

Review: Estimator, Models, and Definitions

Model

Let $\mathcal{F}$ be a function class and $f^\star \in \mathcal{F}$ and $\{\mathbf{x}_i\}_{i=1}^n$ be a set of n covariates. Assume that $y_i = f^\star(\mathbf{x}_i) + \sigma w_i$ where $w_i \sim \mathcal{N}(0,1)$.

Estimator

Given the set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - f(\mathbf{x}_i)\Big)^2.$$

Measure of Goodness

Let $\{\mathbf{x}_i\}_{i=1}^n$ be the set of fixed covariates:

$$||f - f^\star||_n := \left[\frac{1}{n} \sum_{i=1}^{n} \big(f(\mathbf{x}_i) - f^\star(\mathbf{x}_i)\big)^2\right]^{1/2}.$$

Local Gaussian Complexity

$$\mathcal{G}_n(\delta)\mathbb{E}_w\left[\sup_{\substack{g\in\mathcal{F}\\ ||g||_n\leq\delta}}\left|\frac{1}{n}\sum_{i=1}^{n}w_i f(\mathbf{x}_i)\right|\right]$$

Critical Inequality

A central object in our analysis is the set of positive scalars δ that satisfy the *critical inequality*:

$$\frac{\delta}{2\sigma}\geq\frac{\mathcal{G}_n(\delta;\mathcal{F}^\star)}{\delta}$$

Existence of Critical Radius

For any star shaped class $\mathcal{F}^\star$, the function $\delta\to\frac{\mathcal{G}_n(\delta,\mathcal{F}^\star)}{\delta}$ is non-increasing on $(0,\infty)$. Consequently, for any $c>0$, the inequality $\frac{\mathcal{G}_n(\delta,\mathcal{F}^\star)}{\delta}\leq c\delta$ has a smallest positive solution.

Review: Main Results

Theorem

Suppose that the shifted function class $\mathcal{F}^\star$ is star-shaped, and let δ be any solution to the critical inequality. Then for any $t \geq \delta$, the nonparametric least-squares estimate $\widehat{f}_n$ satisfies the bound

$$\mathbb{P}\left[||\widehat{f}_n - f^\star||_n^2 \geq 16t\delta\right] \leq \exp\left(\frac{-nt\delta}{2\sigma^2}\right)$$

Auxiliary Lemma

Let $\mathcal{F}^\star$ be a star-shaped class and δ satisfy the critical inequality. Define the bad event $A(u)$ as

$$A(u) = \left\{\exists g \in \mathcal{F}^\star, ||g||_n \geq u \ : \ \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(\mathbf{x}_i)\right| \geq 2u||g||_n\right\}.$$

For a given $u \geq \delta$, we have $\mathbb{P}[A(u)] \leq \exp\left(\frac{-nu^2}{2\sigma^2}\right)$.

Review: How to Compute a Critical Radius

For any star-shaped function class $\mathcal{F}^\star$, define:

- $B(\delta\,; \mathcal{F}^\star) = \{h \in \mathcal{F}^\star \;\big|\; ||h||_n \leq \delta\}$
- $\mathcal{N}\left(t; B(\delta; \mathcal{F}^\star)\right)$ be the t-convering number of $B(\delta; \mathcal{F})$ in norm $||.||_n$.

Critical Inequality via Metric Entropy

Any $\delta \in (0, \infty]$ satisfying

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}\left(t\,; B(\delta; \mathcal{F}^\star)\right)}\, dt \leq \frac{\delta^2}{4\sigma},$$

satisfies the critical inequality.

Parametric Example: Linear Regression

- Consider $y_i = \langle \boldsymbol{\theta}^\star, \mathbf{x}_i \rangle + w_i$ with $\boldsymbol{\theta}^\star \in \mathbb{R}^d$.
- Hypothesis class: $\mathcal{F} = \{ f_{\boldsymbol{\theta}}(.) = \langle \boldsymbol{\theta}, . \rangle \mid \boldsymbol{\theta} \in \mathbb{R}^d \}$.
- Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ be the design matrix with rank r.
- $\mathcal{F} = \mathcal{F}^\star$ and convex. Thus star-shaped around any point.
- $\forall f_{\boldsymbol{\theta}} \in \mathcal{F}$, the function $\boldsymbol{\theta} \to ||f_{\boldsymbol{\theta}}||_n^2 = \frac{||\mathbf{X}\boldsymbol{\theta}||_2^2}{n}$.
- $B(\delta; \mathcal{F})$ is a δ-ball in $\mathrm{Range}(\mathbf{X})$.
- Dimension of $\mathrm{Range}(\mathbf{X})$ is r. By a volume argument:

$$\log \Big(\mathcal{N}\big(t, B(\delta; \mathcal{F})\big) \Big) \leq r \log \big(1 + \frac{2\delta}{t}\big).$$

