

2 Non-parametric least squares

3 In this chapter, we consider the problem of non-parametric regression, in which the goal
 4 is to estimate a (possibly non-linear) function on the basis of noisy observations. Using
 5 results developed in previous chapters, we analyze the convergence rates of procedures
 6 based on solving non-parametric versions of least-squares problems.

7 ■ 13.1 Problem set-up

Regression is a type of prediction problem, in which the goal is to use observations of predictors or covariates $x \in \mathcal{X}$ in order to predict a response variable $y \in \mathcal{Y}$. Throughout this chapter, we focus on the case of real-valued outputs, in which the space \mathcal{Y} is the real line or some subset thereof. Our goal is to estimate a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that the error $y - f(x)$ is as small as possible over some range of pairs (x, y) . In the *random design* version of regression, we model both the response and covariate as random quantities, in which case it is reasonable to measure the quality of f in terms of its *mean-squared error*

$$\bar{\mathcal{L}}_f := \mathbb{E}_{X,Y}[(Y - f(X))^2]. \quad (13.1)$$

The function f^* minimizing this criterion is known as the *Bayes' least-squares estimate* or the *regression function*, and it is given by the conditional expectation

$$f^*(x) = \mathbb{E}[Y \mid X = x], \quad (13.2)$$

8 assuming that all relevant expectations exist. See Exercise 13.1 for further details.

In practice, the expectation defining the MSE (13.1) cannot be computed, since the distribution over (X, Y) is not known. Instead, we are given a collection of samples $\{(x_i, y_i)\}_{i=1}^n$, which can be used to compute an empirical analogue of the mean-squared error, namely

$$\hat{\mathcal{L}}_f := \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2. \quad (13.3)$$

9 The method of *non-parametric least squares*, to be discussed in detail in this chapter, is
 10 based on minimizing this least-squares criterion over some suitably controlled function
 11 class.

12 ■ 13.1.1 Different measures of quality

Given an estimate f of the regression function, it is natural to measure its quality in terms of the *excess risk*—namely, the difference between the optimal MSE achieved by f^* and that achieved by the given estimate f . In the special case of the least-squares cost function, it can be shown (see Exercise 13.1) that this excess risk can be written as

$$\bar{\mathcal{L}}_f - \bar{\mathcal{L}}_{f^*} = \underbrace{\mathbb{E}_X[(f(X) - f^*(X))^2]}_{\|f^* - f\|_{L^2(\mathbb{P})}^2}, \quad (13.4)$$

13 where \mathbb{P} denotes the distribution over the covariates. We frequently adopt the shorthand
 14 notation $\|f - f^*\|_2$ for the $L^2(\mathbb{P})$ norm when this underlying distribution is clear from
 the context. 1

In this chapter, we measure the error using a closely related but slightly different measure, one that is defined by the samples $\{x_i\}_{i=1}^n$ of the covariates. In particular, they define the empirical distribution $\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ that places a weight $1/n$ on each sample, and the associated $L^2(\mathbb{P}_n)$ norm is given by

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

In order to lighten notation, we frequently use $\|\hat{f} - f^*\|_n$ as a shorthand for the more
 cumbersome $\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}$. Throughout the remainder of this chapter, we will view the
 samples $\{x_i\}_{i=1}^n$ as being fixed, a set-up known as regression with a *fixed design*. The
 theory in this chapter focuses on error bounds in terms of the empirical $L^2(\mathbb{P}_n)$ -norm.
 Results from Chapter 14 to follow can be used to translate these bounds into equivalent
 results in the population $L^2(\mathbb{P})$ -norm. 2
3
4
5
6
7

■ 13.1.2 Estimation via constrained least-squares 8

Given a fixed collection $\{x_i\}_{i=1}^n$ of fixed design points, the associated response variables $\{y_i\}_{i=1}^n$ can always be written in the generative form

$$y_i = f^*(x_i) + v_i, \quad \text{for } i = 1, 2, \dots, n, \quad (13.6)$$

where v_i is a random variable representing the “noise” in the i^{th} response variable. Note
 1 that these noise variables must be zero-mean, given the form (13.2) of the regression 9

function f^* . Otherwise, their structure in general depends on the form of the conditional distribution of the random variable $(Y | X = x)$. In the *standard non-parametric regression* model, we assume that an i.i.d. sequence of the form $v_i = \sigma w_i$, where $w_i \sim \mathcal{N}(0, 1)$ is a standard Gaussian variate, and σ is a standard deviation parameter.

Given this set-up, one way in which to estimate the regression function f^* is by constrained least-squares—that is, by solving the problem¹

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

where \mathcal{F} is a suitably chosen subset of functions. When $v_i \sim \mathcal{N}(0, \sigma^2)$, note that the criterion (13.7) is equivalent to constrained maximum likelihood, apart from the scale factor $1/\sigma^2$. However, as with least-squares regression in the parametric setting, the estimator is far more generally applicable.

At least in general, the optimization problem (13.7) defining the non-parametric least squares estimator \hat{f} is infinite-dimensional in nature, since f ranges over a potentially infinite family \mathcal{F} . Typically, we choose \mathcal{F} to be a compact subset of some ambient function class \mathcal{G} —for instance, a ball of radius R in some norm $\|\cdot\|_{\mathcal{G}}$. In certain settings, the ambient function class \mathcal{G} might be a Hilbert space, or even a reproducing kernel Hilbert space, as discussed in Chapter 12. Moreover, for computational reasons, it can be convenient to use regularized estimators of the form

$$\hat{f} \in \arg \min_{f \in \mathcal{G}} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{G}}^2 \right\}, \quad (13.8)$$

where $\lambda_n > 0$ is a suitably chosen regularization weight. See Section 13.4 for analysis of such estimators.

■ 13.1.3 Some examples

Let us illustrate the estimators (13.7) and (13.8) with some examples.

Example 13.1 (Linear regression). For a given vector $\theta \in \mathbb{R}^d$, define the function $f_{\theta}(x) = \langle \theta, x \rangle$. For a compact subset $\mathcal{C} \subseteq \mathbb{R}^d$, define $\mathcal{F}_{\mathcal{C}} := \{f_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R} \mid \theta \in \mathcal{C}\}$. With this choice, the estimator (13.7) reduces to a constrained form of least-squares estimation—namely

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

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix with the vector $x_i \in \mathbb{R}^d$ in its i^{th} row. Particular

¹Although the renormalization by $(2n)^{-1}$ in the definition (13.7) of the non-parametric least squares estimator has no consequence on \hat{f} , we do so in order to emphasize the connection between this method and the $L^2(\mathbb{P}_n)$ -norm.

instances of this estimator include *ridge regression*, obtained by setting

$$\mathcal{C} = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2^2 \leq R_2\}$$

for some (squared) radius $R_2 > 0$. More generally, this class of estimators contains all the *constrained ℓ_q -ball* estimators, obtained by setting

$$\mathcal{C} = \{\theta \in \mathbb{R}^d \mid \sum_{j=1}^d |\theta_j|^q \leq R_q\}$$

- 14 for some $q \in [0, 2]$ and radius $R_q > 0$. See Figure 7-1 for an illustration of these sets. The constrained form of the Lasso (7.20), as analyzed in depth in Chapter 7, is a special but important case, obtained by setting $q = 1$. 1 2

Whereas the previous example was a parametric problem, we now turn to some non-parametric examples: 3 4

Example 13.2 (Cubic smoothing spline). Consider the class of twice continuously differentiable functions $f : [0, 1] \rightarrow \mathbb{R}$, and for a given squared radius $R > 0$, define the function class

$$\mathcal{F}(R) := \{f : [0, 1] \rightarrow \mathbb{R} \mid \int_0^1 (f''(x))^2 dx \leq R\}, \quad (13.9)$$

where f'' denotes the second derivative of f . Note that this constraint can be understood as a Hilbert norm bound in a second-order Sobolev space (recall Example 12.7). For this function class, the penalized form of the non-parametric least squares estimate is given by


$$\hat{f} \in \arg \min_f \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \int_0^1 (f''(x))^2 dx \right\}, \quad (13.10)$$

where $\lambda_n > 0$ is a user-defined regularization parameter. It can be shown that any minimizer \hat{f} is a cubic spline, meaning that it is a piecewise cubic function, with the third derivative changing at each of the distinct design points x_i . In the limit as $R \rightarrow 0$ (or equivalently, as $\lambda_n \rightarrow +\infty$), the cubic spline fit \hat{f} becomes a linear function, since we have $f'' = 0$ only for a linear function. ♣ 5 6 7 8 9

The spline estimator in the previous example turns out to be a special case of a more general class of estimators, based on regularization in a reproducing kernel Hilbert space (see Chapter 12 for background). 10 11 12 13

Example 13.3 (Kernel ridge regression). Let \mathbb{H} be a reproducing kernel Hilbert space, equipped with the norm $\|\cdot\|_{\mathbb{H}}$, and for some radius $R > 0$, consider the estimator

$$\hat{f} \in \arg \min_{\|f\|_{\mathbb{H}} \leq R} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 \right\}.$$

As discussed in Chapter 12, the computation of this estimate can be reduced to solving a quadratic program involving the empirical kernel matrix defined by the design points $\{x_i\}_{i=1}^n$. In particular, if we define the kernel matrix with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$, then the solution takes the form $\hat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\alpha}_i \mathcal{K}(\cdot, x_i)$, where $\hat{\alpha} \in \mathbb{R}^n$ is a solution to the constrained quadratic program $\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|y - \sqrt{n} \mathbf{K} \alpha\|_2^2$ such that $\alpha^T \mathbf{K} \alpha \leq R^2$. In Exercise 13.2, we show how the spline estimator in Example 13.2 can be understood in the context of kernel ridge regression. 

Example 13.4 (Convex regression). Now suppose that $f^* : \mathcal{C} \rightarrow \mathbb{R}$ is known to be a convex function over its domain \mathcal{C} , some convex and open subset of \mathbb{R}^d . In this case, it is natural to consider the least-squares estimator with a convexity constraint—namely

$$\hat{f} \in \arg \min_{\substack{f: \mathcal{C} \rightarrow \mathbb{R} \\ f \text{ is convex}}} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 \right\}.$$

As stated, this optimization problem is infinite-dimensional in nature. Fortunately, by exploiting the structure of convex functions, it can be converted to an equivalent finite-dimensional problem. In particular, any convex function is sub-differentiable at each point in (the interior of) its domain. Applying this fact to each of the sampled points x_i , $i = 1, 2, \dots, n$, there must exist a sub-gradient vector $\tilde{z}_i \in \mathbb{R}^d$ such that

$$f(x) \geq f(x_i) + \langle \tilde{z}_i, x - x_i \rangle \quad \text{for all } x \in \mathcal{C}. \quad (13.11)$$

Since the cost function depends only on the values $\tilde{y}_i := f(x_i)$, the optimum does not depend on the function behavior elsewhere. Consequently, we claim that it suffices to consider the set $\{(\tilde{y}_i, \tilde{z}_i)\}_{i=1}^n$ of function value and subgradient pairs, and solve the optimization problem

$$\begin{aligned} \min_{\{(\tilde{y}_i, \tilde{z}_i)\}_{i=1}^n} \quad & \frac{1}{2n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 \\ \text{such that} \quad & \tilde{y}_j \geq \tilde{y}_i + \langle \tilde{z}_i, x_j - x_i \rangle \quad \text{for all } i, j = 1, 2, \dots, n. \end{aligned} \quad (13.12)$$

Note that this is a convex program in $N = n(d+1)$ variables, with a quadratic cost function and a total of $2\binom{n}{2}$ linear constraints. 21

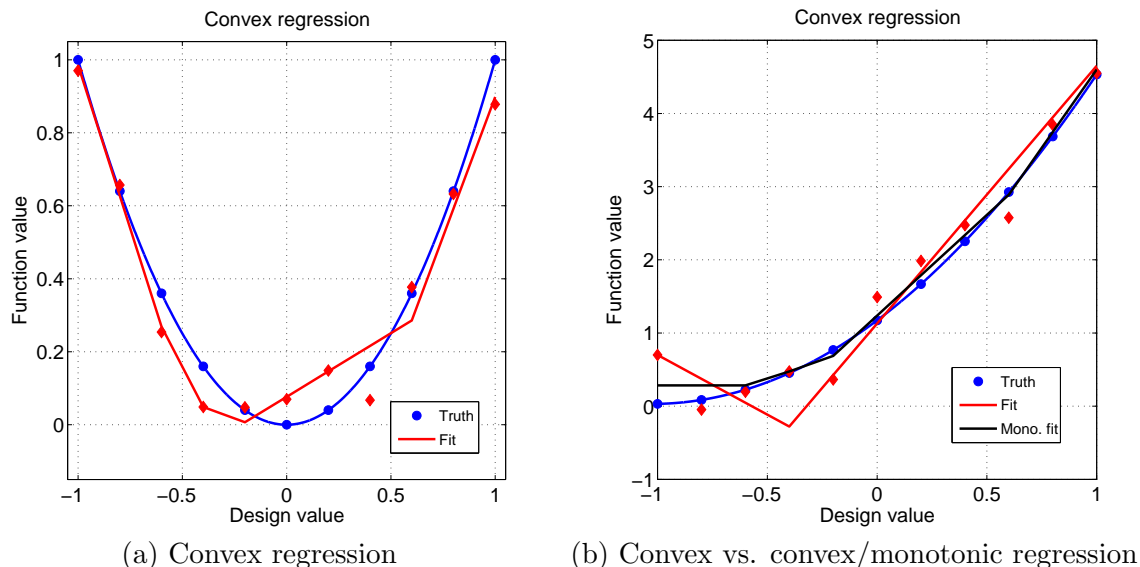


Figure 13-1. (a) Illustration of the convex regression estimate (13.13) based on a fixed design with $n = 11$ equi-distant samples over the interval $\mathcal{C} = [-1, 1]$. (b) Ordinary convex regression compared with convex and monotonic regression estimate.

