Non-Parametric Least Squares

Yu Zhang , Liangchen He

Department of Statistics and Finance, USTC

2022/12/5

1. Problem set-up
2. Bounding the prediction error
3. Oracle inequalities
4. Regularized estimators

1. Problem set-up
   - Different Measures of Quality
   - Estimation via constrained least squares
   - Some examples
2. Bounding the prediction error
3. Oracle inequalities
4. Regularized estimators

- A regression problem is defined by a set of covariates $x \in \mathcal{X}$, along with a response variable $y \in \mathcal{Y}$.
- Our goal is to estimate a function $f : \mathcal{X} \to \mathcal{Y}$ such that the error $y - f(x)$ is as small as possible.
- Mean-squared error (MSE):

$$\overline{\mathcal{L}}_f = \mathbb{E}_{X,Y}\left[(Y - f(X))^2\right] \Longrightarrow f^*(x) = \mathbb{E}[Y \mid X = x]$$

- 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:

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

- Non-Parametric Least Squares : minimizing this least-squares criterion over some suitably controlled function class $\mathcal{F}$. That is

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{L}}_f$$

1. Problem set-up
   - Different Measures of Quality
   - Estimation via constrained least squares
   - Some examples
2. Bounding the prediction error
3. Oracle inequalities
4. Regularized estimators

Given the estimate $\hat{f}$ of the regression, it is natural to measure the difference between the optimal MSE $\overline{\mathcal{L}}_{f^*}$.

- Excess Risk :

$$\overline{\mathcal{L}}_{\hat{f}} - \overline{\mathcal{L}}_{f^*} = \mathbb{E}_X\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] \triangleq \left\|\hat{f} - f^*\right\|_{L^2(\mathbb{P})}^2 \tag{13.4}$$

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

- Sample version: Let $\{x_i\}_{i=1}^n$ be the set of fixed covariates and $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be their empirical measure. Define :

$$\left\|\hat{f} - f^*\right\|_{L^2(\mathbb{P}_n)} = \left[\frac{1}{n}\sum_{i=1}^n \left(\hat{f}(x_i) - f^*(x_i)\right)^2\right]^{1/2} \tag{13.5}$$

  we denote it as $\left\|\hat{f} - f^*\right\|_n$.

1 Problem set-up
Different Measures of Quality
Estimation via constrained least squares
Some examples

2 Bounding the prediction error

3 Oracle inequalities

4 Regularized estimators

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

$$y_i = f^*(\mathbf{x}_i) + \nu_i, \quad \text{for } i = 1, 2, \ldots, n. \tag{13.6}$$

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

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \right\}. \tag{13.7}$$

- When $\nu_i \sim \mathcal{N}\left(0, \sigma^2\right)$, 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:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{F}}^2 \right\}. \tag{13.8}$$

1. Problem set-up
   - Different Measures of Quality
   - Estimation via constrained least squares
   - Some examples
2. Bounding the prediction error
3. Oracle inequalities
4. Regularized estimators

Example : Linear Regression

For a given vector $\theta \in \mathbb{R}^d$, define $f_\theta(x) = < x, \theta >$ and consider the function class $\mathcal{F}_C = \left\{ f_\theta : \mathbb{R}^d \to \mathbb{R} \mid \theta \in C \right\}$ for a compact C.

- The least squares estimate:

$$\hat{\theta} \in \arg\min_{\theta \in C} \left\{ \frac{1}{n} \|y - X\theta\|^2 \right\},$$

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

- The constrained l_q-ball of linear regression:

$$C = \left\{ \theta \in \mathbb{R}^d \mid \|\theta\|_2^q \leq R_q \right\}$$

  for some $q \in [0, 2]$ and radius $R_q > 0$.

Example : Smoothing spline

Consider the class of twice continuously differentiable functions $f : [0,1] \to \mathbb{R}$, define the function class

$$\mathcal{F}(R) = \left\{ f : [0,1] \to \mathbb{R} \mid \int_0^1 (f''(x))^2 \, dx \le R \right\}$$

for a given R, and f'' denotes the second derivative of f. In this case, the penalized form of the nonparametric least-squares estimate is given by

$$\hat{f} \in \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \int_0^1 (f''(x))^2 \right\}$$

where $\lambda_n > 0$ is a user-defined regularization parameter.

1 Problem set-up

2 Bounding the prediction error
Bounds via Gaussian complexity
Bounds via metric entropy
Bounds for high-dimensional parametric problems
Bounds for nonparametric problems

3 Oracle inequalities

4 Regularized estimators

- From a statistical perspective, an essential question associated with the nonparametric least squares estimate (13.7) is how well it approximates the true regression function f^*. Bound the error $\left\|\widehat{f} - f^*\right\|_n$, as measured in the $L^2(\mathbb{P}_n)$ norm.
- Intuitively, the difficulty of estimating the function f^* should depend on the complexity of the function class $\mathcal{F}$ in which it lies. As discussed in Chapter 5, there are a variety of ways of measuring the complexity of a function class, notably by its metric entropy or its Gaussian complexity. We make use of both of these complexity measures in the results to follow.

1 Problem set-up

2 Bounding the prediction error
   Bounds via Gaussian complexity
   Bounds via metric entropy
   Bounds for high-dimensional parametric problems
   Bounds for nonparametric problems

3 Oracle inequalities

4 Regularized estimators

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^*.

Define the set

$$\mathcal{F}^* := \mathcal{F} - \{f^*\} = \{f - f^* \mid f \in \mathcal{F}\},$$

corresponding to an f^*-shifted version of the original function class $\mathcal{F}$.

For a given radius $\delta > 0$, the local Gaussian complexity around f^* at scale δ is given by

$$\mathcal{G}_n(\delta; \mathcal{F}^*) := \mathbb{E}_w \left[\sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_n \le \delta}} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \right],$$

where the variables $\{w_i\}_{i=1}^n$ are i.i.d. $N(0, 1)$.

