13

Nonparametric least squares

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

13.1 Problem set-up

A regression problem is defined by a set of predictors or covariates $x \in \mathcal{X}$, along with a response variable $y \in \mathcal{Y}$. Throughout this chapter, we focus on the case of real-valued response variables, in which the space $\mathcal{Y}$ is the real line or some subset thereof. Our goal is to estimate a function $f\colon \mathcal{X} \to \mathcal{Y}$ 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* (MSE)

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

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

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

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 joint 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 analog of the mean-squared error, namely

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

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

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 $\bar{\mathcal{L}}_{f^*}$ achieved by the

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

regression function f^*, and that achieved by the estimate f. In the special case of the least-squares cost function, it can be shown (see Exercise 13.1) that this excess risk takes the form

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

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

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

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

In order to lighten notation, we frequently use $\|\widehat{f} - f^*\|_n$ as a shorthand for the more cumbersome $\|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}$. 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.

13.1.2 Estimation via constrained least squares

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

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

where v_i is a random variable representing the "noise" in the ith response variable. Note that these noise variables must have zero mean, given the form (13.2) of the regression function f^*. Apart from this zero-mean property, their structure in general depends on the distribution of the conditioned random variable $(Y \mid X = x)$. In the *standard nonparametric regression* model, we assume the noise variables are drawn in an i.i.d. manner from the $\mathcal{N}(0, \sigma^2)$ distribution, where $\sigma > 0$ is a standard deviation parameter. In this case, we can write $v_i = \sigma w_i$, where $w_i \sim \mathcal{N}(0, 1)$ is a Gaussian random variable.

Given this set-up, one way in which to estimate the regression function f^* is by constrained least squares—that is, by solving the problem[1]

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

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

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

Typically, we restrict the optimization problem (13.7) to some appropriately chosen subset of $\mathscr{F}$—for instance, a ball of radius R in an underlying norm $\|\cdot\|_{\mathscr{F}}$. Choosing $\mathscr{F}$ to be a reproducing kernel Hilbert space, as discussed in Chapter 12, can be useful for computational reasons. It can also be convenient to use regularized estimators of the form

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

where $\lambda_n > 0$ is a suitably chosen regularization weight. We return to analyze such estimators in Section 13.4.

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 linear function $f_\theta(x) = \langle \theta,\ x\rangle$. Given a compact subset $C \subseteq \mathbb{R}^d$, consider the function class

$$\mathscr{F}_C \ := \{f_\theta \colon \mathbb{R}^d \to \mathbb{R} \mid \theta \in C\}.$$

With this choice, the estimator (13.7) reduces to a constrained form of least-squares estimation, more specifically

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

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the design matrix with the vector $x_i \in \mathbb{R}^d$ in its ith row. Particular instances of this estimator include *ridge regression*, obtained by setting

$$C = \left\{\theta \in \mathbb{R}^d \mid \|\theta\|_2^2 \le R_2\right\}$$

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

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

for some $q \in [0, 2]$ and radius $R_q > 0$. See Figure 7.1 for an illustration of these sets for $q \in (0, 1]$. The constrained form of the Lasso (7.19), as analyzed in depth in Chapter 7, is a special but important case, obtained by setting $q = 1$.

Whereas the previous example was a parametric problem, we now turn to some nonparametric examples:

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

$$\mathscr{F}(R) := \left\{ f\colon [0,1] \to \mathbb{R} \mid \int_0^1 (f''(x))^2\, dx \le R \right\}, \tag{13.9}$$

where f'' denotes the second derivative of f. The integral constraint on f'' can be understood as a Hilbert norm bound in the second-order Sobolev space $\mathbb{H}^\alpha[0,1]$ introduced in Example 12.17. In this case, the penalized form of the nonparametric least-squares estimate is given by

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

where $\lambda_n > 0$ is a user-defined regularization parameter. It can be shown that any minimizer $\widehat{f}$ 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 $R \to 0$ (or equivalently, as $\lambda_n \to +\infty$), the cubic spline fit $\widehat{f}$ becomes a linear function, since we have $f'' = 0$ only for a linear function. ♣

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). Let us consider this family more generally:

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

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

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 $\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{\alpha}_i \mathcal{K}(\cdot, x_i)$, where $\widehat{\alpha} := (\mathbf{K} + \lambda_n \mathbf{I}_n)^{-1} \frac{y}{\sqrt{n}}$. In Exercise 13.3, we show how the spline estimator from Example 13.2 can be understood in the context of kernel ridge regression. ♣

Let us now consider an example of what is known as *shape-constrained* regression.