An optimal solution $\{(\hat{y}_i, \hat{z}_i)\}_{i=1}^n$ can be used to define the estimate $\hat{f} : \mathcal{C} \rightarrow \mathbb{R}$ via

$$\hat{f}(x) := \max_{i=1, \dots, n} \{\hat{y}_i + \langle \hat{z}_i, x - x_i \rangle\}. \quad (13.13)$$

- 2 As the maximum of a collection of linear functions, the function \hat{f} is convex. Moreover, a
 3 short calculation—using the fact that $\{(\hat{y}_i, \hat{z}_i)\}_{i=1}^n$ are feasible for the program (13.12)—
 4 shows that $\hat{f}(x_i) = \hat{y}_i$ for all $i = 1, 2, \dots, n$. Figure 13-1(a) provides an illustration of
 the convex regression estimate (13.13), showing its piecewise linear nature.

There are various extensions to the basic convex regression estimate. For instance,
 in the one-dimensional setting ($d = 1$), it might be known *a priori* that f is a non-
 decreasing function, so that its derivative (or more generally, sub-gradients) are non-
 negative. In this case, it is natural to impose additional non-negativity constraints
 ($\tilde{z}_j \geq 0$) on the sub-gradients in the estimator (13.12). Panel (b) of Figure 13-1 com-
 pares the standard convex regression estimate with the estimator that imposes these
 additional monotonicity constraints. ♣

■ 13.2 Statistical error bounds

From a statistical perspective, the most important question associated with the non-
 parametric least squares estimate (13.7) is how well it approximates the true regression

function f^* . In this section, we develop some techniques to bound the error $\|\hat{f} - f^*\|_n$, as measured in the $L^2(\mathbb{P}_n)$ -norm. We refer the interested reader to Chapter 14 for results that allow such bounds to be translated into bounds in the $L^2(\mathbb{P})$ -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.

Our first main result is defined in terms of 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^* . In particular, let us define the f^* -shifted version of the function class \mathcal{F} , namely

$$\mathcal{F}^* := \{f - f^* \mid f \in \mathcal{F}\}. \quad (13.14)$$

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

$$\mathcal{G}_n(\delta; \mathcal{F}^*) := \mathbb{E}_w \left[\sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \right], \quad (13.15)$$

where the variables $\{w_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0, 1)$ variates. Note that this complexity measure is a deterministic quantity, since we are considering the case of fixed design.

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

$$\frac{\mathcal{G}_n(\delta; \mathcal{F}^*)}{\delta} \leq \frac{\delta}{2\sigma}. \quad (13.16)$$

As we verify in Lemma 13.1, whenever the shifted function class \mathcal{F}^* is star-shaped,² the left-hand side is a non-increasing function of δ , which ensures that the inequality can be satisfied. We use $\delta_n > 0$ to denote the *critical radius* for which inequality (13.16) holds with equality.

Figure 13-2 illustrates the non-increasing property of $\mathcal{G}_n(\delta)/\delta$ for two different function classes: a first-order Sobolev space in panel (a), and a Gaussian kernel space in panel (b). Both of these function classes are convex, so that the star-shaped property holds for any f^* . Setting $\sigma = 1/2$ for concreteness, the critical radius is determined by finding where this non-increasing function crosses the line with slope 1, as illustrated.

As will be clarified later, the Gaussian kernel class is much smaller than the first-order Sobolev space, so that its critical radius is also much smaller. This ordering reflects the

²A function class \mathcal{H} is star-shaped if for any $h \in \mathcal{H}$ and $\alpha \in [0, 1]$, the rescaled function αh also belongs to \mathcal{H} .

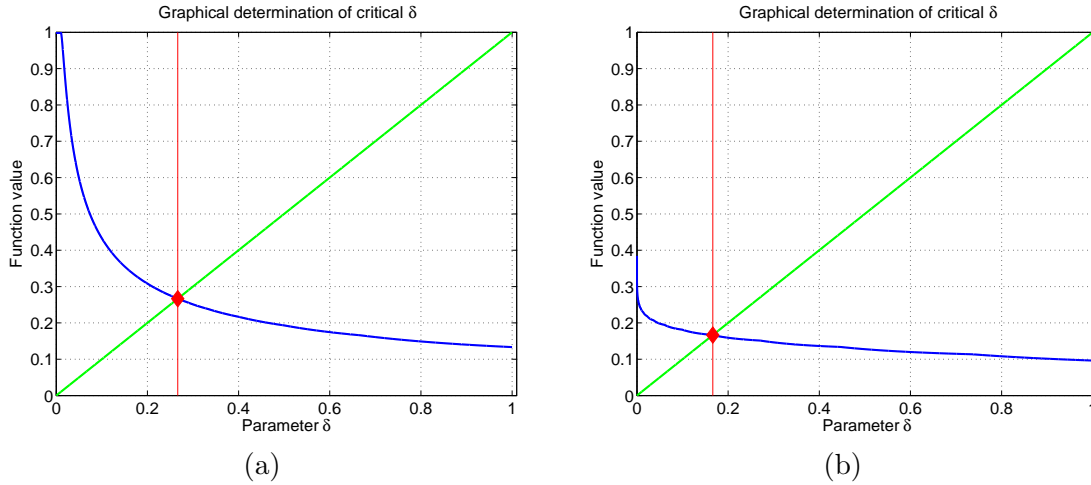


Figure 13-2. Plots of the function $h(\delta) = \frac{\mathcal{G}_n(\delta; \mathcal{F})}{\delta}$ for sample size $n = 100$ and two different function classes. (a) A first-order Sobolev space. (b) A Gaussian kernel class. In both cases, the function h (plotted in blue) is non-increasing, as guaranteed by Lemma 13.1. The critical radius δ_n , marked by a red diamond, is determined by finding its intersection with the line of slope $1/(2\sigma)$, with $\sigma = 1$ in this case (plotted as a green line).

natural intuition that it should be easier to perform regression over a smaller function class. See the discussion following Theorem 13.1 for more details on the star-shaped property and the critical radius δ_n .

Some intuition: Why should the inequality (13.16) be relevant to the analysis of the non-parametric least squares estimator? A little calculation is helpful in gaining intuition. Since \hat{f} and f^* are optimal and feasible, respectively, for the constrained least-squares problem (13.7), we are guaranteed that

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \leq \frac{1}{2n} \sum_{i=1}^n (y_i - f^*(x_i))^2.$$

Recalling that $y_i = f^*(x_i) + \sigma w_i$, some simple algebra leads to the equivalent expression

$$\frac{1}{2} \|\hat{f} - f^*\|_n^2 \leq \frac{\sigma}{n} \sum_{i=1}^n w_i (\hat{f}(x_i) - f^*(x_i)), \quad (13.17)$$

which we call the *basic inequality for non-parametric least squares*.

Now by definition, the difference function $\hat{f} - f^*$ belongs to \mathcal{F}^* , so that we can bound the right-hand side by taking the supremum over all functions $g \in \mathcal{F}^*$ with $\|g\|_n \leq \|\hat{f} - f^*\|_n$. Reasoning heuristically, this observation suggests that the squared

error $\delta^2 := \mathbb{E}[\|\hat{f} - f^*\|_n^2]$ should satisfy a bound of the form

$$\frac{\delta^2}{2} \leq \sigma \mathcal{G}_n(\delta; \mathcal{F}^*), \quad \text{or equivalently} \quad \frac{\delta}{2\sigma} \leq \frac{\mathcal{G}_n(\delta; \mathcal{F}^*)}{\delta}. \quad (13.18)$$

By construction, this inequality can only hold for values of δ that are less than or equal to the smallest positive solution δ_n to the critical inequality (13.16). In summary, this heuristic argument suggests a bound of the form $\mathbb{E}[\|\hat{f} - f^*\|_n^2] \leq \delta_n^2$.

To be clear, the step from the basic inequality (13.17) to the bound (13.18) is *not* rigorously justified for various reasons, but the underlying intuition is correct. Let us now state a rigorous result, one that applies to the least-squares estimator (13.7) based on observations from standard Gaussian noise model $y_i = f^*(x_i) + v_i$, where $v_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. noise variables. Recall the definition (13.16) of the critical radius δ_n .

Theorem 13.1. Suppose that the shifted function class \mathcal{F}^* is star-shaped. Then there are universal positive constants (c_0, c_1, c_2) such that for any $t \geq \delta_n$, the non-parametric least squares estimate \hat{f}_n satisfies

$$\mathbb{P}[\|\hat{f}_n - f^*\|_n^2 \geq c_0 t \delta_n] \leq c_1 e^{-\frac{c_2 n t \delta_n}{\sigma^2}}. \quad (13.19)$$

Remarks: Note that bound (13.19) provides non-asymptotic control on the regression error $\|\hat{f} - f^*\|_2^2$. It also implies that the mean-squared error in the $L^2(\mathbb{P}_n)$ -seminorm is upper bounded as

$$\mathbb{E}[\|\hat{f}_n - f^*\|_n^2] \leq c_3 \left\{ \delta_n^2 + \frac{\sigma^2}{n} \right\} \quad \text{for some universal constant } c_3.$$

As shown in Exercise 13.4, for any function class \mathcal{F} that contains the constant function $f \equiv 1$, we necessarily have $\delta_n^2 \geq \frac{2}{\pi} \frac{\sigma^2}{n}$, so that the δ_n^2 term dominates for any fixed σ .

For concreteness, we have stated the result for the case of noise variables $v_i \sim \mathcal{N}(0, \sigma^2)$. However, as the proof will clarify, all that is required is an upper tail bound on the random variable

$$Z_n(\delta) := \sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n v_i g(x_i) \right|$$

in terms of its expectation. The expectation $\mathbb{E}[Z_n(\delta)]$ defines a more general form of (potentially non-Gaussian) noise complexity that then determines the critical radius.

The star-shaped condition on the shifted function class $\mathcal{F}^* = \mathcal{F} - f^*$ is needed in

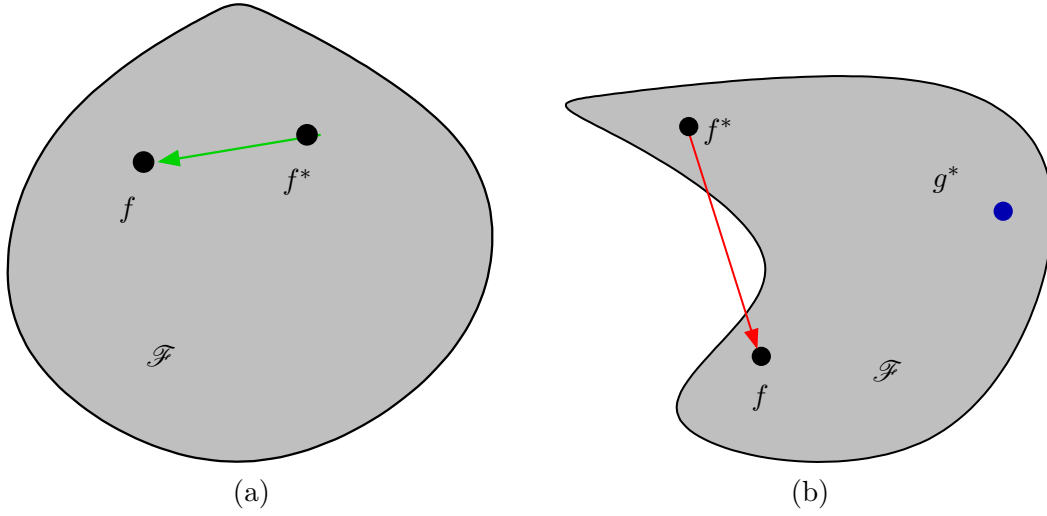


Figure 13-3. Illustration of star-shaped properties of sets. (a) The set \mathcal{F} is convex, and hence is star-shaped around any of its points. The line between f^* and f is contained within \mathcal{F} , and the same is true for any line joining any pair of points in \mathcal{F} . (b) A set \mathcal{F} that is not star-shaped around all its points. It fails to be star-shaped around the point f^* , since the line drawn to $f \in \mathcal{F}$ does not lie within the set. However, this set is star-shaped around the point g^* .

order to ensure the existence of the critical radius δ_n . In explicit terms, the function class \mathcal{F}^* is star-shaped if for any $g \in \mathcal{F}$ and $\alpha \in [0, 1]$, the function αg also belongs to \mathcal{F}^* . Equivalently, we say that \mathcal{F} is star-shaped around f^* . For instance, if \mathcal{F} is convex, then as illustrated in Figure 13-3, then it is necessarily star-shaped around any $f^* \in \mathcal{F}$. See Exercise 13.3 for exploration of these relations. If \mathcal{F} is not convex, then there must exist choices $f^* \in \mathcal{F}$ such that \mathcal{F}^* is not star-shaped. However, given a non-convex function class, it is still possible that the \mathcal{F}^* is star-shaped for *some* choices of f^* . See Figure 13-3(b) for an illustration for an illustration of these two possibilities.

If the star-shaped condition fails to hold, then Theorem 13.1 can instead be applied with δ_n defined in terms of the star-hull

$$\text{star}(\mathcal{F}^*; 0) := \{\alpha g \mid g \in \mathcal{F}^*, \alpha \in [0, 1]\} = \{\alpha(f - f^*) \mid f \in \mathcal{F}, \alpha \in [0, 1]\}. \quad (13.20)$$

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

$$\partial\mathcal{F} := \mathcal{F} - \mathcal{F} = \{f_1 - f_2 \mid f_1, f_2 \in \mathcal{F}\}, \quad (13.21)$$

or its star-hull when necessary. We illustrate these considerations in the concrete examples to follow.

Let us now verify that the star-shaped condition ensures existence of the critical radius:

Lemma 13.1. 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

$$\frac{\mathcal{G}_n(\delta; \mathcal{H})}{\delta} \leq c\delta \quad (13.22)$$

has a smallest positive solution.