- We have

$$\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}\big(t; B\big(\delta; \mathcal{F}^\star\big)\big)}\, dt \leq c\delta \sqrt{\frac{r}{n}} \implies ||\widehat{f} - f^\star||_n \leq c\sigma \sqrt{\frac{r}{n}}.$$

Non-Parametric Example: Lipschitz Functions

- Consider the function class

$$\mathcal{F}_{\mathrm{Lip}}(L) = \{f : [0,1] \to \mathbb{R} \,\big|\, f(0) = 0, f \text{ is L-Lip}\}.$$

- Note that $\mathcal{F}_{\mathrm{Lip}}(L) - \mathcal{F}_{\mathrm{Lip}}(L) = 2\mathcal{F}_{\mathrm{Lip}}(L) \subseteq \mathcal{F}_{\mathrm{Lip}}(2L)$.
- The Dudley's integral can be bounded as follows:

$$\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log N_n\Big(t; B_n\big(\mathcal{F}_{\mathrm{Lip}}(2L)\big)\Big)}\, dt \lesssim \frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log N_\infty\Big(t; B_n\big(\mathcal{F}_{\mathrm{Lip}}(2L)\big)\Big)}\, d$$

$$\lesssim \frac{1}{\sqrt{n}} \int_0^{\delta} \Big(\frac{L}{t}\Big)^{\frac{1}{2}} dt \lesssim \sqrt{\frac{L\delta}{n}}.$$

- It suffices to choose $\delta > 0$ such that $\sqrt{\frac{L\delta}{n}} \leq \frac{\delta^2}{\sigma} \implies \delta^2 \simeq \Big(\frac{L\sigma^2}{n}\Big)^{\frac{2}{3}}$.
- According to the main theorem: $||\widehat{f} - f^\star||_n^2 \lesssim \Big(\frac{L\sigma^2}{n}\Big)^{\frac{2}{3}}$ with probability at least $1 - c_1 \exp\Big(-c_2\big(\frac{n}{L\sigma^2}\big)^{\frac{1}{3}}\Big)$.

Section 6

Oracle Inequalities

Oracle Inequalities

- In our analysis so far, we have assumed that $f^\star \in \mathcal{F}$. If we lift this assumption, we have:
  - Prediction Error
  - Approximation Error
- It is natural to measure the approximation error in terms of the best approximating function in $\mathcal{F}$, i.e. $\inf_{f\in\mathcal{F}} ||f - f^\star||_n$.
- Only an oracle that has access to $f^\star(\mathbf{x}_i)$ can calculate $\inf_{f\in\mathcal{F}} ||f - f^\star||_n$. Hence we call inequalities with $\inf_{f\in\mathcal{F}} ||f - f^\star||_n$, the *Oracle Inequalities*.

Oracle Inequalities

Assume that $y_i = f^\star(\mathbf{x}_i) + \sigma w_i$, where $w_i \sim \mathcal{N}(0,1)$.
Define $\partial\mathcal{F} = \{f_1 - f_2 | f_1, f_2 \in \mathcal{F}\}$. We assume that this set is star-shaped.

Oracle Inequality

Let δ be any positive solution to the critical inequality $\frac{\mathcal{G}_n(\delta, \partial\mathcal{F})}{\delta} \leq \frac{\delta}{2\sigma}$.
For any $t \geq \delta$, the non-parametric least-square estimate $\widehat{f}$ satisfies the bound

$$||\widehat{f} - f^\star||_n^2 \leq \inf_{\gamma \in (0,1)} \left\{ \frac{1+\gamma}{1-\gamma} ||f - f^\star||_n^2 + \frac{c_0}{\gamma(1-\gamma)} t\delta \right\} \text{ for all } f \in \mathcal{F},$$

with probability at least $1 - c_1 \exp(-\frac{c_2 n t \delta}{\sigma^2})$.

When $f^\star \in \mathcal{F}$, we can set $f = f^\star$. Hence, $||\widehat{f} - f^\star||_n^2 \lesssim t\delta$, recovering the previous result up to constants.

Oracle Inequalities

$$||\widehat{f} - f^\star||_n^2 \leq \inf_{\gamma \in (0,1)} \left\{ \frac{1+\gamma}{1-\gamma} ||f - f^\star||_n^2 + \frac{c_0}{\gamma(1-\gamma)} t\delta \right\} \text{ for all } f \in \mathcal{F}.$$

- When $f^\star \notin \mathcal{F}$, setting $t = \delta$ and taking infimum over $f \in \mathcal{F}$, yields

$$||\widehat{f} - f^\star||_n^2 \lesssim \inf_{f \in \mathcal{F}} ||f - f^\star||_n^2 + \delta^2$$

with high probability. The term δ^2 can be viewed as the estimation error and $\inf_{f \in \mathcal{F}} ||\widehat{f} - f^\star||_n^2$ as the approximation error.