Example 13.4 (Convex regression) Suppose that $f^*\colon C \to \mathbb{R}$ is known to be a convex function over its domain 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

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

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 f is subdifferentiable at each point in the (relative) interior of its domain C. More precisely, at any interior point $x \in C$, there exists at least one vector $z \in \mathbb{R}^d$ such that

$$f(y) \geq f(x) + \langle z,\, y - x\rangle \qquad \text{for all } y \in C. \tag{13.11}$$

Any such vector is known as a *subgradient*, and each point $x \in C$ can be associated with the set $\partial f(x)$ of its subgradients, which is known as the *subdifferential* of f at x. When f is actually differentiable at x, then the lower bound (13.11) holds if and only if $z = \nabla f(x)$, so that we have $\partial f(x) = \{\nabla f(x)\}$. See the bibliographic section for some standard references in convex analysis.

Applying this fact to each of the sampled points $\{x_i\}_{i=1}^n$, we find that there must exist subgradient vectors $\widetilde{z}_i \in \mathbb{R}^d$ such that

$$f(x) \geq f(x_i) + \langle \widetilde{z}_i,\, x - x_i\rangle \qquad \text{for all } x \in C. \tag{13.12}$$

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

$$\min_{\{(\widetilde{y}_i, z_i)\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^n (y_i - \widetilde{y}_i)^2 \tag{13.13}$$
$$\text{such that} \quad \widetilde{y}_j \geq \widetilde{y}_i + \langle \widetilde{z}_i,\, x_j - x_i\rangle \qquad \text{for all } i, j = 1, 2, \ldots, n.$$

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.

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

$$\widehat{f}(x) := \max_{i=1,\ldots,n} \{\widehat{y}_i + \langle \widehat{z}_i,\, x - x_i\rangle\}. \tag{13.14}$$

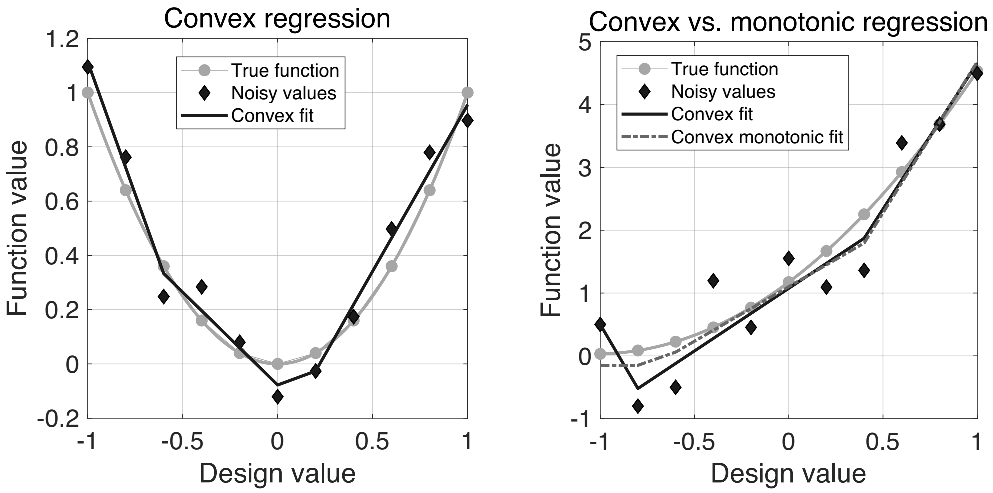
As the maximum of a collection of linear functions, the function $\widehat{f}$ is convex. Moreover, a short calculation—using the fact that $\{(\widehat{y}_i, \widehat{z}_i)\}_{i=1}^n$ are feasible for the program (13.13)—shows that $\widehat{f}(x_i) = \widehat{y}_i$ for all $i = 1, 2, \ldots, n$. Figure 13.1(a) provides an illustration of the convex regression estimate (13.14), 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, subgradients) are non-negative. In this case, it is natural to impose additional non-negativity constraints ($\widetilde{z}_j \geq 0$) on the subgradients in the estimator (13.13). Figure 13.1(b) compares the standard convex regression estimate with the estimator that imposes these additional monotonicity constraints. ♣