Proof. So as to ease notation, we drop the dependence of \mathcal{G}_n on the function class \mathcal{H} throughout this proof. Given a pair $0 < \delta \leq t$, it suffices to show that $\frac{\delta}{t}\mathcal{G}_n(t) \leq \mathcal{G}_n(\delta)$. Given any function $h \in \mathcal{H}$ with $\|h\|_n \leq t$, we may define the rescaled function $\tilde{h} = \frac{\delta}{t}h$, and write

$$\frac{\delta}{t} \frac{1}{n} \sum_{i=1}^n w_i h(x_i) = \frac{1}{n} \sum_{i=1}^n w_i \tilde{h}(x_i).$$

By construction, we have $\|\tilde{h}\|_n \leq \delta$; moreover, since $\delta \leq t$, the star-shaped assumption guarantees that $\tilde{h} \in \mathcal{H}$. Consequently, in expectation, the right-hand side is at most $\mathcal{G}_n(\delta)$ for any \tilde{h} formed in this way. 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. \square

In practice, determining the exact value of the critical radius δ_n may be difficult, so that we seek reasonable upper bounds on it. As shown in Exercise 13.4, we always have $\delta_n \leq \sigma$, but this is a very crude result. By bounding the local Gaussian complexity, we will obtain much finer results, as illustrated in the examples to follow.

■ 13.2.1 Bounds via metric entropy

Note that the localized Gaussian complexity corresponds to expected absolute maximum of a Gaussian process. As discussed in Chapter 5, Dudley's entropy integral can be used to upper bound such quantities.

In order to do so, let us begin by introducing some convenient notation. For any function class \mathcal{H} , we define $\mathbb{B}_n(\delta; \mathcal{H}) := \{h \in \text{star}(\mathcal{H}) \mid \|h\|_n \leq \delta\}$, and we let $N_n(t; \mathbb{B}_n(\delta; \mathcal{H}))$ denote the t -covering number of $\mathbb{B}_n(\delta; \mathcal{H})$ in the norm $\|\cdot\|_n$. With this notation, we have the following corollary:

Corollary 13.1. Under the conditions of Theorem 13.1, the critical radius δ_n is upper bounded by any $\delta \in (0, \sigma]$ such that

$$\frac{32}{\sqrt{n}} \int_{\frac{\delta^2}{2\sigma}}^{\delta} \sqrt{\log N_n(t; \mathbb{B}_n(\delta; \mathcal{F}^*))} dt \leq \frac{\delta^2}{\sigma}. \quad (13.23)$$

12

Proof. For any $\delta \in (0, \sigma]$, we have $\frac{\delta^2}{2\sigma} \leq \delta$. Accordingly, let us construct a $\frac{\delta^2}{2\sigma}$ -covering of the set $\mathbb{B}_n(\delta; \mathcal{F}^*)$ in the $L^2(\mathbb{P}_n)$ -norm, say a $\{g^1, \dots, g^M\}$. For any $g \in \mathbb{B}_n(\delta; \mathcal{F}^*)$, there is an index $j \in [M]$ such that $\|g^j - g\|_n \leq \frac{\delta^2}{2\sigma}$. Consequently, we have

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| &\stackrel{(i)}{\leq} \frac{1}{n} \sum_{i=1}^n |w_i g^j(x_i)| + \frac{1}{n} \sum_{i=1}^n |w_i (g(x_i) - g^j(x_i))| \\ &\stackrel{(ii)}{\leq} \max_{j=1, \dots, M} \left| \frac{1}{n} \sum_{i=1}^n w_i g^j(x_i) \right| + \sqrt{\frac{\sum_{i=1}^n w_i^2}{n}} \sqrt{\frac{\sum_{i=1}^n (g(x_i) - g^j(x_i))^2}{n}} \\ &\stackrel{(iii)}{\leq} \max_{j=1, \dots, M} \left| \frac{1}{n} \sum_{i=1}^n w_i g^j(x_i) \right| + \sqrt{\frac{\sum_{i=1}^n w_i^2}{n}} \frac{\delta^2}{2\sigma}, \end{aligned}$$

where step (i) follows from the triangle inequality; step (ii) follows from the Cauchy-Schwarz inequality; and step (iii) uses the covering property. 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

$$\mathcal{G}_n(\delta) \leq \mathbb{E}_w \left[\max_{j=1, \dots, M} \left| \frac{1}{n} \sum_{i=1}^n w_i g^j(x_i) \right| \right] + \frac{\delta^2}{2\sigma}, \quad (13.24)$$

3 where we have used the fact that $\mathbb{E}_w \sqrt{\frac{\sum_{i=1}^n w_i^2}{n}} \leq 1$.

It remains to upper bound the expected finite maximum over the M functions in the cover, and we do by using the chaining method from Chapter 5. Define the family of Gaussian random variables $Z_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i g^j(x_i)$ for $j = 1, \dots, M$. Some calculation shows that they are zero-mean, and their associated semi-metric is given by

$$\rho_Z^2(g^j, g^k) := \text{var}(Z_f - Z_g) = \|g^j - g^k\|_n^2.$$

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

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

4 Combined with our earlier bound (13.24), this establishes the claim.

5

□

6 Some examples are helpful to illustrate the uses of Theorem 13.1 and Corollary 13.1.

7 ■ 13.2.2 Bounds for high-dimensional parametric problems

We begin with some bounds for parametric problems, allowing for a general dimension. 1

Example 13.5 (Bound for linear regression). As a warm-up, consider the standard linear regression model $y_i = \langle \theta, x_i \rangle + w_i$ where $\theta \in \mathbb{R}^d$. Although it is a parametric model, some insight can be gained by analyzing it using our general theory. The usual least-squares estimate corresponds to optimizing over the function class

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

In this special case, the shifted function class $\mathcal{F}_{\text{lin}}^*$ is equal to \mathcal{F}_{lin} for any choice of f^* . Note that \mathcal{F}_{lin} is convex and hence, from the result of Exercise 13.3, it is also star-shaped so that Corollary 13.1 can be applied. 2
3
4

As usual for linear regression models, we let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the design matrix with x_i^T as its i^{th} row. Note that the mapping $\theta \mapsto \|f_\theta\|_n = \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}}$ defines a norm on $\text{range}(\mathbf{X})$, and the set $\mathbb{B}_n(\delta; \mathcal{F}_{\text{lin}})$ is isomorphic to a δ -ball in $\text{range}(\mathbf{X})$. Since this range space has dimension given by $\text{rank}(\mathbf{X})$, by a volume ratio argument (see Example 5.4), we have

$$\log N_n(t; \mathbb{B}_n(\delta; \mathcal{F}_{\text{lin}})) \leq r \log \left(1 + \frac{4\delta}{t}\right), \quad \text{where } r = \text{rank}(\mathbf{X}).$$

Using this upper bound in Corollary 13.1, we find that

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

where we have made the change of variables $u = t/\delta$ in step (i), and the final inequality (ii) follows by bounding the integral. Putting together the pieces, Corollary 13.1 implies that

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 = \frac{\|\mathbf{X}(\hat{\theta} - \theta^*)\|_2^2}{n} \lesssim \sigma^2 \frac{\text{rank}(\mathbf{X})}{n}$$

with high probability. This bound is minimax-optimal up to constant factors, as we will show in Chapter 15.



Let us now consider another high-dimensional parametric problem, namely that of sparse linear regression.

Example 13.6 (Bounds for linear regression over ℓ_q -“balls”). Consider the case of sparse linear regression, where the d -variate regression vector θ is assumed to lie within the ℓ_q -ball of radius R_q —namely, the set

$$\mathbb{B}_q(R_q) := \left\{ \theta \in \mathbb{R}^d \mid \sum_{j=1}^d |\theta_j|^q \leq R_q \right\}. \quad (13.25)$$

See Figure 7-1 for an illustration of these sets for different choices of $q \in (0, 1]$. Consider class of linear functions $f_\theta(x) = \langle \theta, x \rangle$ given by

$$\mathcal{F}_q(R_q) := \{f_\theta \mid \theta \in \mathbb{B}_q(R_q)\} \quad (13.26)$$

We adopt the shorthand \mathcal{F}_q when the radius R_q is clear from context.

In this example, we focus on the range $q \in (0, 1)$. Suppose that we solve the least-squares problem with ℓ_q regularization—that is, we compute the estimate

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{B}_q(R_q)} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \langle x_i, \theta \rangle)^2 \right\}. \quad (13.27)$$

For $q \in (0, 1)$, the function class $\mathcal{F}_q(R_q)$ is not convex, so that there exist $\theta^* \in \mathbb{B}_q(R_q)$ such that the shifted class $\mathcal{F}_q^* = \mathcal{F}_q - f_{\theta^*}$ is not star-shaped. Let us instead focus on bounding the metric entropy for $\mathcal{F}_q - \mathcal{F}_q = 2\mathcal{F}_q$. For all $q \in (0, 1)$ and numbers $a, b \in \mathbb{R}$, we have $|a + b|^q \leq |a|^q + |b|^q$, which implies that $2\mathcal{F}_q(R_q)$ is contained with $\mathcal{F}_q(2R_q)$. See the bibliographic section for details.

It is known that for $q \in (0, 1)$, and under mild conditions on the choice of t relative to (n, d, R_q) , the metric entropy of the ℓ_q -ball with respect to ℓ_2 -norm is upper bounded by

$$\log N_{2,q}(t) \leq C_q \left[R_q^{\frac{2}{2-q}} \left(\frac{1}{t} \right)^{\frac{2q}{2-q}} \log d \right], \quad (13.28)$$

where C_q is a constant depending only on q .

Given our design vectors $\{x_i\}_{i=1}^n$, consider the $n \times d$ design matrix \mathbf{X} with x_i^T as its i^{th} row, and let $X_j \in \mathbb{R}^n$ denote its j^{th} column. Our objective is to bound the metric

entropy of the set of all vectors of the form

$$\frac{\mathbf{X}\theta}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{j=1}^d X_j \theta_j \quad (13.29)$$

- 1 as θ ranges over $\mathbb{B}_q(R_q)$, an object known as the *q-convex hull* of the renormalized
 2 column vectors $\{X_1, \dots, X_d\}/\sqrt{n}$. As long as there is a numerical constant C such that
 3 $\max_{j=1, \dots, d} \|X_j\|_2/\sqrt{n} \leq C$, then it is known that the metric entropy of this *q-convex hull*
 4 has the same scaling as the original ℓ_q -ball. See the bibliographic section for further
 5 discussion of these facts about metric entropy.

Exploiting this fact and our earlier bound (13.28) on the metric entropy of the ℓ_q -ball, we find that

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{2\sigma}}^{\delta} \sqrt{\log N_n(t; \mathbb{B}_n(\delta; \mathcal{F}_q(2R_q))} dt &\lesssim R_q^{\frac{1}{2-q}} \sqrt{\frac{\log d}{n}} \int_0^{\delta} \left(\frac{1}{t}\right)^{\frac{q}{2-q}} dt \\ &\lesssim R_q^{\frac{1}{2-q}} \sqrt{\frac{\log d}{n}} \delta^{1-\frac{q}{2-q}}, \end{aligned}$$

a calculation valid for all $q \in (0, 1)$. Corollary 13.1 now implies that the critical condition (13.16) is satisfied as long as

$$R_q^{\frac{1}{2-q}} \sqrt{\frac{\sigma^2 \log d}{n}} \lesssim \delta^{1+\frac{q}{2-q}}, \quad \text{or equivalently} \quad R_q \left(\frac{\sigma^2 \log d}{n}\right)^{1-\frac{q}{2}} \lesssim \delta^2.$$

Theorem 13.1 then implies that

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 = \frac{\|\mathbf{X}(\hat{\theta} - \theta^*)\|_2^2}{n} \lesssim R_q \left(\frac{\sigma^2 \log d}{n}\right)^{1-\frac{q}{2}},$$

- 6 with high probability. In fact, although this result is a corollary of our general theorem,
 7 this rate is minimax-optimal up to constant factors, meaning that no estimator can
 achieve a faster rate. See the bibliographic section for further discussion and references

♣

■ 13.2.3 Bounds for non-parametric problems

Let us now illustrate the use of our techniques for some non-parametric problems.

Example 13.7 (Bounds for Lipschitz functions). Consider the class of functions

$$\mathcal{F}_{\text{Lip}}(L) := \{f : [0, 1] \rightarrow \mathbb{R} \mid f(0) = 0, \quad f \text{ is } L\text{-Lipschitz}\}. \quad (13.30)$$

Recall that f is L -Lipschitz means that $|f(x) - f(x')| \leq L|x - x'|$ for all $x, x' \in [0, 1]$.
 Let us analyze the statistical error associated with non-parametric least squares over
 this function class.

Noting the inclusions

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

it suffices to upper bound the metric entropy of $\mathcal{F}_{\text{Lip}}(2L)$. Based on our discussion from Example 5.6, the metric entropy of this class in the supremum norm scales as $\log N_\infty(t; \mathcal{F}_{\text{Lip}}(2L)) \simeq (L/\delta)$. Consequently, we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N_n(t; \mathbb{B}_n(\delta; \mathcal{F}_{\text{Lip}}(2L)))} dt &\leq \int_0^\delta \sqrt{\log N_\infty(t; \mathcal{F}_{\text{Lip}}(2L))} dt \leq \frac{c}{\sqrt{n}} \int_0^\delta (L/t)^{\frac{1}{2}} dt \\ &= \frac{c'}{\sqrt{n}} \sqrt{L\delta}, \end{aligned}$$

where the constants absorb all terms not dependent on the triplet (δ, L, n) . 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 $\delta_n^2 \simeq \left(\frac{L\sigma^2}{n}\right)^{\frac{2}{3}}$. Putting together pieces, Corollary 13.1 implies that the error in the non-parametric least squares estimate satisfies the bound

$$\|\hat{f} - f^*\|_n^2 \lesssim \left(\frac{L\sigma^2}{n}\right)^{2/3} \quad (13.31)$$

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

♣ 9

Example 13.8 (Bounds for convex regression). As a continuation of the previous example, let us consider the class of *convex* 1-Lipschitz functions, namely

$$\mathcal{F}_{\text{conv}}([0, 1]; 1) := \{f : [0, 1] \rightarrow \mathbb{R} \mid f(0) = 0 \text{ and } f \text{ is convex and 1-Lipschitz}\}.$$

As discussed in Example 13.4, computation of the non-parametric least squares estimate over such convex classes can be reduced to a type of quadratic program. Here we consider the statistical rates that are achievable by such an estimator.