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

Critical Inequality

$$\frac{\mathcal{G}_n(\delta; \mathcal{F}^*)}{\delta} \leq \frac{\delta}{2\sigma} \tag{13.17}$$

Remark:

- As we verify in Lemma 13.6, whenever the shifted function class $\mathcal{F}^*$ is star-shaped, the left-hand side is a non-increasing function of δ, which ensures that the inequality can be satisfied.
- We refer to any $\delta_n > 0$ satisfying inequality (13.17) as being valid, and we use $\delta_n^* > 0$ to denote the smallest positive radius.

Some Intuition

- Since $\widehat{f}$ and f^* are optimal and feasible, respectively, for the constrained least-squares problem (13.7), we are guaranteed that
$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\widehat{f}(x_i)\right)^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i-f^*(x_i)\right)^2$$
.
- Recalling that $y_i = f^*(x_i) + \sigma w_i$, some simple algebra leads to the equivalent expression (**Basic Inequality**)

$$\frac{1}{2}\left\|\widehat{f}-f^*\right\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^{n} w_i\left(\widehat{f}(x_i)-f^*(x_i)\right) \tag{13.18}$$

- Note that $\widehat{f} - f^* \in \mathcal{F}^*$, and bound the right-hand side by taking the supremum over all functions $g \in \mathcal{F}^*$ with $\|g\|_n \le \left\|\widehat{f} - f^*\right\|_n$.
- Reasoning heuristically, this observation suggests that the squared error $\delta^2 := \mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_n^2\right]$ should satisfy a bound of the form

$$\frac{\delta^2}{2} \le \sigma \mathcal{G}_n(\delta; \mathcal{F}^*) \quad \text{or equivalently} \quad \frac{\delta}{2\sigma} \le \frac{\mathcal{G}_n(\delta; \mathcal{F}^*)}{\delta}$$

- By definition (13.17) of the critical radius δ_n^*, this inequality can only hold for values of $\delta \le \delta_n^*$. In summary, this heuristic argument suggests a bound of the form $\mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_n^2\right] \le (\delta_n^*)^2$.

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

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

Theorem 13.5

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

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

Remarks:

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

$$\operatorname{star}\left(\mathcal{F}^*\right) = \{\alpha\left(f - f^*\right) \mid f \in \mathcal{F}, \alpha \in [0,1]\}$$

- Moreover, since the function f^* is not known to us, we often replace $\mathcal{F}^*$ with the larger class

$$\partial\mathcal{F} := \mathcal{F} - \mathcal{F} = \{f_1 - f_2 \mid f_1, f_2 \in \mathcal{F}\},$$

or its star hull when necessary.

Proof of Theorem 13.5

- **Establishing a basic inequality**
  Denote $\hat{\Delta} = \widehat{f} - f^*$, the basic inequality can be written as

$$\frac{1}{2}\|\hat{\Delta}\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^{n} w_i \hat{\Delta}\left(x_i\right). \tag{13.36}$$

  By definition, the error function $\hat{\Delta} = \widehat{f} - f^*$ belongs to the shifted function class $\mathcal{F}^*$.

- **Controlling the right-hand side**
  Let $\mathcal{H}$ be an arbitrary star-shaped function class, and let $\delta_n > 0$ satisfy the inequality $\frac{\mathcal{G}_n(\delta;\mathcal{H})}{\delta} \leq \frac{\delta}{2\sigma}$. For a given scalar $u \geq \delta_n$, define the event

$$\mathcal{A}(u) := \left\{ \exists g \in \mathcal{H} \cap \{\|g\|_n \geq u\} \left| \frac{\sigma}{n} \sum_{i=1}^{n} w_i g\left(x_i\right) \right| \geq 2\|g\|_n u \right\}$$

  The following lemma provides control on the probability of this event:

Lemma 13.12

For all $u \geq \delta_n$, we have

$$\mathbb{P}[\mathcal{A}(u)] \leq e^{-\frac{nu^2}{2\sigma^2}}.$$

Now consider two cases:

- $\|\hat{\Delta}\|_n < \sqrt{t\delta_n}$, then the claim is immediate.
- $\|\hat{\Delta}\|_n \geq \sqrt{t\delta_n}$, so that we may condition on $\mathcal{A}^c\left(\sqrt{t\delta_n}\right)$ .Set$\mathcal{H} = \mathcal{F}^*$ and $u = \sqrt{t\delta_n}$ for some $t \geq \delta_n$, then we have

$$\mathbb{P}\left[\mathcal{A}^c\left(\sqrt{t\delta_n}\right)\right] \geq 1 - e^{-\frac{n\delta_n}{2\sigma^2}}.$$

so as to obtain the bound

$$\|\hat{\Delta}\|_n^2 \leq 2\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i\hat{\Delta}\left(x_i\right)\right| \leq 4\|\hat{\Delta}\|_n\sqrt{t\delta_n}.$$

Consequently, $\|\hat{\Delta}\|_n^2 \leq 4\|\hat{\Delta}\|_n\sqrt{t\delta_n}$, or equivalently that $\|\hat{\Delta}\|_n^2 \leq 16t\delta_n$, a bound that holds with probability at least $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$.

Proof of Lemma 13.12

1.Reduce the problem to controlling a supremum over a subset of functions satisfying the upper bound $\|\widetilde{g}\|_n \leq u$.