13.2 Bounding the prediction error

From a statistical perspective, an essential question associated with the nonparametric least-squares estimate (13.7) is how well it approximates the true regression function f^*. In this

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press



(a) Convex regression (b) Convex vs convex/monotonic regression

Figure 13.1 (a) Illustration of the convex regression estimate (13.14) based on a fixed design with $n = 11$ equidistant samples over the interval $C = [-1, 1]$. (b) Ordinary convex regression compared with convex and monotonic regression estimate.

section, we develop some techniques to bound the error $\|\widehat{f} - f^*\|_n$, as measured in the $L^2(\mathbb{P}_n)$-norm. In Chapter 14, we develop 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 $\mathscr{F}$ 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 $\mathscr{F}$, locally in a neighborhood around the true regression function f^*. More precisely, we define the set

$$\mathscr{F}^* := \mathscr{F} - \{f^*\} = \{f - f^* \mid f \in \mathscr{F}\}, \tag{13.15}$$

corresponding to an f^*-shifted version of the original function class $\mathscr{F}$. For a given radius $\delta > 0$, the *local Gaussian complexity* around f^* at scale δ is given by

$$\mathcal{G}_n(\delta; \mathscr{F}^*) := \mathbb{E}_w\Big[\sup_{\substack{g \in \mathscr{F}^* \\ \|g\|_n \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^{n} w_i g(x_i)\Big|\Big], \tag{13.16}$$

where the variables $\{w_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0, 1)$ variates. Throughout this chapter, this complexity measure should be understood as a deterministic quantity, since we are considering the case of fixed covariates $\{x_i\}_{i=1}^n$.

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

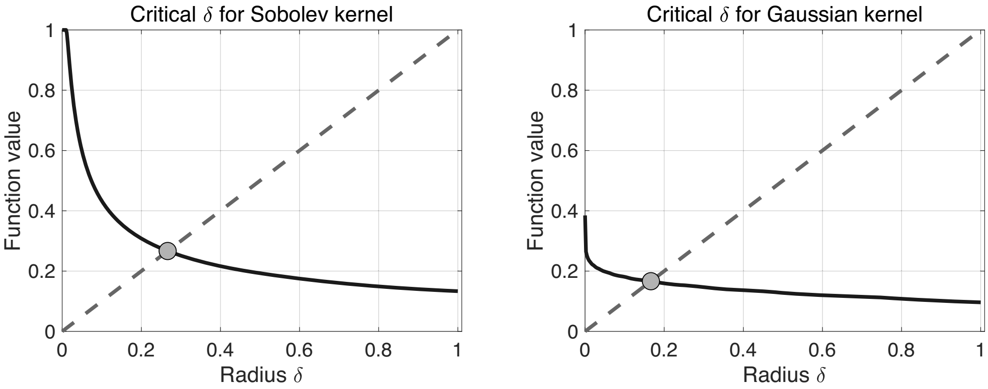
https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

inequality

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

As we verify in Lemma 13.6, whenever the shifted function class $\mathscr{F}^*$ is star-shaped,[2] the left-hand side is a non-increasing function of δ, which ensures that the inequality can be satisfied. We refer to any $\delta_n > 0$ satisfying inequality (13.17) as being *valid*, and we use $\delta_n^* > 0$ to denote the smallest positive radius for which inequality (13.17) holds. See the discussion following Theorem 13.5 for more details on the star-shaped property and the existence of valid radii δ_n.

Figure 13.2 illustrates the non-increasing property of the function $\delta \mapsto \mathcal{G}_n(\delta)/\delta$ for two different function classes: a first-order Sobolev space in Figure 13.2(a), and a Gaussian kernel space in Figure 13.2(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 δ_n^* can be determined by finding where this non-increasing function crosses the line with slope one, 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 correspondingly smaller. This ordering reflects the natural intuition that it should be easier to perform regression over a smaller function class.