It is known that the metric entropy of $\mathcal{F}_{\text{conv}}$, when measured in the infinity norm, satisfies the upper bound

$$\log N(\epsilon; \mathcal{F}_{\text{conv}}, \|\cdot\|_\infty) \lesssim \left(\frac{1}{\epsilon}\right)^{1/2} \quad (13.32)$$

for all $\epsilon > 0$ sufficiently small. (See the bibliographic section for details.) Thus, we can again use an entropy integral approach to derive upper bounds on the statistical error. In particular, a similar calculation to the previous example shows that the conditions

of Corollary 13.1 are satisfied for $\delta_n^2 \simeq \left(\frac{\sigma^2}{n}\right)^{\frac{4}{5}}$, and so we are guaranteed that

$$\|\hat{f} - f^*\|_n^2 \lesssim \left(\frac{\sigma^2}{n}\right)^{4/5} \quad (13.33)$$

with probability at least $1 - c_1 e^{-c_2 \left(\frac{n}{\sigma^2}\right)^{1/5}}$.

13

1 Note that our error bound (13.33) for convex Lipschitz functions is substantially
 2 faster than our earlier bound (13.31) for Lipschitz functions *without a convexity con-*
 3 *straint*—in particular, the respective rates are $n^{-4/5}$ versus $n^{-2/3}$. In Chapter 15, we
 4 show that both of these rates are minimax optimal, meaning that apart from constant
 5 factors, they cannot be improved substantially. Thus, we see that the additional con-
 6 straint of convexity is significant from a statistical point of view. In fact, as we explore
 7 in Exercise 13.7, in terms of their statistical error, convex Lipschitz functions behave
 8 exactly like the class of all twice differentiable functions with bounded second derivative,
 9 so that the convexity constraint amounts to imposing an extra degree of smoothness.

10 ♣

11 ■ 13.2.4 Proof of Theorem 13.1

12 We now turn to the proof of our previously stated theorem.

13 Establishing a basic inequality

Recall the basic inequality (13.17) established in our earlier discussion. In terms of the shorthand notation $\hat{\Delta} = \hat{f} - f^*$, it can be written as

$$\frac{1}{2} \|\hat{\Delta}\|_n^2 \leq \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i). \quad (13.34)$$

14 Note that the error function $\hat{\Delta} = \hat{f} - f^*$ belongs to the shifted function class \mathcal{F}^* .

15 Controlling the right-hand side

In order to control the stochastic component on the right-hand side, we begin by stating an auxiliary lemma in a somewhat more general form, since it will be useful for subsequent arguments. Let \mathcal{H} be an arbitrary star-shaped function class, and let δ_n satisfy the inequality $\frac{\mathcal{G}_n(\delta; \mathcal{H})}{\delta} \leq \frac{\delta}{2\sigma}$. For a given $u > 0$, define the event

$$\mathcal{A}(u) := \left\{ \exists g \in \mathcal{H} \cap \{\|g\|_n \geq u\} \mid \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i) \right| \geq 2\|g\|_n u \right\}. \quad (13.35)$$

16 The following lemma provides control on this event:

17

Lemma 13.2. For all $u \geq \delta_n$, we have

$$\mathbb{P}[\mathcal{A}(u)] \leq e^{-\frac{nu^2}{2\sigma^2}}. \quad (13.36)$$

Let us prove the main result by exploiting this lemma, in particular with the settings $\mathcal{H} = \mathcal{F}^*$ and $u = \sqrt{t\delta_n}$, so that we have

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

If $\|\hat{\Delta}\|_n < \sqrt{t\delta_n}$, then the claim is immediate. Otherwise, we have $\hat{\Delta} \in \mathcal{F}^*$ and $\|\hat{\Delta}\|_n \geq \sqrt{t\delta_n}$, so that we may condition on $\mathcal{A}^c(\sqrt{t\delta_n})$ so as to obtain the bound

$$\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \hat{\Delta}(x_i) \right| \leq \sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_2 \leq \|\hat{\Delta}\|_n}} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i) \right| \leq 2 \|\hat{\Delta}\|_n \sqrt{t\delta_n}.$$

Consequently, the basic inequality (13.34) implies that $\|\hat{\Delta}\|_n^2 \leq 4\|\hat{\Delta}\|_n \sqrt{t\delta_n}$, or equivalently that $\|\hat{\Delta}\|_n^2 \leq 16t\delta_n$, with probability greater than $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$.

In order to complete the proof of Theorem 13.1, it remains to prove Lemma 13.2.

Proof of Lemma 13.2

Our first step is reduce the problem to controlling a supremum over a subset of functions satisfying the upper bound $\|\tilde{g}\|_n \leq u$. Suppose that there exists some $g \in \mathcal{H}$ with $\|g\|_n \geq u$ such that

$$\left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i) \right| \geq 2\|g\|_n u. \quad (13.37)$$

Defining the function $\tilde{g} := \frac{u}{\|g\|_n} g$, we observe that $\|\tilde{g}\|_n = u$. Since $g \in \mathcal{H}$ and $u \leq \|g\|_n$ by construction, the star-shaped assumption implies that $\tilde{g} \in \mathcal{H}$. Consequently, we have shown that if there exists a function g satisfying the inequality (13.37), which occurs whenever the event $\mathcal{A}(u)$ is true, then there exists a function $\tilde{g} \in \mathcal{H}$ with $\|\tilde{g}\|_n = u$ such that

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

We thus conclude that

$$\mathbb{P}[\mathcal{A}(u)] \leq \mathbb{P}[Z_n(u) \geq 2u^2], \quad \text{where} \quad Z_n(u) := \sup_{\substack{\tilde{g} \in \mathcal{H} \\ \|\tilde{g}\|_n \leq u}} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i \tilde{g}(x_i) \right|. \quad (13.38)$$

Since the noise variables $w_i \sim \mathcal{N}(0, 1)$ are i.i.d., for each fixed \tilde{g} , the variable $\frac{\sigma}{n} \sum_{i=1}^n w_i \tilde{g}(x_i)$ is zero-mean and Gaussian, so that the variable $Z_n(u)$ is the supremum of a Gaussian process. If we view this supremum as a function of a standardized Gaussian vector, then it can be verified that the associated Lipschitz constant is at most $\frac{\sigma u}{\sqrt{n}}$. Consequently, by Theorem 2.4, for any $s > 0$, we have the sub-Gaussian tail bound $\mathbb{P}[Z_n(u) \geq \mathbb{E}[Z_n(u)] + s] \leq e^{-\frac{ns^2}{2u^2\sigma^2}}$, and setting $s = u^2$ yields

$$\mathbb{P}[Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2] \leq e^{-\frac{nu^2}{2\sigma^2}}. \quad (13.39)$$

Finally, by definition of $Z_n(u)$ and $\mathcal{G}_n(u)$, we have $\mathbb{E}[Z_n(u)] = \sigma \mathcal{G}_n(u)$. By Lemma 13.1, the function $v \mapsto \frac{\mathcal{G}_n(v)}{v}$ is non-decreasing, and since $u \geq \delta_n$ by assumption, we have

$$\sigma \frac{\mathcal{G}_n(u)}{u} \leq \sigma \frac{\mathcal{G}_n(\delta_n)}{\delta_n} \stackrel{(i)}{\leq} \delta_n/2 \leq \delta_n,$$

where step (i) uses the critical condition (13.16). Putting together the pieces, we have shown that $\mathbb{E}[Z_n(u)] \leq u\delta_n$. Combined with the tail bound (13.39), we obtain

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

where step (ii) uses the inequality $u^2 \geq u\delta_n$. 7

■ 13.3 Oracle inequalities 8

In our analysis thus far, we have assumed that the unknown function f^* belongs to the function class \mathcal{F} over which the constrained least-squares estimator (13.7) is defined. In practice, this assumption might not be satisfied, but we might nonetheless be interested in obtaining bounds on the performance of the estimator (13.7). In such settings, we expect its performance to involve both the *estimation error* that arises in Theorem 13.1, and some additional form of *approximation error*, arising from the fact that $f^* \notin \mathcal{F}$. 9
10
11
12
13

A natural way in which to measure approximation error is in terms of the best approximation to f^* using functions from \mathcal{F} . In the setting of interest in this chapter, the error in this best approximation is given $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$. Note that this error can only be achieved by an “oracle” that has direct access to the samples $\{f^*(x_i)\}_{i=1}^n$. 1

For this reason, results that involve this form of approximation error are referred to as *oracle inequalities*. With this set-up, we have the following generalization of Theorem 13.1. As before, we assume that we observe samples $\{(y_i, x_i)\}_{i=1}^n$ from the model $y_i = f^*(x_i) + \sigma w_i$ where $w_i \sim \mathcal{N}(0, 1)$. The reader should also recall the shorthand notation $\partial\mathcal{F} = \{f_1 - f_2 \mid f_1, f_2 \in \mathcal{F}\}$.

Theorem 13.2. Let δ_n be the smallest positive solution to the inequality

$$\frac{\mathcal{G}_n(\delta; \partial\mathcal{F})}{\delta} \leq \frac{\delta}{2\sigma}. \quad (13.40a)$$

There are universal positive constants (c_0, c_1, c_2) such that for any $t \geq \delta_n$, the non-parametric least squares estimate \hat{f}_n satisfies the bound

$$\|\hat{f} - f^*\|_n^2 \leq \inf_{\gamma \in (0,1)} \left\{ \frac{1+\gamma}{1-\gamma} \|f - f^*\|_n^2 + \frac{c_0 t \delta_n}{\gamma(1-\gamma)} \right\} \quad \text{for all } f \in \mathcal{F} \quad (13.40b)$$

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

Remarks: Note that the guarantee (13.40b) is actually a family of bounds, one for each $f \in \mathcal{F}$. When $f^* \in \mathcal{F}$, then we can set $f = f^*$, so that the bound (13.40b) reduces to asserting that $\|\hat{f} - f^*\|_n^2 \lesssim t \delta_n$ with high probability, where δ_n satisfies our previous critical inequality (13.16). Thus, up to constant factors, we recover Theorem 13.1 as a special case of Theorem 13.2. In the more general setting when $f^* \notin \mathcal{F}$, setting $t = \delta_n$ and taking the infimum over $f \in \mathcal{F}$ yields an upper bound of the form

$$\|\hat{f} - f^*\|_n^2 \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2. \quad (13.41)$$

This form of the bound clarifies the terminology *oracle inequality*: more precisely, the quantity $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$ is the error achievable only by an oracle that has access to uncorrupted samples of the function f^* . The bound (13.41) guarantees that achieves an error that is a constant multiple of the oracle error, plus a term proportional to δ_n^2 . These two terms have concrete interpretations: the error $\inf_{f \in \mathcal{F}} \|f - f^*\|_n^2$ is a form of *approximation error* that decreases as the function class \mathcal{F} grows, whereas the term δ_n^2 is the *estimation error* that increases as \mathcal{F} becomes more complex. This upper bound can thus be used to choose \mathcal{F} as a function of the sample size so as to obtain a desirable trade-off between the two types of error. We will see specific instantiations of this procedure in the examples to follow.

For future reference, it is also useful to note that by integrating the tail bound (13.40b), we obtain an upper bound of the form

$$\mathbb{E}_w \|\hat{f} - f^*\|_n^2 \lesssim \|f - f^*\|_n^2 + \delta_n^2, \quad (13.42)$$

8 again valid for all choices of $f \in \mathcal{F}$.

■ 13.3.1 Some examples of oracle inequalities

Theorem 13.2 as well as oracle inequality (13.41) are best understood by applying them to derive explicit rates for some particular examples.

Example 13.9 (Orthogonal series expansion). Let $\{\phi_m\}_{m=1}^\infty$ be an orthonormal basis of $L^2(\mathbb{P})$, and for each integer $T = 1, 2, \dots$, consider the function class

$$\mathcal{F}_{\text{ortho}}(1; T) := \left\{ f = \sum_{m=1}^T \beta_m \phi_m \mid \|\beta\|_2 \leq 1 \right\}. \quad (13.43)$$

1 For this function class, computation of the estimate \hat{f} is straightforward: it reduces to
2 a version of linear ridge regression (see Exercise 13.9).

Here we consider the guarantees of Theorem 13.2 as applied to \hat{f} as an estimate of an arbitrary function f^* in the unit ball of $L^2(\mathbb{P})$. Since $\{\phi_m\}_{m=1}^\infty$ is an orthonormal basis of $L^2(\mathbb{P})$, we are guaranteed an expansion of the form $f^* = \sum_{m=1}^\infty \theta_m \phi_m$, where Parseval's theorem ensures³ the equivalence $\|f\|_2^2 = \sum_{m=1}^\infty \theta_m^2$. Moreover, as shown in Exercise 13.9, for each $T = 1, 2, \dots$, we have

$$\inf_{f \in \mathcal{F}_{\text{ortho}}(1; T)} \|f - f^*\|_2^2 = \sum_{m=T+1}^\infty (\theta_m^*)^2,$$

3 and this infimum is achieved by the truncated function $\tilde{f}_T = \sum_{m=1}^T \theta_m^* \phi_m$.