- Suppose that there exists some $g \in \mathcal{H}$ with $\|g\|_n \geq u$ such that
$$\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g\left(x_i\right)\right| \geq 2\|g\|_n u$$
- Define $\widetilde{g} := \frac{u}{\|g\|_n} g$, then $\|\widetilde{g}\|_n = u$. Since $g \in \mathcal{H}$ and $\frac{u}{\|g\|_n} \in (0, 1]$, the star-shaped assumption implies that $\widetilde{g} \in \mathcal{H}$.
$$\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{g}\left(x_i\right)\right| = \frac{u}{\|g\|_n}\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g\left(x_i\right)\right| \geq \frac{u}{\|g\|_n} 2\|g\|_n u = 2u^2$$
- $\mathbb{P}[\mathcal{A}(u)] \leq \mathbb{P}\left[Z_n(u) \geq 2u^2\right]$, where $Z_n(u) := \sup_{\substack{\widetilde{g} \in \mathcal{H} \\ \|\widetilde{g}\|_n \leq u}} \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{g}\left(x_i\right)\right|$

2.Concentration of supremum

- Recall that $w_i \sim \mathcal{N}(0,1)$ are i.i.d., the variable $\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{g}(x_i) \sim N(0, \frac{\sigma^2}{n}\|\widetilde{g}\|_n)$ for each fixed $\widetilde{g}$.
- If we view this supremum as a function of the standard Gaussian vector $(w_1, \ldots, w_n)$: $Z_n(u) = h(w_1, \ldots, w_n) := \sup_{\substack{\widetilde{g}\in\mathcal{H} \\ \|\widetilde{g}\|_n \le u}} \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{g}(x_i)\right|$then the Lipschitz constant of h is at most $\frac{\sigma u}{\sqrt{n}}$.
- Theorem 2.26 guarantees the tail bound $\mathbb{P}[Z_n(u) \ge \mathbb{E}[Z_n(u)] + s] \le e^{-\frac{ns^2}{2u^2\sigma^2}}$, valid for any $s > 0$.
- Setting $s = u^2$ yields

$$\mathbb{P}\left[Z_n(u) \ge \mathbb{E}[Z_n(u)] + u^2\right] \le e^{-\frac{nu^2}{2\sigma^2}}$$

3.Bound the expectation

- By definition of $Z_n(u)$ and $\mathcal{G}_n(u)$, we have $\mathbb{E}\left[Z_n(u)\right] = \sigma\mathcal{G}_n(u)$.
- By Lemma 13.6, the function $v \mapsto \frac{\mathcal{G}_n(v)}{v}$ is non-decreasing, and since $u \geq \delta_n$ by assumption, we have

$$\sigma\frac{\mathcal{G}_n(u)}{u} \leq \sigma\frac{\mathcal{G}_n\left(\delta_n\right)}{\delta_n} \overset{\text{(i)}}{\leq} \delta_n/2 \leq \delta_n,$$

- Then we have shown that $\mathbb{E}\left[Z_n(u)\right] \leq u\delta_n$.

4.Combined with the tail bound (13.41), we obtain

$$\mathbb{P}\left[Z_n(u) \geq 2u^2\right] \overset{\text{(ii)}}{\leq} \mathbb{P}\left[Z_n(u) \geq u\delta_n + u^2\right] \leq e^{-\frac{nu^2}{2\sigma^2}},$$

where step (ii) uses the inequality $u^2 \geq u\delta_n$.

Existence of the critical radius

Lemma 13.6

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

$$\frac{\mathcal{G}_n(\delta;\mathcal{H})}{\delta} \leq c\delta \tag{13.23}$$

has a smallest positive solution.

Proof of Lemma 13.6

Given $0 < \delta \leq t$**, we should show that** $\frac{\delta}{t}\mathcal{G}_n(t) \leq \mathcal{G}_n(\delta)$**:**
Given any $h \in \mathcal{H}^*$ with $\|h\|_n \leq t$, we may define the scaled function $\widetilde{h} = \frac{\delta}{t}h \in \mathcal{H}^*$ and write

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

By construction, $\|\widetilde{h}\|_n \leq \delta, \widetilde{h} \in \mathcal{H}^*$. Consequently, for any $\widetilde{h}$ formed in this way, the right-hand side is at most $\mathcal{G}_n(\delta)$ in expectation. Taking the supremum over the set $\mathcal{H} \cap \{\|h\|_n \leq t\}$ followed by expectations yields $\mathcal{G}_n(t)$ on the left-hand side.
Combining the pieces yields the claim.

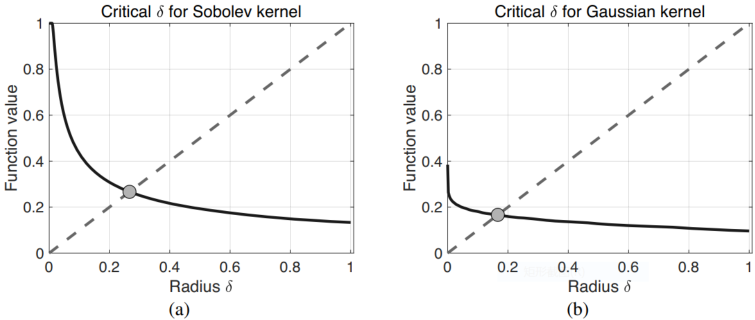

Figure 13.2 Illustration of the critical radius for sample size $n = 100$ and two different function classes. (a) A first-order Sobolev space. (b) A Gaussian kernel class. In both cases, the function $\delta \mapsto \frac{\mathcal{G}_n(\delta;\mathscr{F})}{\delta}$, plotted as a solid line, is non-increasing, 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 $[\delta_n^*, \infty)$.

1 Problem set-up

2 Bounding the prediction error
- Bounds via Gaussian complexity
- Bounds via metric entropy
- Bounds for high-dimensional parametric problems
- Bounds for nonparametric problems

3 Oracle inequalities

4 Regularized estimators

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

- $B_n(\delta; \mathcal{F}^*) = \{h \in \text{star}(\mathcal{F}^*) \mid \|h\|_n \leq \delta\}$, where $\text{star}(\mathcal{F}^*) = \{\alpha f \mid f \in \mathcal{F}^*, \alpha \in [0,1]\}$.
- $\mathcal{N}(t; B_n(\delta; \mathcal{F}^*))$ be the t-covering number of $B_n(\delta; \mathcal{F}^*)$ in the norm $\|\cdot\|_n$

Corollary 13.7 (Critical Inequality via Metric Entropy)

Under the condition of Theorem 13.5, any $\delta \in [0, \sigma)$ such that

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}(t; B_n(\delta; \mathcal{F}^*))} dt \leq \frac{\delta^2}{4\sigma} \tag{13.24}$$

satisfies the critical inequality.