(a) (b)

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

Some intuition: Why should the inequality (13.17) be relevant to the analysis of the nonparametric least-squares estimator? A little calculation is helpful in gaining intuition. Since $\widehat{f}$ and f^* are optimal and feasible, respectively, for the constrained least-squares prob-

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

lem (13.7), we are guaranteed that

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

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

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

which we call the *basic inequality for nonparametric least squares*.

Now, by definition, the difference function $\widehat{f} - f^*$ belongs to $\mathscr{F}^*$, so that we can bound the right-hand side by taking the supremum over all functions $g \in \mathscr{F}^*$ with $\|g\|_n \leq \|\widehat{f} - f^*\|_n$. Reasoning heuristically, this observation suggests that the squared error $\delta^2 := \mathbb{E}[\|\widehat{f} - f^*\|_n^2]$ should satisfy a bound of the form

$$\frac{\delta^2}{2} \leq \sigma\, \mathcal{G}_n(\delta; \mathscr{F}^*) \quad \text{or equivalently} \quad \frac{\delta}{2\sigma} \leq \frac{\mathcal{G}_n(\delta; \mathscr{F}^*)}{\delta}. \tag{13.19}$$

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

To be clear, the step from the basic inequality (13.18) to the bound (13.19) 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 the standard Gaussian noise model $y_i = f^*(x_i) + \sigma w_i$.

Theorem 13.5 *Suppose that the shifted function class $\mathscr{F}^*$ is star-shaped, and let δ_n be any positive solution to the critical inequality* (13.17). *Then for any $t \geq \delta_n$, the nonparametric least-squares estimate $\widehat{f}_n$ satisfies the bound*

$$\mathbb{P}\Big[\|\widehat{f}_n - f^*\|_n^2 \geq 16\, t\, \delta_n\Big] \leq e^{-\frac{n t \delta_n}{2\sigma^2}}. \tag{13.20}$$

Remarks: The bound (13.20) provides non-asymptotic control on the regression error $\|\widehat{f} - f^*\|_2^2$. By integrating this tail bound, it follows that the mean-squared error in the $L^2(\mathbb{P}_n)$-semi-norm is upper bounded as

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

As shown in Exercise 13.5, for any function class $\mathscr{F}$ that contains the constant function $f \equiv 1$, we necessarily have $\delta_n^2 \geq \frac{2}{\pi}\frac{\sigma^2}{n}$, so that (disregarding constants) the δ_n^2 term is always the dominant one.

For concreteness, we have stated the result for the case of additive Gaussian noise ($v_i = \sigma w_i$). However, as the proof will clarify, all that is required is an upper tail bound on the

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

random variable

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

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 $\mathscr{F}^* = \mathscr{F} - f^*$ is needed at various parts of the proof, including in ensuring the existence of valid radii δ_n (see Lemma 13.6 to follow). In explicit terms, the function class $\mathscr{F}^*$ is star-shaped if for any $g \in \mathscr{F}$ and $\alpha \in [0, 1]$, the function αg also belongs to $\mathscr{F}^*$. Equivalently, we say that $\mathscr{F}$ is star-shaped around f^*. For instance, if $\mathscr{F}$ is convex, then as illustrated in Figure 13.3 it is necessarily star-shaped around any $f^* \in \mathscr{F}$. Conversely, if $\mathscr{F}$ is not convex, then there must exist choices $f^* \in \mathscr{F}$ such that $\mathscr{F}^*$ is not star-shaped. However, for a general non-convex set $\mathscr{F}$, it is still possible that $\mathscr{F}^*$ is star-shaped for *some* choices of f^*. See Figure 13.3 for an illustration of these possibilities, and Exercise 13.4 for further details.