On the other hand, since the estimator over $\mathcal{F}_{\text{ortho}}(1; T)$ corresponds to a form of ridge regression in dimension T , the calculations from Example 13.5 imply that the critical equation (13.40a) is satisfied by $\delta_n^2 \simeq \sigma^2 \frac{T}{n}$. Setting $f = \tilde{f}_T$ in the bound (13.42) and then taking expectations over the design X yields that the least-squares estimate \hat{f} over $\mathcal{F}_{\text{ortho}}(1; T)$ satisfies the bound

$$\mathbb{E}_{X,w} [\|\hat{f} - f^*\|_n^2] \lesssim \sum_{m=T+1}^\infty (\theta_m^*)^2 + \sigma^2 \frac{T}{n}. \quad (13.44)$$

4 This oracle inequality allows us to choose the parameter T , which indexes the number
5 of coefficients used in our basis expansion, so as to balance the approximation and
6 estimation errors.

7 The optimal choice of T will depend on the rate at which the basis coefficients
8 $(\theta_m)_{m=1}^\infty$ decay to zero. For example, suppose that they exhibit a polynomial decay,
9 say $|\theta_m| \leq C m^{-\alpha}$ for some $\alpha > 1/2$. See Example 13.10 for a concrete instance of

³Herein we adopt $\|f\|_2$ as a shorthand for the more cumbersome $\|f\|_{L^2(\mathbb{P})}$.

such polynomial decay using Fourier coefficients and α -times differentiable functions. Figure 13-4(a) shows a plot of the upper bound (13.44) as a function of T , with one curve for each the sample sizes $n \in \{100, 250, 500, 1000\}$. The solid markers within each curve show the point $T^* = T^*(n)$ at which the upper bound is minimized, thereby achieving the optimal trade-off between approximation and estimation errors. Note how this optimum grows with the sample size, since more samples allow us to reliably estimate a larger number of coefficients.

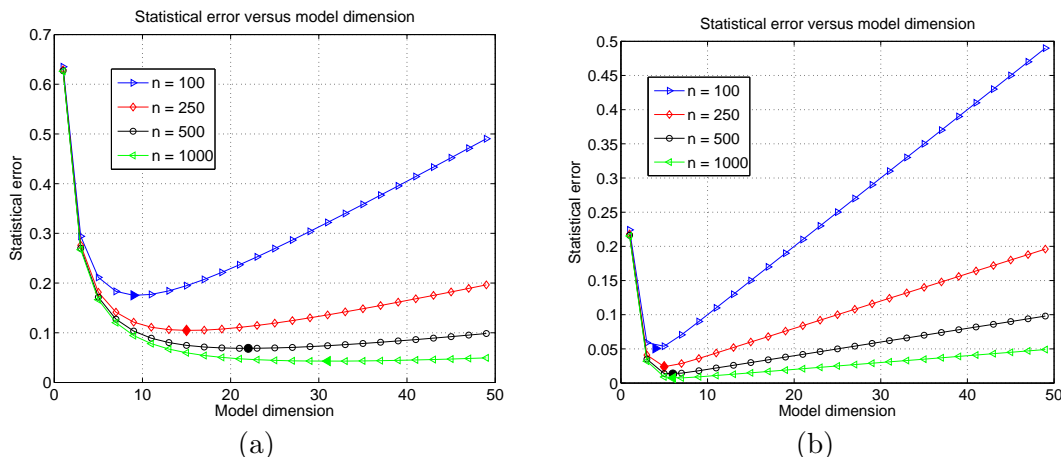


Figure 13-4. Plot of upper bound (13.44) versus the model dimension T , in all cases with noise variance $\sigma^2 = 1$. Each of the four curves corresponds to a different sample size $n \in \{100, 250, 500, 1000\}$. Solid markers within each curve show the optimal choice T^* that minimizes the approximation-estimation trade-off. (a) Polynomial decaying coefficients $|\theta_m| \leq m^{-1}$. (b) Exponential decaying coefficients $|\theta_m| \leq e^{-m/2}$.

As a more concrete instantiation of the previous example, let us consider the approximation of differentiable functions over the space $L^2[0, 1]$.

Example 13.10 (Fourier bases and differentiable functions). Define the constant function $\phi_0(x) = 1$ for all $x \in [0, 1]$, and the sinusoidal functions

$$\phi_m(x) := \sqrt{2} \cos(2m\pi x), \quad \text{and} \quad \tilde{\phi}_m(x) := \sqrt{2} \sin(2m\pi x) \quad \text{for } m = 1, 2, \dots$$

It can be verified that the collection $\{\phi_0\} \cup \{\phi_m\}_{m=1}^\infty \cup \{\tilde{\phi}_m\}_{m=1}^\infty$ forms an orthonormal basis of $L^2[0, 1]$. Consequently, any element $f^* \in L^2[0, 1]$ has the series expansion

$$f^* = \theta_0^* + \sum_{m=1}^{\infty} \{\theta_m^* \phi_m + \tilde{\theta}_m^* \tilde{\phi}_m\}.$$

For each $M = 1, 2, \dots$, define the function class

$$\mathcal{G}(1; M) = \left\{ \beta_0 + \sum_{m=1}^M (\beta_m \phi_m + \tilde{\beta}_m \tilde{\phi}_m) \mid \beta_0^2 + \sum_{m=1}^M (\beta_m^2 + \tilde{\beta}_m^2) \leq 1 \right\}. \quad (13.45)$$

Note that this is simply a re-indexing of a function class $\mathcal{F}_{\text{ortho}}(1; T)$ of the form (13.43) with $T = 2M + 1$. 11
12

Now suppose that for some integer $\alpha \geq 1$, the target function f^* is α -times differentiable with $\int_0^1 (\frac{d^\alpha f^*(t)}{dx^\alpha})^2 dt \leq R$ for some constant R . In this case, it can be verified that there is a constant c such that $(\beta_m^*)^2 + (\tilde{\beta}_m^*)^2 \leq \frac{c}{m^{2\alpha}}$ for all $m \geq 1$, and moreover, we can find a function $f \in \mathcal{G}(1; M)$ such that

$$\|f - f^*\|_2^2 \leq \frac{c'}{M^{2\alpha}}. \quad (13.46)$$


See Exercise 13.10 for details on these properties. 13

Putting together the pieces, the bound (13.44) combined with the approximation-theoretic guarantee (13.46) implies that the least-squares estimate \hat{f}_M over $\mathcal{G}(1; M)$ satisfies the bound

$$\mathbb{E}_{X,w} [\|\hat{f}_M - f^*\|_n^2] \lesssim \frac{c'}{M^{2\alpha}} + \sigma^2 \frac{(2M+1)}{n}.$$

Thus, for a given sample size n and assuming knowledge of the smoothness α and noise variance σ^2 , we can choose $M = M(n, \alpha, \sigma^2)$ so as to balance the approximation and estimation error terms. A little algebra shows that the optimal choice is $M \simeq (n/\sigma^2)^{\frac{1}{2\alpha+1}}$, which leads to the overall rate

$$\mathbb{E}_{X,w} [\|\hat{f}_M - f^*\|_n^2] \lesssim \left(\frac{\sigma^2}{n} \right)^{\frac{2\alpha}{2\alpha+1}}.$$

As will be clarified in Chapter 15, this $n^{-\frac{2\alpha}{2\alpha+1}}$ decay in mean-squared error is the best that can be expected for general univariate α -smooth functions. 14
15 

We now turn to the use of oracle inequalities in high-dimensional sparse linear regression. 16
17

Example 13.11 (Best sparse approximation). Consider the standard linear model $y_i = f_{\theta^*}(x_i) + \sigma w_i$, where $f_{\theta^*}(x) := \langle \theta^*, x \rangle$ is an unknown linear regression function, and $w_i \sim \mathcal{N}(0, 1)$ is an i.i.d. noise sequence. For some sparsity index $s \in \{1, 2, \dots, d\}$, consider the class of all linear regression functions based on s -sparse vectors—namely, the class

$$\mathcal{F}_{\text{spar}}(s) := \{f_\theta \mid \|\theta\|_0 \leq s\},$$

- 1 where $\|\theta\|_0 = \sum_{j=1}^d \mathbb{I}[\theta_j \neq 0]$ counts the number of non-zero coefficients in the vector
 2 $\theta \in \mathbb{R}^d$.

Disregarding computational considerations, a natural estimator is given by

$$\hat{\theta} \in \arg \min_{\theta \in \mathcal{F}_{\text{spar}}(s)} \|y - \mathbf{X}\theta\|_n^2, \quad (13.47)$$

corresponding to performing least-squares over the set of all regression vectors with at most s non-zero coefficients. As corollary of Theorem 13.2, we claim that the $L^2(\mathbb{P}_n)$ -error of this estimator is upper bounded as

$$\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 \lesssim \inf_{\theta \in \mathcal{F}_{\text{spar}}(s)} \|f_{\theta} - f_{\theta^*}\|_n^2 + \underbrace{\sigma^2 \frac{s \log(\frac{ed}{s})}{n}}_{\delta_n^2} \quad (13.48)$$

- 3 with high probability. Consequently, up to constant factors, its error is as good as the
 4 best s -sparse predictor plus the penalty term δ_n^2 . Note that the penalty term grows
 5 linearly with the sparsity s , but only logarithmically in the dimension d , so that it can
 6 be very small even when the dimension is exponentially larger than the sample size n .
 7 In essence, this result guarantees that we pay a relatively small price for not knowing
 8 in advance the best s -sized subset of coefficients to use.

In order to derive this result as a corollary of Theorem 13.2, we need to compute the local Gaussian complexity (13.40a) for our function class. Making note of the inclusion $\partial \mathcal{F}_{\text{spar}}(s) \subset \mathcal{F}_{\text{spar}}(2s)$, we have $\mathcal{G}_n(\delta; \partial \mathcal{F}_{\text{spar}}(s)) \leq \mathcal{G}_n(\delta; \mathcal{F}_{\text{spar}}(2s))$. Now let $S \subset \{1, 2, \dots, d\}$ be an arbitrary $2s$ -sized subset of indices, and let $\mathbf{X}_S \in \mathbb{R}^{n \times 2s}$ denote the sub-matrix with columns indexed by S . We can then write

$$\mathcal{G}_n(\delta; \mathcal{F}_{\text{spar}}(2s)) = \mathbb{E}_w \left[\max_{|S|=2s} Z_n(S) \right], \quad \text{where} \quad Z_n(S) := \sup_{\substack{\theta_S \in \mathbb{R}^{2s} \\ \|\mathbf{X}_S \theta_S\|_2 / \sqrt{n} \leq \delta}} \left| \frac{w^T \mathbf{X}_S \theta_S}{n} \right|.$$

Viewed as a function of the standard Gaussian vector w , the variable $Z_n(S)$ is Lipschitz with constant at most $\frac{\delta}{\sqrt{n}}$, from which Theorem 2.4 implies the tail bound

$$\mathbb{P}[Z_n(S) \geq \mathbb{E}[Z_n(S)] + t\delta] \leq e^{-\frac{nt^2}{2}} \quad \text{for all } t > 0. \quad (13.49)$$

We now upper bound the expectation. Consider the singular value decomposition $\mathbf{X}_S = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times 2s}$ and $\mathbf{V} \in \mathbb{R}^{d \times 2s}$ are matrices of left and right singular vectors, respectively, and $\mathbf{D} \in \mathbb{R}^{2s \times 2s}$ is a diagonal matrix of the singular values. Noting

that $\|\mathbf{X}_S \theta_S\|_2 / \sqrt{n} = \|\mathbf{D} \mathbf{V}^T \theta_S\|_2 / \sqrt{n}$, we arrive at the upper bound

$$\mathbb{E}[Z_n(S)] \leq \mathbb{E}\left[\sup_{\|\beta\|_2 \leq \delta} \left| \frac{1}{\sqrt{n}} \langle \mathbf{U}^T w, \beta \rangle \right|\right],$$

where $\beta \in \mathbb{R}^{2s}$. Since $w \in \mathbb{R}^n$ is standard Gaussian and \mathbf{U} has orthonormal columns, the vector $\mathbf{U}^T w \in \mathbb{R}^{2s}$ has i.i.d. Gaussian entries, and therefore $\mathbb{E}[Z_n(S)] \leq \sqrt{\frac{2s}{n}} \delta$. Combining this upper bound with the earlier tail bound (13.49), an application of the union bound yields

$$\mathbb{P}\left[\max_{|S|=2s} Z_n(S) \geq \sqrt{\frac{2s}{n}} \delta + t\delta\right] \leq \binom{d}{2s} e^{-\frac{nt^2}{2}}, \quad \text{valid for all } t \geq 0.$$

By integrating this tail bound, we find that

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

9 so that the critical inequality (13.16) is satisfied for $\delta_n^2 \simeq \sigma^2 \frac{s \log(ed/s)}{n}$, as claimed. ♣

10 ■ 13.3.2 Proof of Theorem 13.2

We now turn to the proof of our oracle inequality, which is a relatively straightforward extension of the proof of Theorem 13.1. Given an arbitrary $\tilde{f} \in \mathcal{F}$, since it is feasible and \hat{f} is optimal, we have

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \leq \frac{1}{2n} \sum_{i=1}^n (y_i - \tilde{f}(x_i))^2.$$

Using the relation $y_i = f^*(x_i) + \sigma w_i$, some algebra then yields

$$\|\hat{\Delta}\|_n^2 \leq \|\Delta^*\|_n^2 + 2 \left| \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \right| + 2 \underbrace{\left| \frac{\sigma}{n} \sum_{i=1}^n w_i \Delta^*(x_i) \right|}_{Z^*},$$

11 where we have defined $\hat{\Delta} := \hat{f} - f^*$, and $\Delta^* = f^* - \tilde{f}$.