Oracle Inequalities: Proof

Given an arbitrary $\widetilde{f} \in \mathcal{F}$, define $\widehat{\Delta} = \widehat{f} - f^\star$ and $\widetilde{\Delta} = \widehat{f} - \widetilde{f}$. We have

$$0 \geq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_i\right)^2 - \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widetilde{f}_i\right)^2$$

$$= \frac{1}{2}||\widehat{\Delta}||_n^2 + \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - f_i{}^\star\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\underbrace{(y_i - f_i{}^\star)}_{=\sigma w_i}(f_i{}^\star - \widehat{f}_i) - \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widetilde{f}_i\right)^2$$

Hence $\frac{1}{2}||\widehat{\Delta}||_n^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left(\widetilde{\Delta}_i - \widehat{\Delta}_i\right)(2y_i - \widetilde{f}_i - f_i{}^\star) - \frac{1}{n}\sum_{i=1}^{n}\sigma w_i(f_i^\star - \widehat{f}_i)$.

$$\frac{1}{2}||\widehat{\Delta}||_n^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left[\left(\widetilde{\Delta}_i - \widehat{\Delta}_i\right)(2\sigma w_i - \widetilde{\Delta}_i - \widehat{\Delta}_i) + 2\sigma w_i\widehat{\Delta}_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sigma w_i\widetilde{\Delta}_i + \frac{1}{2}||f_i^\star - \widetilde{f}_i||_n^2 \leq \frac{1}{2}||f_i^\star - \widetilde{f}_i||_n^2 + \frac{\sigma}{n}\left|\sum_{i=1}^{n} w_i\widetilde{\Delta}_i\right|.$$

Oracle Inequalities: Proof

To bound the error, we consider two cases:

- **Case 1**: Suppose that $||\widetilde{\Delta}||_n \leq \sqrt{t\delta}$.

$$||\widehat{\Delta}||_n^2 = ||(\widetilde{f} - f^\star) + \widetilde{\Delta}||_n^2 \leq \left(||\widetilde{f} - f^\star||_n + \sqrt{t\delta}\right)^2$$
$$\leq (1 + 2\beta)||\widetilde{f} - f^\star||_n^2 + \left(1 + \frac{2}{\beta}\right) t\delta.$$

Set $\beta = \frac{\gamma}{1-\gamma}$. This yields

$$||\widehat{f} - f^\star||_n^2 \leq \inf_{\gamma \in (0,1)} \left\{ \frac{1+\gamma}{1-\gamma} ||\widetilde{f} - f^\star||_n^2 + \frac{c_0}{\gamma(1-\gamma)} t\delta \right\}.$$

- **Case 2**: Suppose that $||\widetilde{\Delta}||_n > \sqrt{t\delta}$.

Auxiliary Lemma

Let $\mathcal{H}$ be a star-shaped class and δ satisfy the critical inequality:

$$\forall u \geq \delta: \quad \mathbb{P}\left[\exists g \in \mathcal{H}, ||g||_n \geq u \ : \ \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(\mathbf{x}_i)\right| \geq 2u||g||_n\right] \leq \exp\left(\frac{-nu^2}{2\sigma^2}\right).$$

Let $u = \sqrt{t\delta}$ and $\mathcal{H} = \partial\mathcal{F}$. We have

$$\mathbb{P}\left[2\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{\Delta}(\mathbf{x}_i)\right| \geq 4\sqrt{t\delta}||\widetilde{\Delta}||_n\right] \leq \exp\left(-\frac{nt\delta}{2\sigma^2}\right).$$

Hence, with high probability

$$||\widehat{\Delta}||_n^2 \leq ||\widetilde{f} - f^\star||_n^2 + 4\sqrt{t\delta}||\widetilde{\Delta}||_n^2 \leq ||\widetilde{f} - f^\star||_n^2 + 4\sqrt{t\delta}\left\{||\widehat{\Delta}||_n^2 + ||\widetilde{f} - f^\star||_n^2\right\}$$

$$\leq ||\widetilde{f} - f^\star||_n^2 + \left[4\beta||\widehat{\Delta}||_n^2 + \frac{4}{\beta}t\delta\right] + \left[4\beta||\widetilde{f} - f^\star||_n^2 + \frac{4}{\beta}t\delta\right].$$

Rearrenging terms: $||\widehat{\Delta}||_n^2 \leq \frac{1+4\beta}{1-4\beta}||\widetilde{f} - f^\star||_n^2 + \frac{8}{\beta(1-4\beta)}t\delta$. □

Section 7

Examples

Example: Best Sparse Approximation

- Consider $\mathcal{F}(s) = \{f = \langle \boldsymbol{\theta}, . \rangle \,\big|\, ||\boldsymbol{\theta}||_0 \leq s\}$.
- Disregarding the computational complexity, let

$$\widehat{\theta} \in \arg \min_{||\boldsymbol{\theta}||_0 \leq s} ||y - \mathbf{X}\boldsymbol{\theta}||_0^2.$$

- We will prove that with high probability:

$$||\widehat{f} - f^\star||_n^2 \lesssim \inf_{f \in \mathcal{F}(s)} ||f - f^\star||_n^2 + \underbrace{\frac{\sigma^2 s \log(ed/s)}{n}}_{\delta^2}$$

- The penalty grows linearly with s and logarithmic with d. We only pay a logarithmic price for not knowing the support set in advance.