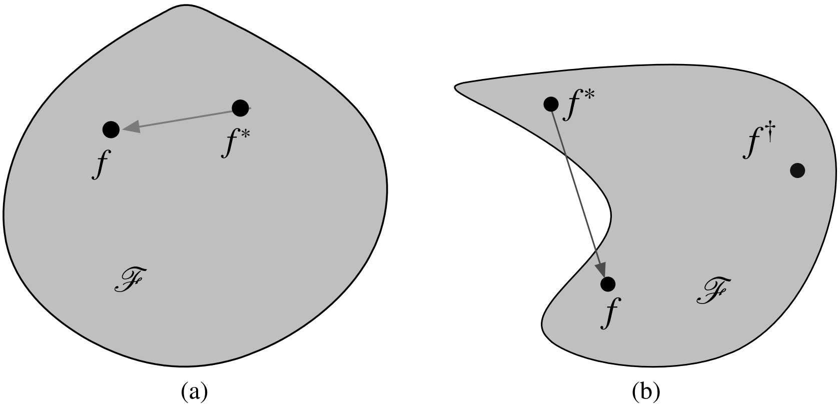


Figure 13.3 Illustration of star-shaped properties of sets. (a) The set $\mathscr{F}$ is convex, and hence is star-shaped around any of its points. The line between f^* and f is contained within $\mathscr{F}$, and the same is true for any line joining any pair of points in $\mathscr{F}$. (b) A set $\mathscr{F}$ that is not star-shaped around all its points. It fails to be star-shaped around the point f^*, since the line drawn to $f \in \mathscr{F}$ does not lie within the set. However, this set is star-shaped around the point $f^\dagger$.

If the star-shaped condition fails to hold, then Theorem 13.5 can instead by applied with δ_n defined in terms of the *star hull*

$$\text{star}(\mathscr{F}^*; 0) := \{\alpha g \mid g \in \mathscr{F}^*, \alpha \in [0,1]\} = \{\alpha(f - f^*) \mid f \in \mathscr{F}, \alpha \in [0,1]\}. \tag{13.21}$$

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

$$\partial\mathscr{F} := \mathscr{F} - \mathscr{F} = \{f_1 - f_2 \mid f_1, f_2 \in \mathscr{F}\}, \tag{13.22}$$

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

to follow.

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

Lemma 13.6 *For any star-shaped function class $\mathscr{H}$, the function $\delta \mapsto \frac{\mathcal{G}_n(\delta;\mathscr{H})}{\delta}$ is non-increasing on the interval $(0, \infty)$. Consequently, for any constant $c > 0$, the inequality*

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

has a smallest positive solution.

Proof So as to ease notation, we drop the dependence of $\mathcal{G}_n$ on the function class $\mathscr{H}$ 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 $h \in \mathscr{H}$ with $\|h\|_n \leq t$, we may define the rescaled function $\widetilde{h} = \frac{\delta}{t}h$, and write

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

By construction, we have $\|\widetilde{h}\|_n \leq \delta$; moreover, since $\delta \leq t$, the star-shaped assumption guarantees that $\widetilde{h} \in \mathscr{H}$. Consequently, for any $\widetilde{h}$ formed in this way, the right-hand side is at most $\mathcal{G}_n(\delta)$ in expectation. Taking the supremum over the set $\mathscr{H} \cap \{\|h\|_n \leq t\}$ followed by expectations yields $\mathcal{G}_n(t)$ on the left-hand side. Combining the pieces yields the claim. □

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 13.5, 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 the 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 $\mathscr{H}$, we define $\mathbb{B}_n(\delta;\mathscr{H}) := \{h \in \text{star}(\mathscr{H}) \mid \|h\|_n \leq \delta\}$, and we let $N_n(t;\mathbb{B}_n(\delta;\mathscr{H}))$ denote the t-covering number of $\mathbb{B}_n(\delta;\mathscr{H})$ in the norm $\|\cdot\|_n$. With this notation, we have the following corollary:

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Corollary 13.7 *Under the conditions of Theorem 13.5, any* $\delta \in (0, \sigma]$ *such that*

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t;\ \mathbb{B}_n(\delta; \mathscr{F}^*))}\, dt \leq \frac{\delta^2}{4\sigma} \tag{13.24}$$

satisfies the critical inequality (13.17), *and hence can be used in the conclusion of Theorem 13.5.*

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

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

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

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

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

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

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

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

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

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Some examples are helpful in understanding the uses of Theorem 13.5 and Corollary 13.7, and we devote the following subsections to such illustrations.

13.2.2 Bounds for high-dimensional parametric problems

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

Example 13.8 (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

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

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

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

with high probability. To be clear, in this special case, this bound (13.26) can be obtained by a direct linear algebraic argument, as we explore in Exercise 13.2. However, it is instructive to see how our general theory leads to concrete predictions in a special case.