Proof of Corollary 13.7

For any $\delta \in (0, \sigma]$, we have $\frac{\delta^2}{4\sigma} < \delta$, so that we can construct a minimal $\frac{\delta^2}{4\sigma}$-covering of the set $\mathbb{B}_n(\delta; \mathcal{F}^*)$ in the $L^2(\mathbb{P}_n)$-norm, say $\{g^1, \ldots, g^M\}$. For any function $g \in \mathbb{B}_n(\delta; \mathcal{F}^*)$, there is an index $j \in [M]$ such that $\|g^j - g\|_n \leq \frac{\delta^2}{4\sigma}$.

Consequently, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} w_i g(x_i)\right| \underset{\text{(i)}}{\leq} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g^j(x_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} w_i \left(g(x_i) - g^j(x_i)\right)\right|$$

$$\overset{\text{(ii)}}{\leq} \max_{j=1,\ldots,M} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g^j(x_i)\right| + \sqrt{\frac{\sum_{i=1}^{n} w_i^2}{n}} \sqrt{\frac{\sum_{i=1}^{n} \left(g(x_i) - g^j(x_i)\right)^2}{n}}$$

$$\overset{\text{(iii)}}{\leq} \max_{j=1,\ldots,M} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g^j(x_i)\right| + \sqrt{\frac{\sum_{i=1}^{n} w_i^2}{n}} \frac{\delta^2}{4\sigma},$$

Taking the supremum over $g \in \mathbb{B}_n(\delta; \mathcal{F}^*)$ on the left-hand side and then expectation over the noise, we obtain

$$\mathcal{G}_n(\delta) \leq \mathbb{E}_w\left[\max_{j=1,\ldots,M} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g^j(x_i)\right|\right] + \frac{\delta^2}{4\sigma} \tag{13.25}$$

where we have used the fact that $\mathbb{E}_w \sqrt{\frac{\sum_{i=1}^n w_i^2}{n}} \leq 1$.
Upper bound the expected maximum over the M functions in the cover, and we do this by using the chaining method from Chapter 5. Define the family of Gaussian random variables
$Z\left(g^j\right) := \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i g^j\left(x_i\right)$ for $j = 1, \ldots, M$.
Some calculation shows that they are zero-mean, and their associated semi-metric is given by

$$\rho_Z^2\left(g^j, g^k\right) := \operatorname{var}\left(Z\left(g^j\right) - Z\left(g^k\right)\right) = \left\|g^j - g^k\right\|_n^2$$

Since $\|g\|_n \le \delta$ for all $g \in \mathbb{B}_n(\delta; \mathcal{F}^*)$, the coarsest resolution of the chaining can be set to δ, and we can terminate it at $\frac{\delta^2}{4\sigma}$, since any member of our finite set can be reconstructed exactly at this resolution. Working through the chaining argument, we find that

$$\begin{aligned}
\mathbb{E}_w\left[\max_{j=1,\ldots,M}\left|\frac{1}{n}\sum_{i=1}^{n} w_i g^j(x_i)\right|\right] &= \mathbb{E}_w\left[\max_{j=1,\ldots,M}\frac{\left|Z\left(g^j\right)\right|}{\sqrt{n}}\right] \\
&\le \frac{16}{\sqrt{n}}\int_{\frac{\delta^2}{4\sigma}}^{\delta}\sqrt{\log N_n\left(t; \mathbb{B}_n\left(\delta; \mathcal{F}^*\right)\right)}dt
\end{aligned}$$

Combined with our earlier bound (13.25), this establishes the claim.

1 Problem set-up

2 Bounding the prediction error
Bounds via Gaussian complexity
Bounds via metric entropy
Bounds for high-dimensional parametric problems
Bounds for nonparametric problems

3 Oracle inequalities

4 Regularized estimators

Example 13.8 (Bound for linear regression)

Consider the standard linear regression model $y_i = \langle \theta^*, x_i \rangle + w_i$, where $\theta^* \in \mathbb{R}^d$. The function class

$$\mathcal{F}_{\text{lin}} = \left\{ f_\theta(\cdot) = \langle \theta, \cdot \rangle \mid \theta \in \mathbb{R}^d \right\}.$$

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the design matrix, with $x_i \in \mathbb{R}^d$ as its i th row. Use our general theory to show that the least-squares estimate satisfies a bound of the form

$$\left\| f_{\widehat{\theta}} - f_{\theta^*} \right\|_n^2 = \frac{\left\| \mathbf{X} \left(\widehat{\theta} - \theta^* \right) \right\|_2^2}{n} \lesssim \sigma^2 \frac{\operatorname{rank}(\mathbf{X})}{n} \tag{13.26}$$

with high probability.

Observe that the shifted function class $\mathscr{F}_{\text{lin}}^{*}$ is equal to $\mathscr{F}_{\text{lin}}$ for any choice of f^{*}. Moreover, the set $\mathscr{F}_{\text{lin}}$ is convex and hence star-shaped around any point (see Exercise 13.4), so that Corollary 13.7 can be applied.