Example: Best Sparse Approximation

- Note that $\partial\mathcal{F}(s) \subset \mathcal{F}(2s)$. Hence $\mathcal{G}_n(\delta; \partial\mathcal{F}(s)) \leq \mathcal{G}_n(\delta; \mathcal{F}(2s))$.
- Let $S \subseteq \{1, \ldots, d\}$ be an arbitrary $2s$-sized subset.
- Let $\mathbf{X}_S \in \mathbb{R}^{n \times 2s}$ the sub-matrix with columns in S.
- Define

$$Z_n(S) = \sup_{\substack{\boldsymbol{\theta}_S \in \mathbb{R}^{2s} \\ ||\mathbf{X}_S\boldsymbol{\theta}_S||_2/\sqrt{n} \leq \delta}} \left| \frac{\mathbf{w}^T \mathbf{X}_S \boldsymbol{\theta}_S}{n} \right|.$$

- We have $\mathcal{G}_n\big(\delta; \mathcal{F}(2s)\big) = \mathbb{E}_w\Big[\max_{|S|=2s} Z_n(S)\Big]$
- Viewed as a function of $\mathbf{w}$, $Z_n(S)$ is $\frac{\delta}{\sqrt{n}}$ Lipschitz. Hence

$$\mathbb{P}\big[Z_n(S) \geq \mathbb{E}(Z_n(S)) + t\delta\big] \leq \exp\left(\frac{-nt^2}{2}\right)$$

- Consider the SVD of $\mathbf{X}_S = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. We have $||\mathbf{X}_S\theta_S||_2 = ||\mathbf{D}\mathbf{V}^\top\theta_S||_2$:

$$\mathbb{E}\big[Z_n(S)\big] = \mathbb{E}\Big[\sup_{\substack{\beta\in\mathbb{R}^{2s}\\ ||\beta||_2\le\delta}} \big|\frac{1}{\sqrt{n}}\langle \mathbf{U}^\top\mathbf{w},\beta\rangle\big|\Big] \le \frac{\delta}{\sqrt{n}}\mathbb{E}\big[||\mathbf{U}^\top\mathbf{w}||_2\big].$$

- Since $\mathbf{U}$ is orthogonal and $\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_n)$, thus $\mathbf{U}^\top\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_{2s})$.
- Therefore $\mathbb{E}[||\mathbf{U}^\top\mathbf{w}||]\le\sqrt{2s}$. Applying the union bound to the Lip upper bound:

$$\mathbb{P}\Big[\max_{|S|=2s} Z_n(S)\ge\delta\Big(\sqrt{\frac{2s}{n}}+t\Big)\Big]\le\binom{d}{2s}\exp\Big(\frac{-nt^2}{2}\Big).$$

- By integrating:

$$\frac{\mathbb{E}[\max_{|S|=2s} Z_n(S)]}{\delta} = \frac{\mathcal{G}_n(\delta)}{\delta} \lesssim \sqrt{\frac{s}{n}} + \sqrt{\frac{\log\binom{d}{2s}}{n}} \lesssim \sqrt{\frac{s\log\left(\frac{ed}{s}\right)}{n}}.$$

- Thus $\delta\simeq\sigma^2\frac{s\log(ed/s)}{n}$ satisfies critical inequality.

Section 8

Regularized Estimators

Oracle Inequalities for Regularized Estimators

- Given a space $\mathcal{F}$ of real-valued functions, an associated norm $||.||_{\mathcal{F}}$, consider the family of regularized least-square problems:

$$\widehat{f} \in \arg\min_{\mathcal{F}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i) \right)^2 + \lambda_n ||f||_{\mathcal{F}}^2 \right\}.$$

- A Local Gaussian Complexity Measure:

$$\mathcal{G}_n(\delta) \triangleq \mathcal{G}_n\big(\delta; B_{\partial\mathcal{F}}(3)\big) = \mathbb{E}_w \left[\sup_{\substack{g \in \partial\mathcal{F} \\ ||g||_n \leq \delta \\ ||g||_{\mathcal{F}} \leq 3}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i f(\mathbf{x}_i) \right| \right]$$

- The critical inequality for a user defined R:

$$\frac{\mathcal{G}_n(\delta)}{\delta} \leq \frac{R}{2\sigma} \delta.$$

Oracle Inequalities for Regularized Estimators

Theorem

Consider a convex function class $\mathcal{F}$ and assume that δ satisfies the critical inequality

$$\frac{\mathcal{G}_n(\delta)}{\delta} \leq \frac{R}{2\sigma}\delta.$$

If $\lambda_n \geq 2\delta^2$. There exists universal constants c_0, c_1, c_2, c_3 such that

$$||\widehat{f} - f^\star||_n^2 \leq c_0 \inf_{||f||_{\mathcal{F}} \leq R} ||f - f^\star||_n^2 + c_1 R^2(\delta^2 + \lambda_n)$$

with probability grater than $1 - c_2 \exp\big(-c_3 \frac{nR^2\delta^2}{\sigma^2}\big)$.