The term denoted Z^* is straightforward to analyze. Since the function Δ^* is a deterministic object, the variable Z^* is zero-mean Gaussian with variance $\frac{\sigma^2}{n} \|\Delta^*\|_n^2$, and hence

$$\mathbb{P}\left[|Z^*| > \sqrt{t\delta_n} \|\Delta^*\|_n\right] \leq e^{-\frac{nt\delta_n}{2\sigma^2}}.$$

On the other hand, the term involving $\hat{\Delta}$ can be analyzed using Lemma 13.2. In particular, applying this lemma with $u = \sqrt{t\delta_n}$ for some $t \geq \delta_n$, we either have $\|\hat{\Delta}\|_n^2 < t\delta_n$, in which case the claim follows, or

$$\mathbb{P}\left[2\left|\frac{\sigma}{n}\sum_{i=1}^n w_i \hat{\Delta}(x_i)\right| \geq 4\sqrt{t\delta_n}\|\hat{\Delta}\|_n\right] \leq e^{-\frac{nt\delta_n}{2\sigma^2}}.$$

Putting together the pieces, we see that the inequality

$$\|\hat{\Delta}\|_n^2 \leq \|\Delta^*\|_n^2 + 4\sqrt{t\delta_n}(\|\hat{\Delta}\|_n + \|\Delta^*\|_n)$$

holds with probability at least $1 - 2e^{-\frac{nt\delta_n}{2\sigma^2}}$. We now apply the inequality $ab \leq \frac{a^2}{\gamma} + \gamma b^2$, valid for any positive triplet of numbers (a, b, γ) , to obtain

$$\|\hat{\Delta}\|_n^2 \leq \|\Delta^*\|_n^2 + \frac{32t\delta_n}{\gamma} + \gamma(\|\hat{\Delta}\|_n^2 + \|\Delta^*\|_n^2).$$

For any $\gamma \in (0, 1)$, some algebra then yields

$$\|\hat{\Delta}\|_n^2 \leq \frac{1+\gamma}{1-\gamma}\|\Delta^*\|_n^2 + \frac{32t\delta_n}{\gamma(1-\gamma)},$$

which completes the proof. 1

■ 13.4 Regularized estimators 2

Up to this point, we have analyzed least-squares estimators based on imposing explicit constraints on the function class. From the computational point of view, it is often more convenient to implement estimators based on explicit penalization or regularization terms. As we will see, these estimators enjoy statistical behavior similar to their constrained analogues. 3
4
5

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

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{F}}^2 \right\}, \quad (13.50)$$

where $\lambda_n \geq 0$ is a regularization weight to be chosen by the statistician. We state a general oracle-type result that does not require f^* to be a member of \mathcal{F} . 3
4

5 ■ 13.4.1 A general statistical error bound

Recall the compact notation $\partial\mathcal{F} = \mathcal{F} - \mathcal{F}$. As in our previous theory, the statistical error involves a local Gaussian complexity over this class, which in this case takes the form

$$\mathcal{G}_n(\delta; \mathbb{B}_{\partial\mathcal{F}}(3)) := \mathbb{E}_w \left[\sup_{\substack{g \in \partial\mathcal{F} \\ \|g\|_{\mathcal{F}} \leq 3, \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n w_i f(x_i) \right| \right]. \quad (13.51)$$

When the function class \mathcal{F} and ball $\mathbb{B}_{\partial\mathcal{F}}(3) = \{g \in \partial\mathcal{F} \mid \|g\|_{\mathcal{F}} \leq 1\}$ are clear from the context, we adopt $\mathcal{G}_n(\delta)$ as a convenient shorthand. For a user-defined radius $R > 0$, we let $\delta_n > 0$ the smallest positive solution to the inequality

$$\frac{\mathcal{G}_n(\delta)}{\delta} \leq \frac{R}{2\sigma} \delta. \quad (13.52)$$

Theorem 13.3. Given the previously described observation model, suppose that we solve the convex program (13.50) with $\lambda_n \geq 2\delta_n^2$. Then there are universal positive constants (c_0, c_1, c_2) such that

$$\|\hat{f} - f^*\|_n^2 \leq c_0 \inf_{\|f\|_{\mathcal{F}} \leq R} \|f - f^*\|_n^2 + c_1 R^2 \{\delta_n^2 + \lambda_n\} \quad (13.53a)$$

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

$$\mathbb{E} \|\hat{f} - f^*\|_n^2 \leq c'_0 \inf_{\|f\|_{\mathcal{F}} \leq R} \|f - f^*\|_n^2 + c'_1 R^2 \{\delta_n^2 + \lambda_n\}. \quad (13.53b)$$

8 ■ 13.4.2 Consequences for kernel ridge regression

9 Recall from Chapter 12 our discussion of the kernel ridge regression estimate (12.27).
 10 There we showed that this KRR estimate has attractive computational properties, in
 11 that it only requires computing the empirical kernel matrix, and then solving a linear
 12 system (see Proposition 12.4). Here we turn to the complementary question of under-
 13 standing its statistical error. Since it is a special case of the general estimator (13.50),
 14 Theorem 13.3 can be used to derive upper bounds on its statistical error. Interestingly,
 15 this bounds have a very intuitive interpretation, one involving the eigenvalues of the
 16 empirical kernel matrix.

17 Recall that the empirical kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ is symmetric and positive semidef-
 18 inite, with entries of the form $K_{ij} = \mathcal{K}(x_i, x_j)/n$. It is thus diagonalizable with non-
 negative eigenvalues, which we take to be ordered as $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \hat{\mu}_n \geq 0$. The following
 corollary of Theorem 13.3 provides bounds on the performance of the kernel ridge re-
 gression estimate in terms of these eigenvalues:

Corollary 13.2. For the KRR estimate (12.27), the bounds of Theorem 13.3 hold for any $\delta_n > 0$ satisfying the inequality

$$\sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}} \leq \frac{R}{\sigma} \delta^2. \quad (13.54)$$


We provide the proof in Section 13.4.3. Before doing so, let us examine the implications of Corollary 13.2 for some specific choices of kernels.

Example 13.12 (Rates for polynomial regression). Given some integer $m \geq 2$, consider the kernel function $\mathcal{K}(x, z) = (1 + xz)^{m-1}$. The associated RKHS corresponds to the space of all polynomials of degree at most $m - 1$, which is a vector space with dimension m . Consequently, the empirical kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ can have rank at most $\min\{n, m\}$, so that for all $n \geq m$, we have

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}} \leq \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^m \min\{\delta^2, \hat{\mu}_j\}} \leq \delta \sqrt{\frac{m}{n}}$$

Consequently, the critical inequality (13.54) is satisfied for all $\delta \geq \frac{\sigma}{R} \sqrt{\frac{m}{n}}$, so that the KRR estimate satisfies the bound

$$\|\hat{f} - f^*\|_n^2 \lesssim \inf_{\|f\|_{\mathcal{H}} \leq R} \|f - f^*\|_n^2 + \sigma^2 \frac{m}{n},$$

both in high probability and in expectation. This bound is intuitively reasonable, since the space of $m - 1$ polynomials has a total of m free parameters, so that we expect that the ratio m/n should converge to zero in order for consistent estimation to be possible. More generally, this same bound with $m = r$ holds for any kernel function that has some finite rank $r \geq 1$. 

We now turn to a kernel function with an infinite number of eigenvalues:

Example 13.13 (First-order Sobolev space). Recall the kernel function defined on the unit square $[0, 1] \times [0, 1]$ via $\mathcal{K}(x, z) = \min\{x, z\}$. As discussed in Example 12.5, the associated RKHS corresponds to a first-order Sobolev space

$$\mathbb{H}^1[0, 1] := \{f : [0, 1] \rightarrow \mathbb{R} \mid f(0) = 0, \text{ and } f \text{ is absolutely continuous with } f' \in L^2[0, 1]\}.$$

As shown in Example 12.11, the kernel operator has the eigendecomposition

$$\phi_j(x) = \sin(x/\sqrt{\mu_j}), \quad \mu_j = \left(\frac{2}{(2j-1)\pi}\right)^2 \quad \text{for } j = 1, 2, \dots,$$

so that the eigenvalues drop off at the rate j^{-2} . As the sample size increases, the eigenvalues of the empirical kernel matrix \mathbf{K} approach those of the population kernel operator. For the purposes of calculation, Figure 13-5(a) suggests the heuristic of assuming that $\hat{\mu}_j \leq \frac{c}{j^2}$ for some universal constant c . Our later analysis in Chapter 14 will provide a rigorous way of making such an argument.⁴

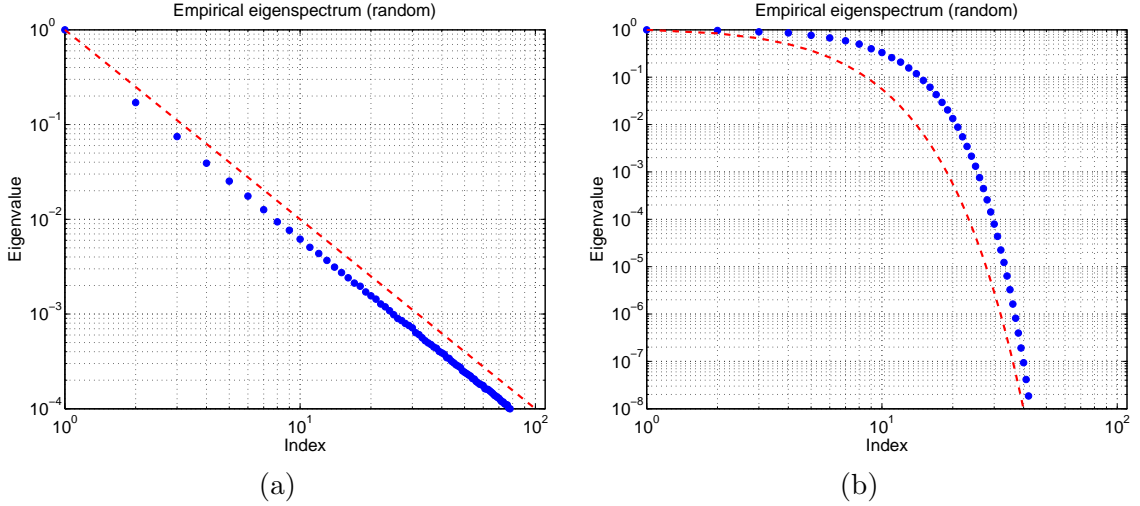


Figure 13-5. Log-log behavior of the eigenspectrum of the empirical kernel matrix based on $n = 2000$ samples drawn i.i.d. from the uniform distribution the interval \mathcal{X} for two different kernel functions. Blue dots correspond to empirical eigenvalues, whereas the red line shows the theoretically predicted drop-off of the population operator. (a) The first-order Sobolev kernel $\mathcal{K}(x, z) = \min\{x, z\}$ on the interval $\mathcal{X} = [0, 1]$. (b) The Gaussian kernel $\mathcal{K}(x, z) = \exp(-\frac{(x-z)^2}{2\sigma^2})$ with $\sigma = 0.5$ on the interval $\mathcal{X} = [-1, 1]$.

5

Under our heuristic assumption, we have

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}} \leq \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, c j^{-2}\}} \leq \frac{1}{\sqrt{n}} \sqrt{k\delta^2 + c \sum_{j=k+1}^n j^{-2}},$$


where k is the smallest positive integer such that $ck^{-2} \leq \delta^2$. Upper bounding the final sum by an integral, we have $c \sum_{j=k+1}^n j^{-2} \leq c \int_{k+1}^{\infty} t^{-2} dt \leq ck^{-1} \leq k\delta^2$, and hence

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}} \leq c' \sqrt{\frac{k}{n}} \delta \leq c'' \sqrt{\frac{\delta}{n}}.$$

⁴In particular, Proposition 14.1 shows that the critical radii computed using the population and empirical kernel eigenvalues are equivalent up to constant factors.

Consequently, the critical inequality (13.54) is satisfied by $\delta_n^{3/2} \simeq \frac{\sigma}{R\sqrt{n}}$, or equivalently $\delta_n^2 \simeq \left(\frac{\sigma^2}{R^2} \frac{1}{n}\right)^{2/3}$. Putting together the pieces, Corollary 13.2 implies that the KRR estimate will satisfy the upper bound

$$\|\hat{f} - f^*\|_n^2 \lesssim \inf_{\|f\|_{\mathbb{H}} \leq R} \|f - f^*\|_n^2 + R^2 \delta_n^2 \simeq \inf_{\|f\|_{\mathbb{H}} \leq R} \|f - f^*\|_n^2 + R^{2/3} \left(\frac{\sigma^2}{n}\right)^{2/3},$$

both with high probability and in expectation. As will be seen later in Chapter 15, this rate is minimax-optimal for the first-order Sobolev space. 

Example 13.14 (Gaussian kernel). Now let us consider the same issues for 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 12.19, the eigenvalues of the associated kernel operator scale as $\mu_j \simeq e^{-c_1 j \log j}$ as $j \rightarrow +\infty$. Accordingly, let us adopt the heuristic that the empirical eigenvalues satisfy a bound of the form $\hat{\mu}_j \leq c_0 e^{-c_1 j \log j}$. Figure 13-5(b) provides empirical justification of this scaling for the Gaussian kernel: notice how the empirical plots on the log-log scale agree qualitatively with the theoretical prediction. Again, Proposition 14.1 in Chapter 14 allows us to make a rigorous argument that reaches the conclusion to be sketched here.


Under our heuristic assumption, for a given $\delta > 0$, we have

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

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 $\delta_n^2 \simeq \frac{\sigma^2}{R^2} \frac{\log(\frac{Rn}{\sigma})}{n}$, so that non-parametric regression over the Gaussian kernel class satisfies the bound

$$\|\hat{f} - f^*\|_n^2 \lesssim \inf_{\|f\|_{\mathbb{H}} \leq R} \|f - f^*\|_n^2 + R^2 \delta_n^2 = \inf_{\|f\|_{\mathbb{H}} \leq R} \|f - f^*\|_n^2 + c \sigma^2 \frac{\log(\frac{Rn}{\sigma})}{n},$$

for some universal constant c . The estimation error component of this upper bound is very fast—within a logarithmic factor of the n^{-1} parametric rate—thereby revealing that the Gaussian kernel class is much smaller than the first-order Sobolev space from Example 13.13. However, the trade-off is that the approximation error decays very slowly as a function of the radius R . See the bibliographic section for further discussion of this important trade-off. 