The mapping $\theta \mapsto \|f_{\theta}\|_{n} = \frac{\|\mathbf{X}\theta\|_{2}}{\sqrt{n}}$ defines a norm on the subspace range $(\mathbf{X})$, and the set $\mathbb{B}_{n}(\delta; \mathscr{F}_{\text{lin}})$ is a δ-ball within the space range $(\mathbf{X})$. By a volume ratio argument (see Example 5.8), we have

$$\log N_{n}(t; \mathbb{B}_{n}(\delta; \mathscr{F}_{\text{lin}})) \leq r \log\left(1 + \frac{2\delta}{t}\right), \quad \text{where } r := \operatorname{rank}(\mathbf{X})$$

Using this upper bound in Corollary 13.7, we find that

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log N_n\left(t ; \mathbb{B}_n\left(\delta ; \mathscr{F}_{\text{lin}}\right)\right.} dt &\leq \sqrt{\frac{r}{n}} \int_0^{\delta} \sqrt{\log \left(1+\frac{2 \delta}{t}\right)} dt \\
&\stackrel{\text{(i)}}{=} \delta \sqrt{\frac{r}{n}} \int_0^1 \sqrt{\log \left(1+\frac{2}{u}\right)} du \\
&\stackrel{\text{(ii)}}{=} c \delta \sqrt{\frac{r}{n}},
\end{aligned}
$$

Putting together the pieces, an application of Corollary 13.7 yields the claim (13.26).

1 Problem set-up

2 Bounding the prediction error
Bounds via Gaussian complexity
Bounds via metric entropy
Bounds for high-dimensional parametric problems
Bounds for nonparametric problems

3 Oracle inequalities

4 Regularized estimators

Example 13.10 (Bounds for Lipschitz functions)

Consider the class of functions

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

Recall that f is L-Lipschitz means that $\left|f(x) - f(x')\right| \leq L\,|x - x'|$ for all $x, x' \in [0,1]$.
Noting the inclusion

$$\mathcal{F}_{\text{Lip}}(L) - \mathcal{F}_{\text{Lip}}(L) = 2\mathcal{F}_{\text{Lip}}(L) \subseteq \mathcal{F}_{\text{Lip}}(2L),$$

it suffices to upper bound the metric entropy of $\mathcal{F}_{\mathrm{Lip}}(2L)$.

Based on our discussion from Example 5.10 , we have

$$\frac{1}{\sqrt{n}}\int_0^{\delta}\sqrt{\log N_n\left(t;\mathbb{B}_n\left(\delta;\mathcal{F}_{\mathrm{Lip}}(2L)\right)\right)}d \lesssim \int_0^{\delta}\sqrt{\log N_\infty\left(t;\mathcal{F}_{\mathrm{Lip}}(2L)\right)}dt$$
$$\lesssim \frac{1}{\sqrt{n}}\int_0^{\delta}(L/t)^{\frac{1}{2}}dt \lesssim \frac{1}{\sqrt{n}}\sqrt{L\delta},$$

Thus, it suffices to choose $\delta_n > 0$ such that $\frac{\sqrt{L\delta_n}}{\sqrt{n}} \lesssim \frac{\delta_n^2}{\sigma}$, or equivalently $\delta_n^2 \gtrsim \left(\frac{L\sigma^2}{n}\right)^{2/3}$. Putting together the pieces, Corollary 13.7 implies that the error in the nonparametric leastsquares estimate satisfies the bound

$$\left\|\widehat{f}-f^*\right\|_n^2 \lesssim \left(\frac{L\sigma^2}{n}\right)^{2/3}$$

with probability at least $1 - c_1 e^{-c_2\left(\frac{n}{L\sigma^2}\right)^{1/3}}$.

1. Problem set-up
2. Bounding the prediction error
3. Oracle inequalities
4. Regularized estimators

- $f^* \notin \mathcal{F}$
- approximation error: $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$
- model: $y_i = f^*(x_i) + \sigma w_i$, where $w_i \sim \mathcal{N}(0, 1)$

Theorem (13.13)

Let δ_n be any positive solution to the inequality

$$\frac{\mathcal{G}_n(\delta; \partial\mathcal{F})}{\delta} \leq \frac{\delta}{2\sigma}. \tag{13.42a}$$

There are universal positive constants (c_0, c_1, c_2) such that for any $t \geq \delta_n$, the nonparametric least-squares estimate $\widehat{f_n}$ satisfies the bound

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq \inf_{\gamma\in(0,1)} \left\{\frac{1+\gamma}{1-\gamma}\left\|f - f^*\right\|_n^2 + \frac{c_0}{\gamma(1-\gamma)} t\delta_n\right\} \quad \text{for all } f \in \mathcal{F} \tag{13.42b}$$

with probability greater than $1 - c_1 e^{-c_2 \frac{nt\delta_n}{\sigma^2}}$.

Proof

- $\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2 \leq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \widetilde{f}(x_i)\right)^2$
- $\widehat{\Delta} := \widehat{f} - f^*, \widetilde{\Delta} := \widehat{f} - \widetilde{f},$

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}\left\|\widetilde{f} - f^*\right\|_n^2 + \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{\Delta}(x_i)\right| \tag{13.51}$$

- $\|\widetilde{\Delta}\|_n \leq \sqrt{t\delta_n}$

$$\begin{aligned}
\|\widehat{\Delta}\|_n^2 = \left\|\widehat{f} - f^*\right\|_n^2 &= \left\|\left(\widetilde{f} - f^*\right) + \widetilde{\Delta}\right\|_n^2 \\
&\overset{\text{(i)}}{\leq} \left\{\left\|\widetilde{f} - f^*\right\|_n + \sqrt{t\delta_n}\right\}^2 \\
&\overset{\text{(ii)}}{\leq} (1+2\beta)\left\|\widetilde{f} - f^*\right\|_n^2 + \left(1 + \frac{2}{\beta}\right) t\delta_n \text{ for any } \beta > 0 \\
&\text{setting } \beta = \frac{\gamma}{1-\gamma} \text{ for some } \gamma \in (0,1)
\end{aligned}$$