Oracle Inequalities for Regularized Estimators: Proof

- $y_i = f^\star(\mathbf{x}_i) + \sigma w_i$. Rescale the model by R.
  - Rescaled Noise variance: $\tilde{\sigma}^2 = \left(\frac{\sigma}{R}\right)^2$.
  - Rescaled approx. error: $\inf_{||f||_{\mathcal{F}} \leq 1} ||f - f^\star||_n^2$.
  - The final MSE should be multiplied by R^2.
- Let $\widetilde{f}$ be an arbitrary element in $\mathcal{F}$ with $||\widetilde{f}||_{\mathcal{F}} \leq 1$.
- We have

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \widehat{f}(\mathbf{x}_i)\right)^2 + \lambda_n ||\widehat{f}||_{\mathcal{F}}^2 \leq \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \widetilde{f}(\mathbf{x}_i)\right)^2 + \lambda_n ||\widetilde{f}||_{\mathcal{F}}^2.$$

- Denote $\widehat{\Delta} = \widehat{f} - f^\star$ and $\widetilde{\Delta} = \widehat{f} - \widetilde{f}$. With a simple calculation (next slide):

$$\frac{1}{2}||\widehat{\Delta}||_n^2 \leq \frac{1}{2}||\widetilde{f} - f^\star||_n^2 + \frac{\widetilde{\sigma}}{n}\left|\sum_{i=1}^{n} w_i \widetilde{\Delta}(\mathbf{x}_i)\right| + \lambda_n\left[||\widetilde{f}||_{\mathcal{F}}^2 - ||\widehat{f}||_{\mathcal{F}}^2\right].$$

Given an arbitrary $\widetilde{f} \in \mathcal{F}$, we have

$$
\begin{aligned}
0 &\geq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_i\right)^2 + \lambda_n||\widehat{f}||_{\mathcal{F}}^2 - \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widetilde{f}_i\right)^2 - \lambda_n||\widetilde{f}||_{\mathcal{F}}^2 \\
&= \frac{1}{2}||\widehat{\Delta}||_n^2 + \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - f_i{}^\star\right)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i{}^\star)(f_i{}^\star - \widehat{f}_i) \\
&\qquad - \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widetilde{f}_i\right)^2 + \lambda_n||\widehat{f}||_{\mathcal{F}}^2 - \lambda_n||\widetilde{f}||_{\mathcal{F}}^2
\end{aligned}
$$

Rearranging terms:

$$
\begin{aligned}
\frac{1}{2}||\widehat{\Delta}||_n^2 &\leq \frac{1}{2n}\sum_{i=1}^{n}(\widetilde{\Delta}_i - \widehat{\Delta}_i)(2y_i - \widetilde{f}_i - f_i{}^\star) - \frac{1}{n}\sum_{i=1}^{n}\sigma w_i(f_i^\star - \widehat{f}_i) + \lambda_n||\widetilde{f}||_{\mathcal{F}}^2 - \lambda_n||\widehat{f}||_{\mathcal{F}}^2 \\
&\leq \frac{1}{2n}\sum_{i=1}^{n}\left[(\widetilde{\Delta}_i - \widehat{\Delta}_i)(2\sigma w_i - \widetilde{\Delta}_i - \widehat{\Delta}_i) + 2\sigma w_i\widehat{\Delta}_i\right] + \lambda_n||\widetilde{f}||_{\mathcal{F}}^2 - \lambda_n||\widehat{f}||_{\mathcal{F}}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sigma w_i\widetilde{\Delta}_i + \frac{1}{2}||f_i^\star - \widetilde{f}_i||_n^2 + \lambda_n||\widetilde{f}||_{\mathcal{F}}^2 - \lambda_n||\widehat{f}||_{\mathcal{F}}^2 \\
&\leq \frac{1}{2}||f^\star - \widetilde{f}||_n^2 + \frac{\sigma}{n}\left|\sum_{i=1}^{n}w_i\widetilde{\Delta}_i\right| + \lambda_n||\widetilde{f}||_{\mathcal{F}}^2 - \lambda_n||\widehat{f}||_{\mathcal{F}}^2.
\end{aligned}
$$

Parts of Proof

- If $||\widetilde{\Delta}||_n \leq \sqrt{t\delta}$: It is trivial.
- Assume that $||\widetilde{\Delta}||_n > \sqrt{t\delta}$.
  - **Case 1** Suppose that $||\widehat{f}||_{\mathcal{F}} \leq 2$. This is also very similar to the previous session.
  - **Case 2** Suppose that $||\widehat{f}||_{\mathcal{F}} > 2$. We will prove this statement here.