■ 13.4.3 Proof of Corollary 13.2

The proof of this corollary is based on a bound on the local Gaussian complexity (13.51) of the unit ball of a RKHS. Since it is of independent interest, let us state it as a separate result:

Lemma 13.3. Consider a RKHS with kernel function \mathcal{K} . For a given set of design points $\{x_i\}_{i=1}^n$, let $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \dots \hat{\mu}_n \geq 0$ be the eigenvalues of the normalized kernel matrix \mathbf{K} with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$. Then for all $\delta > 0$, we have

$$\mathbb{E} \left[\sup_{\substack{\|f\|_{\mathbb{H}} \leq 1 \\ \|f\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n w_i f(x_i) \right| \right] \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}}, \quad (13.55)$$

where $w_i \sim \mathcal{N}(0, 1)$ are i.i.d. Gaussian variates.

Proof. It suffices to restrict our attention to functions of the form

$$g(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i). \quad (13.56)$$

some vector of coefficients $\alpha \in \mathbb{R}^n$. Indeed, as argued in our proof of Proposition 12.4, any function f in the Hilbert space can be written in the form $f = g + g_{\perp}$, where g_{\perp} is a function orthogonal to all functions of the form (13.56). Thus, we must have $g_{\perp}(x_i) = \langle g_{\perp}, \mathcal{K}(\cdot, x_i) \rangle_{\mathbb{H}} = 0$, so that neither the objective nor the constraint $\|f\|_n \leq \delta$ have any dependence on g_{\perp} . Lastly, by the Pythagorean theorem, we have $\|f\|_{\mathbb{H}}^2 = \|g\|_{\mathbb{H}}^2 + \|g_{\perp}\|_{\mathbb{H}}^2$, so that we may assume without loss of generality that $g_{\perp} = 0$.

In terms of the coefficient vector $\alpha \in \mathbb{R}^n$ and kernel matrix \mathbf{K} , the constraint $\|g\|_n \leq \delta$ is equivalent to $\|\mathbf{K}\alpha\|_2 \leq \delta$, whereas inequality $\|g\|_{\mathbb{H}}^2 \leq 1$ corresponds to $\|g\|_{\mathbb{H}}^2 = \alpha^T \mathbf{K} \alpha \leq 1$. Thus, we can write the local Gaussian complexity with a linear cost function and quadratic constraints—namely,

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

Since the kernel matrix \mathbf{K} is symmetric and positive semidefinite, it has an eigendecomposition of the form $\mathbf{K} = \mathbf{U}^T \Lambda \mathbf{U}$, where \mathbf{U} is orthogonal and Λ is diagonal with entries $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \dots \geq \hat{\mu}_n > 0$. If we then define the transformed vector $\beta = \mathbf{K}\alpha$, some algebra leads to $\mathcal{G}_n(\delta) = \frac{1}{\sqrt{n}} \mathbb{E}_w [\sup_{\beta \in \mathcal{D}} |w^T \beta|]$, where \mathcal{D} is the intersection of two ellipses given by $\mathcal{D} := \{\beta \in \mathbb{R}^n \mid \|\beta\|_2^2 \leq \delta^2, \sum_{j=1}^n \frac{\beta_j^2}{\hat{\mu}_j} \leq 1\}$. Now let us define the sequence

$\eta_j = \max\{\delta^{-2}, \widehat{\mu}_j^{-1}\}$, and the associated ellipse

$$\mathcal{E} = \{\beta \in \mathbb{R}^n \mid \sum_{j=1}^n \eta_j \beta_j^2 \leq 2\}. \quad (13.57)$$

We claim that $\mathcal{D} \subset \mathcal{E}$; indeed, for any $\beta \in \mathcal{D}$, we have

$$\sum_{j=1}^n \max\{\delta^{-2}, \widehat{\mu}_j^{-1}\} \beta_j^2 \leq \sum_{j=1}^n \frac{\beta_j^2}{\delta^2} + \sum_{j=1}^n \frac{\beta_j^2}{\widehat{\mu}_j} \leq 2.$$

Applying Hölder's inequality with the norm induced by \mathcal{E} and its dual, we find that

$$\mathcal{G}_n(\delta) \leq \frac{1}{\sqrt{n}} \mathbb{E}[\sup_{\beta \in \mathcal{E}} |\langle w, \beta \rangle|] \leq \sqrt{\frac{2}{n}} \mathbb{E} \sqrt{\sum_{j=1}^n \frac{w_j^2}{\eta_j}}.$$

Jensen's inequality allows us to move the expectation inside the square root, so that

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

and substituting $(\eta_j)^{-1} = (\max\{\delta^{-2}, \widehat{\mu}_j^{-1}\})^{-1} = \min\{\delta^2, \widehat{\mu}_j\}$ yields the claim. \square

18

■ 13.4.4 Proof of Theorem 13.3

19

Finally, we turn to the proof of our general theorem on regularized M -estimators. By rescaling the observation model by R , we can analyze an equivalent model with noise variance $(\frac{\sigma}{R})^2$, and with the rescaled approximation error $\inf_{\|f\|_{\mathcal{F}} \leq 1} \|f - f^*\|_n^2$. Our final mean-squared error then should be multiplied by R^2 in order to obtain a result for the original problem.

20

21

22

23

24

To keep the notation stream-lined, we introduce the shorthand $\tilde{\sigma} = \sigma/R$. Let \tilde{f} be any element of \mathcal{F} such that $\|\tilde{f}\|_{\mathcal{F}} \leq 1$. At the end of the proof, we optimize this choice. Since \widehat{f} and \tilde{f} are optimal and feasible (respectively) for the convex program (13.50), we have

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

Defining the errors $\widehat{\Delta} = \widehat{f} - f^*$ and $\widetilde{\Delta} = \tilde{f} - f^*$ and recalling that $y_i = f^*(x_i) + \tilde{\sigma} w_i$,

performing some algebra yields the *modified basic inequality*

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}\|\widetilde{f} - f^*\|_n^2 + \frac{\tilde{\sigma}}{n} \left| \sum_{i=1}^n w_i \widetilde{\Delta}(x_i) \right| + \lambda_n \{ \|\widetilde{f}\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2 \}, \quad (13.58)$$

where $w_i \sim \mathcal{N}(0, 1)$ are i.i.d. Gaussian variables. We split the remainder of the proof into two cases. 25
26

Case 1: First, suppose that $\|\widehat{f}\|_{\mathcal{F}} \leq 2$. In this case, the modified basic inequality (13.58) implies that

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

using the fact that $\|\widetilde{f}\|_{\mathcal{F}} \leq 1$. Since $\|\widehat{f}\|_{\mathcal{F}} \leq 2$, we are guaranteed that $\|\widetilde{\Delta}\|_{\mathcal{F}} \leq 3$. Consequently, by applying Lemma 13.2 over the set of functions $\{g \in \mathcal{F} \mid \|g\|_{\mathcal{F}} \leq 3\}$, we conclude that $\frac{\tilde{\sigma}}{n} \left| \sum_{i=1}^n w_i \widetilde{\Delta}(x_i) \right| \leq c_0 \sqrt{t\delta_n} \|\widetilde{\Delta}\|_n$ with probability at least $1 - e^{-\frac{t^2}{2\tilde{\sigma}^2}}$. By the triangle inequality, we have

$$\begin{aligned} 2\sqrt{t\delta_n} \|\widetilde{\Delta}\|_n &\leq 2\sqrt{t\delta_n} \|\widehat{\Delta}\|_n + 2\sqrt{t\delta_n} \|\widetilde{f} - f^*\|_n \\ &\leq 2\sqrt{t\delta_n} \|\widehat{\Delta}\|_n + 2t\delta_n + \frac{\|\widetilde{f} - f^*\|_n^2}{2}, \end{aligned} \quad (13.59)$$

where the second step uses the Fenchel-Young inequality. Putting together the pieces yields an upper bound of the form

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

so that the claim follows by the quadratic formula, modulo different values of the numerical constants. 27
28

Case 2: Otherwise, we may assume that $\|\widehat{f}\|_{\mathcal{F}} > 2 > 1 \geq \|\widetilde{f}\|_{\mathcal{F}}$. In this case, we have

$$\|\widetilde{f}\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2 = \{\|\widetilde{f}\|_{\mathcal{F}} + \|\widehat{f}\|_{\mathcal{F}}\} \{\|\widetilde{f}\|_{\mathcal{F}} - \|\widehat{f}\|_{\mathcal{F}}\} \leq \{\|\widetilde{f}\|_{\mathcal{F}} - \|\widehat{f}\|_{\mathcal{F}}\}.$$

Writing $\widehat{f} = \widetilde{f} + \widetilde{\Delta}$ and noting that $\|\widehat{f}\|_{\mathcal{F}} \geq \|\widetilde{\Delta}\|_{\mathcal{F}} - \|\widetilde{f}\|_{\mathcal{F}}$ by the triangle inequality, we obtain

$$\lambda_n \{ \|\widetilde{f}\|_{\mathcal{F}}^2 - \|\widehat{f}\|_{\mathcal{F}}^2 \} \leq \lambda_n \{ 2\|\widetilde{f}\|_{\mathcal{F}} - \|\widetilde{\Delta}\|_{\mathcal{F}} \} \leq 2\lambda_n - \lambda_n \|\widetilde{\Delta}\|_{\mathcal{F}}.$$

Substituting this upper bound into our modified basic inequality (13.58) yields the upper bound

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

Our next step to upper bound the stochastic component in the inequality (13.60).

Lemma 13.4. There are universal positive constants (c_1, c_2) such that

$$\left| \frac{\tilde{\sigma}}{n} \sum_{i=1}^n w_i \Delta(x_i) \right| \leq 2\delta_n \|\Delta\|_n + 2\delta_n^2 \|\Delta\|_{\mathcal{F}} + \frac{1}{16} \|\Delta\|_n^2 \quad \text{for all } \|\Delta\|_{\mathcal{F}} \geq 1 \quad (13.61)$$

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

We now complete the proof of the theorem using this lemma. We begin by observing that since $\|\tilde{f}\|_{\mathcal{F}} \leq 1$ and $\|\widehat{f}\|_{\mathcal{F}} > 2$, the triangle inequality implies that $\|\tilde{\Delta}\|_{\mathcal{F}} \geq \|\widehat{f}\|_{\mathcal{F}} - \|\tilde{f}\|_{\mathcal{F}} > 1$, so that Lemma 13.4 may be applied. Substituting the upper bound (13.61) into the inequality (13.60) yields

$$\begin{aligned} \frac{1}{2}\|\widehat{\Delta}\|_n^2 &\leq \frac{1}{2}\|\tilde{f} - f^*\|_n^2 + 2\delta_n \|\tilde{\Delta}\|_n + \{2\delta_n^2 - \lambda_n\} \|\tilde{\Delta}\|_{\mathcal{F}} + 2\lambda_n + \frac{\|\tilde{\Delta}\|_n^2}{16} \\ &\leq \frac{1}{2}\|\tilde{f} - f^*\|_n^2 + 2\delta_n \|\tilde{\Delta}\|_n + 2\lambda_n + \frac{\|\tilde{\Delta}\|_n^2}{16} \end{aligned} \quad (13.62)$$

where the second step uses the fact that $2\delta_n^2 - \lambda_n \leq 0$ by assumption.

Our next step is to convert the terms involving $\tilde{\Delta}$ into quantities involving $\widehat{\Delta}$: in particular, by the triangle inequality, we have $\|\tilde{\Delta}\|_n \leq \|\tilde{f} - f^*\|_n + \|\widehat{\Delta}\|_n$. Thus, we have

$$2\delta_n \|\tilde{\Delta}\|_n \leq 2\delta_n \|\tilde{f} - f^*\|_n + 2\delta_n \|\widehat{\Delta}\|_n, \quad (13.63a)$$

and in addition, combined with the inequality $(a + b)^2 \leq 2a^2 + 2b^2$, we find that

$$\frac{\|\tilde{\Delta}\|_n^2}{16} \leq \frac{1}{8} \left\{ \|\tilde{f} - f^*\|_n^2 + \|\widehat{\Delta}\|_n^2 \right\}. \quad (13.63b)$$

Substituting inequalities (13.63a) and (13.63b) into the earlier bound (13.62) and performing some algebra yields

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

The claim (13.53a) follows by applying the quadratic formula to this inequality. 34

It remains to prove Lemma 13.4. We claim that it suffices to prove the bound (13.61) 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 (13.61) to the new function $g := \Delta/\|\Delta\|_{\mathcal{F}}$, which belongs to $\partial\mathcal{F}$ by the star-shaped assumption. Applying the bound (13.61) to g and then multiplying both sides by $\|\Delta\|_{\mathcal{F}}$, we obtain 35

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

where the second inequality uses the fact that $\|\Delta\|_{\mathcal{F}} > 1$ by assumption. 36

In order to establish the bound (13.61) for functions with $\|g\|_{\mathcal{F}} = 1$, we first consider it over the ball $\{\|g\|_n \leq t\}$, for some fixed radius $t > 0$. Defining the random variable

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

we observe that, viewed as a function of the standard Gaussian vector w , it is Lipschitz with parameter at most $\tilde{\sigma}t/\sqrt{n}$. Consequently, Theorem 2.4 implies that

$$\mathbb{P}[Z_n(t) \geq \mathbb{E}[Z_n(t)] + u] \leq e^{-\frac{nu^2}{2\tilde{\sigma}^2 t^2}}. \quad (13.64)$$