- $\|\widetilde{\Delta}\|_n > \sqrt{t\delta_n}$

$$\mathbb{P}\left[2\left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i\widetilde{\Delta}\left(x_i\right)\right| \geq 4\sqrt{t\delta_n}\|\widetilde{\Delta}\|_n\right] \leq e^{-\frac{n\delta_n}{2\sigma n}} \text{ by lemma 13.12,}$$

$$\begin{aligned}\|\widehat{\Delta}\|_n^2 &\leq \left\|\widetilde{f} - f^*\right\|_n^2 + 4\sqrt{t\delta_n}\|\widetilde{\Delta}\|_n \\ &\leq \left\|\widetilde{f} - f^*\right\|_n^2 + 4\sqrt{t\delta_n}\left\{\|\widehat{\Delta}\|_n + \left\|\widetilde{f} - f^*\right\|_n\right\}\end{aligned}$$

with probability at least $1 - 2e^{-\frac{n\Delta s_n}{2\sigma^2}}$ by (13.51).

$$\|\widehat{\Delta}\|_n^2 \leq (1 + 4\beta)\left\|\widetilde{f} - f^*\right\|_n^2 + 4\beta\|\widehat{\Delta}\|_n^2 + \frac{8}{\beta}t\delta_n,$$

rearranging and setting $\gamma = 4\beta$.

1. Problem set-up
2. Bounding the prediction error
3. Oracle inequalities
4. **Regularized estimators**

Given a space $\mathcal{F}$ of real-valued functions with an associated semi-norm $\|\cdot\|_{\mathcal{F}}$, consider the family of regularized least-squares problems

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{F}}^2 \right\}. \tag{13.52}$$

local Gaussian complexity:

$$\mathcal{G}_n(\delta; \mathbb{B}_{\partial\mathcal{F}}(3)) := \mathbb{E}_w \left[\sup_{\substack{g \in \partial\mathcal{F} \\ \|g\|_{\mathcal{F}} \leq 3, \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i f(x_i) \right| \right], \tag{13.53}$$

For a user-defined radius $R > 0$, we let $\delta_n > 0$ be any number satisfying the inequality

$$\frac{\mathcal{G}_n(\delta)}{\delta} \leq \frac{R}{2\sigma}\delta. \tag{13.54}$$

Theorem (13.17)

Given the previously described observation model and a convex function class $\mathcal{F}$, suppose that we solve the convex program (13.52) with some regularization parameter $\lambda_n \geq 2\delta_n^2$. Then there are universal positive constants $\left(c_j, c_j'\right)$ such that

$$\left\|\widehat{f} - f^*\right\|_n^2 \leq c_0 \inf_{\|f\|\mathcal{F}_{\mathcal{F}} \leq R} \left\|f - f^*\right\|_n^2 + c_1 R^2 \left\{\delta_n^2 + \lambda_n\right\} \tag{13.55a}$$

with probability greater than $1 - c_2 e^{-c_3 \frac{nR^2\delta_n^2}{\sigma^2}}$. Similarly, we have

$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 \leq c_0' \inf_{\|f\|\|\|_{\mathcal{F}} \leq R} \left\|f - f^*\right\|_n^2 + | \; c_1' R^2 \left\{\delta_n^2 + \lambda_n\right\}. \tag{13.55b}$$

Proof

We introduce the shorthand $\tilde{\sigma} = \sigma / R$. Let $\widetilde{f}$ be any element of $\mathcal{F}$ such that $\|\widetilde{f}\|_{\mathcal{F}} \leq 1$. At the end of the proof, we optimize this choice.

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

modified basic inequality:

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}\left\|\widetilde{f} - f^*\right\|_n^2 + \frac{\tilde{\sigma}}{n}\left|\sum_{i=1}^{n} w_i \widetilde{\Delta}(x_i)\right| + \lambda_n\left\{\|\widetilde{f}\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2\right\} \quad (13.59)$$

$$\leq \frac{1}{2}\left\|\widetilde{f} - f^*\right\|_n^2 + \frac{\tilde{\sigma}}{n}\left|\sum_{i=1}^{n} w_i \widetilde{\Delta}(x_i)\right| + \lambda_n, \quad (13.60)$$

- $\|\widetilde{\Delta}\|_n \leq \sqrt{t\delta_n}$, same as theorem 13.13.
- $\|\widetilde{\Delta}\|_n > \sqrt{t\delta_n}$
  - $\|\widehat{f}\|_{\mathcal{F}} \leq 2$, then we have $\|\widetilde{\Delta}\|_{\mathcal{F}} \leq 3$. By applying Lemma 13.12 over the set of functions $\{g \in \partial\mathcal{F} \mid \|g\|_{\mathcal{F}} \leq 3\}$, we conclude

$$\frac{\tilde{\sigma}}{n}\left|\sum_{i=1}^{n} w_i \widetilde{\Delta}(x_i)\right| \leq c_0 \sqrt{t\delta_n}\|\widetilde{\Delta}\|_n \quad \text{with probability at least } 1 - e^{-\frac{t^2}{2\sigma^2}}.$$

We also have

$$\begin{aligned} 2\sqrt{t\delta_n}\|\widetilde{\Delta}\|_n &\leq 2\sqrt{t\delta_n}\|\widehat{\Delta}\|_n + 2\sqrt{t\delta_n}\left\|\widetilde{f} - f^*\right\|_n \\ &\leq 2\sqrt{t\delta_n}\|\widehat{\Delta}\|_n + 2t\delta_n + \frac{\left\|\widetilde{f} - f^*\right\|_n^2}{2}, \end{aligned}$$

Consequently,

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}(1 + c_0)\left\|\widetilde{f} - f^*\right\|_n^2 + 2c_0 t\delta_n + 2c_0\sqrt{t\delta_n}\|\widehat{\Delta}\|_n + \lambda_n.$$

- $\|\widehat{f}\|_{\mathcal{F}} > 2$, we have

$$\|\widetilde{f}\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2 = \underbrace{\|\widetilde{f}\|_{\mathcal{F}} + \|\widehat{f}\|_{\mathcal{F}}\}}_{>1} \underbrace{\left\{\|\widetilde{f}\|_{\mathcal{F}} - \|\widehat{f}\|_{\mathcal{F}}\right\}}_{<0} \leq \underbrace{\left\{\|\widetilde{f}\|_{\mathcal{F}} - \|\widehat{f}\|_{\mathcal{F}}\right\}}_{<0}.$$