Case 2: $||\widehat{f}||_{\mathcal{F}} > 2 > 1 \geq ||\widetilde{f}||_{\mathcal{F}}$.

$$\frac{1}{2}||\widehat{\Delta}||_n^2 \leq \frac{1}{2}||f^\star - \widetilde{f}||_n^2 + \frac{\tilde{\sigma}}{n}\Big|\sum_{i=1}^n w_i\widetilde{\Delta}_i\Big| + \lambda_n||\widetilde{f}||_n^2 - \lambda_n||\widehat{f}||_n^2$$

We have

$$||\widetilde{f}||_{\mathcal{F}}^2 - ||\widehat{f}||_{\mathcal{F}}^2 = \underbrace{\Big[||\widetilde{f}||_{\mathcal{F}} + ||\widehat{f}||_{\mathcal{F}}\Big]}_{>1} \cdot \underbrace{\Big[||\widetilde{f}||_{\mathcal{F}} - ||\widehat{f}||_{\mathcal{F}}\Big]}_{<0} \leq \Big[||\widetilde{f}||_{\mathcal{F}} - ||\widehat{f}||_{\mathcal{F}}\Big].$$

Writing $\widehat{f} = \widetilde{f} + \widetilde{\Delta} \implies ||\widehat{f}||_{\mathcal{F}} \geq ||\widetilde{\Delta}||_{\mathcal{F}} - ||\widetilde{f}||_{\mathcal{F}}$:

$$\begin{aligned}\lambda_n\Big[||\widetilde{f}||_{\mathcal{F}}^2 - ||\widehat{f}||_{\mathcal{F}}^2\Big] &\leq \lambda_n\Big[||\widetilde{f}||_{\mathcal{F}} - ||\widehat{f}||_{\mathcal{F}}\Big] \\ &\leq \lambda_n\Big[2||\widetilde{f}||_{\mathcal{F}} - ||\widetilde{\Delta}||_{\mathcal{F}}\Big] \leq \lambda_n\Big[2 - ||\widetilde{\Delta}\,||_{\mathcal{F}}\Big].\end{aligned}$$

Substituting in the basic inequality:

$$\frac{1}{2}||\widehat{\Delta}||_n^2 \leq \frac{1}{2}||f^\star - \widetilde{f}||_n^2 + \frac{\tilde{\sigma}}{n}\Big|\sum_{i=1}^n w_i\widetilde{\Delta}_i\Big| + 2\lambda_n - \lambda_n||\widetilde{\Delta}||_{\mathcal{F}}.$$

Auxiliary Lemma 2

There exists positive constants c_1, c_2 such that with prob. at least $1 - c_1 \exp\left(\frac{-n\delta^2}{c_2\tilde{\sigma}^2}\right)$, we have

$$\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i\,\Delta(\mathbf{x}_i)\right| \leq 2\delta||\Delta||_n + 2\delta^2||\Delta||_{\mathcal{F}} + \frac{1}{16}||\Delta||_n^2,$$

for all $\Delta \in \partial\mathcal{F}$ with $||\Delta||_{\mathcal{F}} \geq 1$.

$||\widetilde{f}||_{\mathcal{F}} \leq 1, ||\widehat{f}||_{\mathcal{F}} > 2$, the trig. ineq. yields: $||\widetilde{\Delta}||_{\mathcal{F}} \geq ||\widehat{f}||_{\mathcal{F}} - ||\widetilde{f}||_{\mathcal{F}} > 1$.

$$\begin{aligned}\frac{1}{2}||\widehat{\Delta}||_n^2 &\leq \frac{1}{2}||f^\star - \widetilde{f}||_n^2 + 2\delta||\widetilde{\Delta}||_n + (2\delta^2 - \lambda_n)||\widetilde{\Delta}||_{\mathcal{F}} + 2\lambda_n + \frac{||\widetilde{\Delta}||_n^2}{16} \\ &\leq \frac{1}{2}||f^\star - \widetilde{f}||_n^2 + \underbrace{2\delta||\widetilde{\Delta}||_n}_{\leq 2\delta||\widetilde{f}-f^\star||_n + 2\delta||\widehat{\Delta}||_n} + 2\lambda_n + \underbrace{\frac{||\widetilde{\Delta}||_n^2}{16}}_{\leq \frac{1}{8}\left[||\widetilde{f}-f^\star||_n^2 + ||\widehat{\Delta}||_n^2\right]}. \quad \square\end{aligned}$$

Proof of Auxiliary Lemma 2

Auxiliary Lemma 2

There exists positive constants c_1, c_2 such that with prob. at least $1 - c_1 \exp\left(\frac{-n\delta^2}{c_2\tilde{\sigma}^2}\right)$, we have

$$\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i\,\Delta(\mathbf{x}_i)\right| \leq 2\delta||\Delta||_n + 2\delta^2||\Delta||_{\mathcal{F}} + \frac{1}{16}||\Delta||_n^2,$$

for all $\Delta \in \partial\mathcal{F}$ with $||\Delta||_{\mathcal{F}} \geq 1$.