We first derive a bound for $t = \delta_n$. By the definitions of \mathcal{G}_n and the critical radius δ_n , we have $\mathbb{E}[Z_n(\delta_n)] \leq \tilde{\sigma}\mathcal{G}_n(\delta_n) \leq \delta_n^2$. Setting $u = \delta_n$ in the tail bound (13.64), we find that

$$\mathbb{P}[Z_n(\delta_n) \geq 2\delta_n^2] \leq e^{-\frac{n\delta_n^2}{2\tilde{\sigma}^2}}. \quad (13.65a)$$

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

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

where inequality (i) follows from Lemma 13.1, and inequality (ii) follows by our choice of δ_n . Using this upper bound on the mean and setting $u = t^2/32$ in the tail bound (13.64)

yields

$$\mathbb{P}\left[Z_n(t) \geq t\delta_n + \frac{t^2}{32}\right] \leq e^{-c_2 \frac{nt^2}{\bar{\sigma}^2}} \quad \text{for each } t > \delta_n. \quad (13.65b)$$

We are now equipped to complete the proof by a “peeling” argument. Let \mathcal{E} denote the event that the bound (13.61) is violated for some function g 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 $m = 0, 1, 2, \dots$, 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 union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}(0, t_0)] + \sum_{m=0}^{\infty} \mathbb{P}[\mathcal{E}(t_m, t_{m+1})]. \quad (13.66)$$

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

$$\mathbb{P}[\mathcal{E}(0, t_0)] \leq \mathbb{P}[Z_n(\delta_n) \geq 2\delta_n^2] \leq e^{-\frac{n\delta_n^2}{2\bar{\sigma}^2}}, \quad (13.67)$$

using our earlier tail bound (13.65a). 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

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

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 (13.65b) yields

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

Substituting this inequality and our earlier bound (13.67) into equation (13.66) yields

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

where the reader should recall that the precise values of universal constants may change from line-to-line. 37

38

■ 13.5 Bibliographic details and background

39

1 Non-parametric regression is a classical problem in statistics with a lengthy and rich
 2 history. Although this chapter is limited to the method of non-parametric least squares,
 3 there are a variety of other cost functions that can be used for regression, which might
 4 be preferable for reasons of robustness. The techniques described in this chapter can be
 5 applied to analyze M -estimators—that is, methods based on minimizing or maximizing
 6 some criterion of fit—based on cost functions other than least squares. In addition,
 7 non-parametric regression can be tackled via methods that are not naturally viewed as
 8 M -estimators, including orthogonal function expansions, local polynomial representa-
 9 tions, kernel density estimators, nearest neighbor methods, and scatterplot smoothing
 10 methods, among others. We refer the reader to the books [EL07, GKKW02, HMSW04,
 11 Tsy09, Was06] for further background on these and other methods.

12 Optimal rates for non-parametric regression were first obtained by Stone [Sto82]
 13 using techniques for lower bounds to be discussed in Chapter 15. Stone [Sto85] in-
 14 troduced the class of additive non-parametric regression models discussed in Exer-
 15 cise 13.8, and subsequent work has explored many extensions and variants of these
 16 models (e.g., [HT86, BHT89, MvdGB09, RLLW09, RWY12]). Exercise 13.8 in this
 17 chapter and Exercise 14.7 in Chapter 14 explore some properties of the standard addi-
 18 tive model. The ridge regression estimator from Examples 13.1 and 13.5 was introduced
 19 by Hoerl and Kennard [HK70]. The Lasso estimator from Example 13.1 was discussed
 20 at length in Chapter 7. The ℓ_q -ball constrained estimators from Examples 13.1 and 13.6
 21 were analyzed by Raskutti et al. [RWY11], who also used information-theoretic meth-
 22 ods, to be discussed in Chapter 15, in order to derive matching lower bounds. The
 results on metric entropies of q -convex hulls in this example are based on results from
 Carl and Pajor [CP88], and Guédon and Litvak [GL00]; see also the arguments in the
 paper [RWY11] for details on the specific claims given here.

The idea of localization plays an important role in empirical process theory, and
 we embark on a more in-depth study of it in Chapter 14 to follow. Local func-
 tion complexities of the form given in Corollary 13.1 are used extensively by van de
 Geer in her book [vdG00], whereas other authors have studied localized forms of the
 Rademacher and Gaussian complexities [Kol01, Kol06, BBM05]. The bound on the
 localized Rademacher complexity of reproducing kernel Hilbert spaces, as stated in
 Lemma 13.3, is due Mendelson [Men02]; see also the paper by Bartlett and Mendel-
 son [BM02] for related results. The peeling technique used in the proof of Lemma 13.4
 is widely used in empirical process theory [Ale87, vdG00].

3 The bound (13.32) on the sup-norm (L_∞) metric entropy for bounded convex Lips-
 4 chitz functions is due to Bronshtein [Bro76]; see also Section 8.4 of Dudley [Dud99] for
 5 more details. Note that the class of all convex functions $f : [0, 1] \rightarrow [0, 1]$ without any
 6 Lipschitz constraint is *not* totally bounded in the sup-norm metric; see Exercise 5.1 for
 7 details. Guntuboyina and Sen [GS13] provide bounds on the entropy in the L_p -metrics

8 over the range $p \in [1, \infty)$ for convex functions without the Lipschitz condition.

9 ■ 13.6 Exercises

10 **Exercise 13.1.** (a) Given a random variable Z with finite second moment, show that the function $G(t) = \mathbb{E}[(Z - t)^2]$ is minimized at $t = \mathbb{E}[Z]$. 1

(b) Assuming that all relevant expectations exist, show that the minimizer of the population mean-squared error (13.1) is given by the conditional expectation $f^*(x) = \mathbb{E}[Y \mid X = x]$. (*Hint:* The tower property and part (a) may be useful to you.) 2
3
4
5

(c) Let f be any other function for which the mean-squared error $\mathbb{E}_{X,Y}[(Y - f(X))^2]$ is finite. Show that the excess risk of f is given by $\|f - f^*\|_2^2$, as in equation (13.4). 6
7

Exercise 13.2. Recall the cubic spline estimate (13.10) from Example 13.2, as well as the kernel function $\mathcal{K}(x, z) = \int_0^1 (x - z)_+ (y - z)_+ dz$ from Example 12.21. 8
9

(a) Show that the optimal solution must take the form

$$\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 x + \frac{1}{\sqrt{n}} \sum_{i=1} \hat{\alpha}_i \mathcal{K}(x, x_i)$$

for some vectors $\hat{\theta} \in \mathbb{R}^2$ and $\hat{\alpha} \in \mathbb{R}^n$. 10

(b) Show that these vectors can be obtained by solving the quadratic program

$$(\hat{\theta}, \hat{\alpha}) = \arg \min_{(\theta, \alpha) \in \mathbb{R}^2 \times \mathbb{R}^n} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta - \sqrt{n}\mathbf{K}\alpha\|_2^2 + \lambda_n \alpha^T \mathbf{K}\alpha \right\}$$

where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix defined by the kernel function in part (a), 11
and $\mathbf{X} \in \mathbb{R}^{n \times 2}$ is a design matrix with i^{th} row given by $[1 \ x_i]$. 12

Exercise 13.3. (a) Show that a set \mathcal{C} is star-shaped around one of its point x^* if and only if the points $\alpha x + (1 - \alpha)x^*$ belongs to \mathcal{C} for all $x \in \mathcal{C}$ and scalars $\alpha \in [0, 1]$. 13
14

(b) Show that a set \mathcal{C} is convex if and only if it is star-shaped around each one of its points. 15
16

1 **Exercise 13.4.** Consider the critical inequality (13.16) in the case $f^* = 0$, so that
2 $\mathcal{F}^* = \mathcal{F}$.

3 (a) Show that the critical inequality (13.16) is always satisfied for $\delta^2 = 4\sigma^2$.

4 (b) Suppose that \mathcal{F} contains the constant function $f \equiv 1$. Show that any $\delta \in (0, 1]$
5 satisfying the critical inequality (13.16) must be lower bounded as $\delta^2 \geq \min \left\{ 1, \frac{8}{\pi} \frac{\sigma^2}{n} \right\}$.

Exercise 13.5. This exercise illustrates how, even for a fixed base function class, the local Gaussian complexity $\mathcal{G}_n(\delta; \mathcal{F}^*)$ of the shifted function class can vary dramatically as the target function f^* is changed.

For each $\theta \in \mathbb{R}^n$, let $f_\theta(x) = \langle \theta, x \rangle$ be a linear function, and consider the class $\mathcal{F} = \{f_\theta \mid \|\theta\|_1 \leq 1\}$. Suppose that we observe samples of some unknown f_{θ^*} in this class, evaluated at the standard basis vectors $\{e_i\}_{i=1}^n$ —that is, $y_i = f_{\theta^*}(e_i) + \sigma w_i$ for $i = 1, \dots, n$, where $w_i \sim N(0, 1)$ is an i.i.d. noise sequence.

- (a) For any $f_{\theta^*} \in \mathcal{F}$, show that $\mathcal{G}_n(\delta; \mathcal{F}^*) \leq c_1 \sqrt{\frac{\log n}{n}}$ for some universal constant c_1 . Explain how this result can be used to derive an estimator $\hat{\theta}$ such that

$$\|\hat{\theta} - \theta^*\|_2^2 \leq c'_1 \sqrt{\frac{\sigma^2 \log n}{n}}$$

with high probability.

- (b) Now consider some f_{θ^*} such that $\|\theta^*\|_0 = 1$ —that is, θ^* has exactly one non-zero entry. Show that there is a universal constant c_2 such that $\mathcal{G}_n(\delta; \mathcal{F}^*) \leq c_2 \delta \sqrt{\frac{\log n}{n}}$, and that this calculation leads to an estimator such that

$$\|\hat{\theta} - \theta^*\|_2^2 \leq c'_2 \frac{\sigma^2 \log n}{n}$$

with high probability.

Exercise 13.6. Consider the class of all $m - 1$ -degree polynomials

$$\mathcal{P}_m = \{f_\theta : \mathbb{R} \rightarrow \mathbb{R} \mid \theta \in \mathbb{R}^m\}, \quad \text{where } f_\theta(x) = \sum_{j=0}^{m-1} \theta_j x^j,$$

and suppose that $f^* \in \mathcal{P}_m$. Show that there are universal positive constants (c_0, c_1, c_2) such that the least-squares estimator satisfies

$$\mathbb{P} \left[\|\hat{f} - f^*\|_n^2 \geq c_0 \frac{\sigma^2 m \log n}{n} \right] \leq c_1 e^{-c_2 m \log n}.$$

Exercise 13.7. Consider the function class \mathcal{F} of functions $f : [0, 1] \rightarrow \mathbb{R}$ that are twice differentiable with $\|f\|_\infty + \|f'\|_\infty + \|f''\|_\infty \leq C$ for some constant $C < \infty$. Show that there are positive constants (c_0, c_1, c_2) , which may depend on C but not on (n, σ^2) , such that the non-parametric least squares estimate satisfies

$$\mathbb{P} \left[\|\hat{f} - f^*\|_n^2 \geq c_0 \left(\frac{\sigma^2}{n} \right)^{\frac{4}{5}} \right] \leq c_1 e^{-c_2 (n/\sigma^2)^{1/5}}.$$

(Hint: Results from Chapter 5 may be useful to you.)

Exercise 13.8. Given a class \mathcal{G} of univariate functions $g : \mathbb{R} \rightarrow \mathbb{R}$, consider the class of additive functions over \mathbb{R}^d , namely

$$\mathcal{F}_{\text{additive}} = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f = \sum_{j=1}^d g_j \text{ for some } g_j \in \mathcal{G}. \right\}. \quad (13.68)$$

Under the conditions of Theorem 13.1, show that the non-parametric least-squares estimate \hat{f} over the class $\mathcal{F}_{\text{additive}}$ satisfies the bound

$$\|\hat{f}_n - f^*\|_n^2 \leq c d \varepsilon^2, \quad (13.69)$$

with high probability, where ε is the critical rate (13.16) determined by the univariate function class \mathcal{G} .

Exercise 13.9. Recall the function class $\mathcal{F}_{\text{ortho}}(1; T)$ from Example 13.9 defined by orthogonal series expansion with T coefficients.

- (a) Given a set of design points $\{x_1, \dots, x_n\}$, define the $n \times T$ matrix $\Phi \equiv \Phi(x_1^n)$ with $(i, j)^{\text{th}}$ entry $\Phi_{ij} = \phi_j(x_i)$. Show that the non-parametric least squares estimate \hat{f} over $\mathcal{F}_{\text{ortho}}(1; T)$ can be obtained by solving the ridge regression problem

$$\min_{\theta \in \mathbb{R}^T} \left\{ \frac{1}{2n} \|y - \Phi \theta\|_2^2 + \lambda_n \|\theta\|_2^2 \right\}$$

for a suitable choice of regularization parameter $\lambda_n \geq 0$.

- (b) Show that $\inf_{f \in \mathcal{F}_{\text{ortho}}(1; T)} \|f - f^*\|_2^2 = \sum_{j=T+1}^{\infty} \theta_j^2$.

Exercise 13.10. For a given integer $\alpha \geq 1$ and radius $R > 0$, consider the class of functions

$$\mathcal{F}_{\alpha}(R) = \{f \in L^2[0, 1] \mid f \text{ is } \alpha\text{-times differentiable, with } \int_0^1 (f^{(\alpha)}(x))^2 dx \leq R.\}$$

- (a) For a function $f \in \mathcal{F}_{\alpha}(R)$, let $\{\beta_0, (\beta_m, \beta_m)_{m=1}^{\infty}\}$ be its Fourier coefficients as previously defined in Example 13.10. Show that there is a constant c such that $\beta_m^2 + \tilde{\beta}_m^2 \leq \frac{c}{m^{2\alpha}}$ for all $m \geq 1$.
- (b) Verify the approximation-theoretic guarantee (13.46).