Then we obtain

$$\begin{aligned}
\lambda_n \left\{\|\widetilde{f}\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2\right\} &\leq \lambda_n \left\{\|\widetilde{f}\|_{\mathcal{F}} - \|\widehat{f}\|_{\mathcal{F}}\right\} \\
&\leq \lambda_n \left\{2\|\widetilde{f}\|_{\mathcal{F}} - \|\widetilde{\Delta}\|_{\mathcal{F}}\right\} \\
&\leq \lambda_n \left\{2 - \|\widetilde{\Delta}\|_{\mathcal{F}}\right\},
\end{aligned}$$

Now we get

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}\left\|\widetilde{f} - f^*\right\|_n^2 + \left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i \widetilde{\Delta}\left(x_i\right)\right| + 2\lambda_n - \lambda_n\|\widetilde{\Delta}\|_{\mathcal{F}}. \tag{13.62}$$

Lemma (13.23)

There are universal positive constants (c_1, c_2) such that, with probability greater than $1 - c_1 e^{-\frac{n\delta_n^2}{c_2\tilde{\sigma}^2}}$, we have

$$\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i \Delta(x_i)\right| \le 2\delta_n \|\Delta\|_n + 2\delta_n^2 \|\Delta\|_{\mathcal{F}} + \frac{1}{16}\|\Delta\|_n^2, \tag{13.63}$$

a bound that holds uniformly for all $\Delta \in \partial\mathcal{F}$ with $\|\Delta\|_{\mathcal{F}} \ge 1$.

Since $\|\widetilde{\Delta}\|_{\mathcal{F}} \geq \|\widehat{f}\|_{\mathcal{F}} - \|\widetilde{f}\|_{\mathcal{F}} > 1$, applying lemma 13.23 and substituting (13.63) into (13.62) yields

$$\begin{aligned}
\frac{1}{2}\|\widehat{\Delta}\|_n^2 &\leq \frac{1}{2}\left\|\widetilde{f} - f^*\right\|_n^2 + 2\delta_n\|\widetilde{\Delta}\|_n + \left\{2\delta_n^2 - \lambda_n\right\}\|\widetilde{\Delta}\|_{\mathcal{F}} + 2\lambda_n + \frac{\|\widetilde{\Delta}\|_n^2}{16} \\
&\leq \frac{1}{2}\left\|\widetilde{f} - f^*\right\|_n^2 + 2\delta_n\|\widetilde{\Delta}\|_n + 2\lambda_n + \frac{\|\widetilde{\Delta}\|_n^2}{16}.
\end{aligned}$$

$$2\delta_n\|\widetilde{\Delta}\|_n \leq 2\delta_n\left\|\widetilde{f} - f^*\right\|_n + 2\delta_n\|\widehat{\Delta}\|_n,$$

$$\frac{\|\widetilde{\Delta}\|_n^2}{16} \leq \frac{1}{8}\left\{\left\|\widetilde{f} - f^*\right\|_n^2 + \|\widehat{\Delta}\|_n^2\right\}.$$

$$\Rightarrow \left\{\frac{1}{2} - \frac{1}{8}\right\}\|\widehat{\Delta}\|_n^2 \leq \left\{\frac{1}{2} + \frac{1}{8}\right\}\left\|\widetilde{f} - f^*\right\|_n^2 + 2\delta_n\left\|\widetilde{f} - f^*\right\|_n + 2\delta_n\|\widehat{\Delta}\|_n + 2\lambda_n.$$

Proof of lemma 13.23

We claim that it suffices to prove the bound (13.63) for functions $g \in \partial\mathcal{F}$ such that $\|g\|_{\mathcal{F}} = 1$. Indeed, suppose that it holds for all such functions, and that we are given a function Δ with $\|\Delta\|_{\mathcal{F}} > 1$. By assumption, we can apply the inequality (13.63) to the new function $g := \Delta/\|\Delta\|_{\mathcal{F}}$, which belongs to $\partial\mathcal{F}$ by the star-shaped assumption. Applying the bound (13.63) to g and then multiplying both sides by $\|\Delta\|_{\mathcal{F}}$, we obtain

$$\begin{aligned}\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i \Delta\left(x_i\right)\right| &\leq c_1\delta_n\|\Delta\|_n + c_2\delta_n^2\|\Delta\|_{\mathcal{F}} + \frac{1}{16}\frac{\|\Delta\|_n^2}{\|\Delta\|_{\mathcal{F}}} \\ &\leq c_1\delta_n\|\Delta\|_n + c_2\delta_n^2\|\Delta\|_{\mathcal{F}} + \frac{1}{16}\|\Delta\|_n^2.\end{aligned}$$

In order to establish the bound (13.63) for functions with $\|g\|_{\mathcal{F}} = 1$, we first consider it over the ball $\{\|g\|_n \leq t\}$, for some fixed radius $t > 0$. Define the random variable

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

Viewed as a function of the standard Gaussian vector w , it is Lipschitz with parameter at most $\tilde{\sigma} t / \sqrt{n}$. Consequently, Theorem 2.26 implies that

$$\mathbb{P}\left[Z_n(t) \geq \mathbb{E}\left[Z_n(t)\right] + u\right] \leq e^{-\frac{nu^2}{2\tilde{\sigma}^2 t^2}} . \tag{13.66}$$

We first derive a bound for $t = \delta_n$. By the definitions of $\mathcal{G}_n$ and the critical radius, we have $\mathbb{E}\left[Z_n\left(\delta_n\right)\right] \leq \tilde{\sigma}\mathcal{G}_n\left(\delta_n\right) \leq \delta_n^2$. Setting $u = \delta_n$ in the tail bound (13.66), we find that

$$\mathbb{P}\left[Z_n\left(\delta_n\right) \geq 2\delta_n^2\right] \leq e^{-\frac{n\delta_n^2}{2\tilde{\sigma}^2}}. \tag{13.67a}$$