It suffices to prove the theorem for $g \in \partial\mathcal{F}$ such that $||g||_{\mathcal{F}} = 1$. Given $\Delta \in \partial\mathcal{F}$ with $||\Delta||_{\mathcal{F}} > 1$, apply the lemma for $g = \frac{\Delta}{||\Delta||_{\mathcal{F}}} \in \mathcal{F}$. Hence:

$$\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i\,\Delta(\mathbf{x}_i)\right| \leq c_1\delta||\Delta||_n + c_2\delta^2||\Delta||_n + \frac{1}{16}\frac{\overbrace{||\Delta||_n^2}^{\leq||\Delta||_n^2}}{||\Delta||_{\mathcal{F}}}.$$

Proof of Auxiliary Lemma 2

- We first consider it over $\{||g||_n \leq t\}$. Define

$$Z_n(t) = \sup_{\substack{||g||_{\mathcal{F}} \leq 1 \\ ||g||_n \leq t}} \left| \frac{\tilde{\sigma}}{n} \sum_{i=1}^{n} w_i \, g(\mathbf{x}_i) \right|.$$

  Lipschitz with constant $\frac{\tilde{\sigma} t}{\sqrt{n}}$. Consequently,

$$\mathbb{P}\big[Z_n(t) \geq \mathbb{E}[Z_n(t)] + u\big] \leq \exp\left(\frac{-nu^2}{2t^2\tilde{\sigma}^2}\right).$$

- Let $t = \delta$. Note that $\mathbb{E}[Z_n(\delta)] \leq \tilde{\sigma}\mathcal{G}_n(\delta) \leq \delta^2$.

$$\mathbb{P}\big[Z_n(\delta) \geq 2\delta^2\big] \leq \exp\left(\frac{-n\delta^2}{2\tilde{\sigma}^2}\right).$$

- Also note that $\mathbb{E}[Z_n(t)] \leq \tilde{\sigma}\mathcal{G}_n(t) = t\frac{\tilde{\sigma}\mathcal{G}_n(t)}{t} \leq t\frac{\tilde{\sigma}\mathcal{G}_n(\delta)}{\delta} \leq t\delta$:

$$\mathbb{P}\left[Z_n(t) \geq t\delta + \frac{t^2}{32}\right] \leq \exp\left(\frac{-c_2 n t^2}{\tilde{\sigma}^2}\right), \text{ for } t > \delta.$$

Proof of Auxiliary Lemma 2

Auxiliary Lemma 2

$$\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i\,\Delta(\mathbf{x}_i)\right| \leq 2\delta||\Delta||_n + 2\delta^2||\Delta||_{\mathcal{F}} + \frac{1}{16}||\Delta||_n^2, \tag{2}$$

for all $\Delta \in \partial\mathcal{F}$ with $||\Delta||_{\mathcal{F}} \geq 1$ with prob. at least $1 - c_1 \exp\left(\frac{-n\delta^2}{c_2\tilde{\sigma}^2}\right)$.

- We will complete the proof by a *peeling* argument.
- Let $\mathcal{E}$ the event that (2) is violated for some $g \in \partial\mathcal{F}$ with $||g||_{\mathcal{F}} = 1$.
- For $a, b \in \mathbb{R}$, let $\mathcal{E}(a, b)$ be the event that (2) is violated for some function such that $||g||_n \in [a, b]$ and $||g||_{\mathcal{F}} = 1$.
- For $m \in \mathbb{I}$, define $t_m = 2^m\delta$. We have $\mathcal{E} = \mathcal{E}(0, t_0) \cup \left(\bigcup_{m=0}^{\infty} \mathcal{E}(t_m, t_{m+1})\right)$.
- Hence, $\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}(0, t_0)] + \sum_{m=0}^{\infty} \mathbb{P}[\mathcal{E}(t_m, t_{m+1})]$.
- Since $t_0 = \delta$, we have $\mathbb{P}[\mathcal{E}(0, t_0)] \leq \mathbb{P}\left[Z_n(\delta) \geq 2\delta^2\right] \leq \exp\left(\frac{-n\delta^2}{2\tilde{\sigma}^2}\right)$.