On the other hand, for any $t > \delta_n$, we have

$$\mathbb{E}\left[Z_n(t)\right] = \tilde{\sigma}\mathcal{G}_n(t) = t\frac{\tilde{\sigma}\mathcal{G}_n(t)}{t} \overset{\text{(i)}}{\leq} t\frac{\tilde{\sigma}\mathcal{G}_n\left(\delta_n\right)}{\delta_n} \overset{\text{(ii)}}{\leq} t\delta_n,$$

Using this upper bound on the mean and setting $u = t^2/32$ in the tail bound (13.66) yields

$$\mathbb{P}\left[Z_n(t) \geq t\delta_n + \frac{t^2}{32}\right] \leq e^{-c\frac{n^2}{\sigma^2}} \quad \text{for each } t > \delta_n. \tag{13.67b}$$

Peeling

Let $\mathcal{E}$ denote the event that the bound (13.63) is violated for some function $g \in \partial\mathcal{F}$ with $\|g\|_{\mathcal{F}} = 1$. For real numbers $0 \leq a < b$, let $\mathcal{E}(a, b)$ denote the event that it is violated for some function such that $\|g\|_n \in [a, b]$ and $\|g\|_{\mathcal{F}} = 1$. For $m = 0, 1, 2, \ldots$, define $t_m = 2^m \delta_n$. We then have the decomposition $\mathcal{E} = \mathcal{E}(0, t_0) \cup \left(\bigcup_{m=0}^{\infty} \mathcal{E}(t_m, t_{m+1})\right)$ and hence, by the union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}(0, t_0)] + \sum_{m=0}^{\infty} \mathbb{P}[\mathcal{E}(t_m, t_{m+1})]. \tag{13.68}$$

The final step is to bound each of the terms in this summation. Since $t_0 = \delta_n$, we have

$$\mathbb{P}[\mathcal{E}(0, t_0)] \leq \mathbb{P}\left[Z_n(\delta_n) \geq 2\delta_n^2\right] \leq e^{-\frac{n\delta_n^2}{2\sigma^2}}, \tag{13.69}$$

using our earlier tail bound (13.67a). On the other hand, suppose that $\mathcal{E}(t_m, t_{m+1})$ holds, meaning that there exists some function g with $\|g\|_{\mathcal{F}} = 1$ and $\|g\|_n \in [t_m, t_{m+1}]$ such that

$$\left|\frac{\tilde{\sigma}}{n}\sum_{i=1}^{n} w_i g(x_i)\right| \geq 2\delta_n\|g\|_n + 2\delta_n^2 + \frac{1}{16}\|g\|_n^2 \stackrel{\text{(i)}}{\geq} 2\delta_n t_m + 2\delta_n^2 + \frac{1}{8}t_m^2 \stackrel{\text{(ii)}}{=} \delta_n t_{m+1} + 2\delta_n^2 + \frac{1}{32}t_{m+1}^2$$

where step (i) follows since $\|g\|_n \geq t_m$, and step (ii) follows since $t_{m+1} = 2t_m$. This lower bound implies that $Z_n(t_{m+1}) \geq \delta_n t_{m+1} + \frac{t_{m+1}^2}{32}$, and applying the tail bound (13.67b) yields

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

Substituting this inequality and our earlier bound (13.69) into equation (13.68) yields

$$\mathbb{P}[\mathcal{E}] \leq e^{-\frac{n\delta_n^2}{2\tilde{\sigma}^2}} + \sum_{m=0}^{\infty} e^{-c_2 \frac{n 2^{2m+2} \delta_n^2}{\tilde{\sigma}^2}} \leq c_1 e^{-c_2 \frac{n\delta_n^2}{\tilde{\sigma}^2}}.$$

Corollary (13.18)

For the KRR estimate (12.28), the bounds of Theorem 13.17 hold for any $\delta_n > 0$ satisfying the inequality

$$\sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^{n}\min\left\{\delta^2, \hat{\mu}_j\right\}} \leq \frac{R}{4\sigma}\delta^2. \tag{13.56}$$

Example 13.21 (Gaussian kernel)

The Gaussian kernel $\mathcal{K}(x,z) = e^{-\frac{(x-z)^2}{2\sigma^2}}$ on the square $[-1,1] \times [-1,1]$. As discussed in Example 12.25 , the eigenvalues of the associated kernel operator scale as $\mu_j \simeq e^{-cj\log j} asj \to +\infty$. Accordingly, let us adopt the heuristic that the empirical eigenvalues satisfy a bound of the form $\hat{\mu}_j \leq c_0 e^{-c_1 j \log j}$. For a given $\delta > 0$, we have

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{n} \min\left\{\delta^2, \hat{\mu}_j\right\}} &\leq \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{n} \min\left\{\delta^2, c_0 e^{-c_1 j \log j}\right\}} \\
&\leq \frac{1}{\sqrt{n}} \sqrt{k\delta^2 + c_0 \sum_{j=k+1}^{n} e^{-c_1 j \log j}}
\end{aligned}
$$

where k is the smallest positive integer such that $c_0 e^{-c_1 k \log k} \leq \delta^2$. Some algebra shows that the critical inequality will be satisfied by $\delta_n^2 \simeq \frac{\sigma^2}{R^2} \frac{\log\left(\frac{Rn}{\sigma}\right)}{n}$, so that nonparametric regression over the Gaussian kernel class satisfies the bound

$$\left\|\widehat{f} - f^*\right\|_n^2 \lesssim \inf_{\|f\|_H \leq R} \left\|f - f^*\right\|_n^2 + R^2 \delta_n^2 = \inf_{\|f\|_H \leq R} \left\|f - f^*\right\|_n^2 + c\sigma^2 \frac{\log\left(\frac{Rn}{\sigma}\right)}{n}.$$