Proof of Auxiliary Lemma 2

- Assume that $\mathcal{E}(t_m, t_{m+1})$ holds. Meaning there exists g with $||g||_{\mathcal{F}} = 1$, and $||g||_n \in [t_m, t_{m+1}]$, such that

$$\begin{aligned}\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i\, g(\mathbf{x}_i)\right| &\geq 2\delta||g||_n + 2\delta^2 + \frac{1}{16}||g||_n^2 \\ &\geq 2\delta t_m + 2\delta^2 + \frac{1}{8}t_m^2 \\ &= \delta t_{m+1} + 2\delta^2 + \frac{1}{32}t_{m+1}^2.\end{aligned}$$

- This lower bound implies that $Z_n(t_{m+1}) \geq \delta t_{m+1} + \frac{t_{m+1}^2}{32}$. Thus

$$\mathbb{P}[\mathcal{E}(t_m, t_{m+1})] \leq \exp\left(\frac{-c_2 n 2^{2m+2}\delta^2}{\tilde{\sigma}^2}\right)$$

- Wrapping up:

$$\mathbb{P}[\mathcal{E}] \leq \exp\left(\frac{-n\delta^2}{2\tilde{\sigma}^2}\right) + \sum_{m=0}^{\infty}\exp\left(\frac{-c_2 n 2^{2m+2}\delta^2}{\tilde{\sigma}^2}\right) \leq c_1 \exp\left(\frac{-c_2 n\delta^2}{\tilde{\sigma}^2}\right).$$

Section 9

Kernel Ridge Regression

Local Gaussian Complexity of unit Ball of an RKHS

Theorem

Consider an RKHS with kernel k. For a given set of points $\{\mathbf{x}_i\}_{i=1}^n$, let $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n \geq 0$ be the eigenvalues of normalized kernel matrix $\mathbf{K}$ with entries $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)/n$. For all $\delta > 0$, we have

$$\mathbb{E}\left[\sup_{\substack{||f||_{\mathcal{F}} \leq 1 \\ ||f||_n \leq \delta}} \left|\frac{1}{n}\sum_{i=1}^n w_i f(\mathbf{x}_i)\right|\right] \leq \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^n \min\left(\delta^2, \hat{\mu}_j\right)}$$

It suffices to consider functions of the form $g(.) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_i k(., \mathbf{x}_i)$. Any function of the Hilbert space $\mathcal{F}$ can be written as $f = g + g^{\perp}$. We must have $g^{\perp}(\mathbf{x}_i) = \langle g^{\perp}, k(., \mathbf{x})\rangle = 0$. We also have $||f||_{\mathcal{F}}^2 = ||g||_{\mathcal{F}}^2 + ||g^{\perp}||_{\mathcal{F}}^2$. Without loss of generality, we can assume $g^{\perp} = 0$.

- The constraint $||g||_n \leq \delta$ is equivalent to $||\mathbf{K}\alpha||_2 \leq \delta$.
- The constraint $||g||_{\mathcal{F}} \leq 1$ is equivalent to $\alpha^\top \mathbf{K}\alpha \leq 1$.
- The local complexity:

$$\mathcal{G}_n(\delta) = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{\substack{\alpha^\top \mathbf{K}\alpha \leq 1 \\ \alpha \mathbf{K}^2 \alpha \leq \delta^2}} \left| w^\top \mathbf{K}\alpha \right|\right]$$

- $\mathbf{K}$ is PSD, hence $\mathbf{K} = \mathbf{U}^\top \Lambda \mathbf{U}$: $\mathcal{G}_n(\delta) = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{\beta \in D} |w^\top \beta|\right]$ where $D = \left\{\beta \in \mathbb{R}^n \;\middle|\; ||\beta||_2^2 \leq \delta^2, \sum_{j=1}^n \frac{\beta_j^2}{\hat{\mu}_j} \leq 1\right\}$
- Define $\mathcal{E} = \{\beta \in \mathbb{R}^d \;|\; \sum_{j=1}^n \eta_j \beta_j^2 \leq 2\}$, where $\eta_j = \max\{\delta^{-2}, \hat{\mu}_j^{-1}\}$
- For any $\beta \in D$, we have $\sum_{j=1}^n \max\{\delta^{-2}, \hat{\mu}_j^{-1}\}\beta_j^2 \leq \sum_{j=1}^m \frac{\beta_j^2}{\delta^2} + \frac{\beta_j^2}{\hat{\mu}_j} \leq 2$. Hence $D \subseteq D$.
- As a result, by Hölder Inequality

$$\mathcal{G}_n(\delta) \leq \sqrt{\frac{2}{n}}\mathbb{E}\sqrt{\sum_{j=1}^n \frac{w_j^2}{\eta_j}} \leq \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^n \frac{1}{\eta_j}}. \quad \square$$

Corollary

Any $\delta > 0$ satisfying

$$\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n}\min\left(\delta^2, \hat{\mu}_j\right)} \leq \frac{R}{4\sigma}\delta,$$

satisfies the critical inequality.

Some examples from the book here...