We begin by observing that the shifted function class $\mathscr{F}^*_{\text{lin}}$ is equal to $\mathscr{F}_{\text{lin}}$ for any choice of f^*. Moreover, the set $\mathscr{F}_{\text{lin}}$ is convex and hence star-shaped around any point (see Exercise 13.4), so that Corollary 13.7 can be applied. The mapping $\theta \mapsto \|f_\theta\|_n = \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}}$ defines a norm on the subspace range($\mathbf{X}$), and the set $\mathbb{B}_n(\delta; \mathscr{F}_{\text{lin}})$ is isomorphic to a δ-ball within the space range($\mathbf{X}$). Since this range space has dimension given by rank($\mathbf{X}$), by a volume ratio argument (see Example 5.8), we have

$$\log N_n(t; \mathbb{B}_n(\delta; \mathscr{F}_{\text{lin}})) \leq r \log\Big(1 + \frac{2\delta}{t}\Big), \qquad \text{where } r := \text{rank}(\mathbf{X}).$$

Using this upper bound in Corollary 13.7, we find that

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

where we have made the change of variables $u = t/\delta$ in step (i), and the final step (ii) follows since the integral is a constant. Putting together the pieces, an application of Corollary 13.7 yields the claim (13.26). In fact, the bound (13.26) 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.

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Example 13.9 (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

$$\mathbb{B}_q(R_q) := \Big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q\Big\}. \tag{13.27}$$

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

$$\mathscr{F}_q(R_q) := \Big\{f_\theta \mid \theta \in \mathbb{B}_q(R_q)\Big\}. \tag{13.28}$$

We adopt the shorthand $\mathscr{F}_q$ 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

$$\widehat{\theta} \in \arg \min_{\theta\in\mathbb{B}_q(R_q)} \left\{\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \langle x_i,\ \theta\rangle\big)^2\right\}. \tag{13.29}$$

Unlike the ℓ_1-constrained Lasso analyzed in Chapter 7, note that this is *not* a convex program. Indeed, for $q \in (0, 1)$, the function class $\mathscr{F}_q(R_q)$ is not convex, so that there exists $\theta^* \in \mathbb{B}_q(R_q)$ such that the shifted class $\mathscr{F}_q^* = \mathscr{F}_q - f_{\theta^*}$ is not star-shaped. Accordingly, we instead focus on bounding the metric entropy of the function class $\mathscr{F}_q(R_q) - \mathscr{F}_q(R_q) = 2\mathscr{F}_q(R_q)$. Note that 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 $2\mathscr{F}_q(R_q)$ is contained with $\mathscr{F}_q(2R_q)$.

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

$$\log N_{2,q}(t) \leq C_q\Big[R_q^{\frac{2}{2-q}}\big(\frac{1}{t}\big)^{\frac{2q}{2-q}}\log d\Big], \tag{13.30}$$

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 ith row, and let $X_j \in \mathbb{R}^n$ denote its jth column. Our objective is to bound the metric entropy of the set of all vectors of the form

$$\frac{\mathbf{X}\theta}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{j=1}^{d} X_j\theta_j \tag{13.31}$$

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

Exploiting this fact and our earlier bound (13.30) on the metric entropy of the ℓ_q-ball, we find that

$$\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n\big(t;\ \mathbb{B}_n(\delta; \mathscr{F}_q(2R_q))\big)}\, dt \precsim R_q^{\frac{1}{2-q}} \sqrt{\frac{\log d}{n}} \int_0^{\delta} \Big(\frac{1}{t}\Big)^{\frac{q}{2-q}} dt$$

$$\precsim R_q^{\frac{1}{2-q}} \sqrt{\frac{\log d}{n}}\, \delta^{1-\frac{q}{2-q}},$$

a calculation valid for all $q \in (0,1)$. Corollary 13.7 now implies that the critical condition (13.17) is satisfied as long as

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

Theorem 13.5 then implies that

$$\|f_{\widehat{\theta}} - f_{\theta^*}\|_n^2 = \frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} \precsim R_q\Big(\frac{\sigma^2\ \log d}{n}\Big)^{1-\frac{q}{2}},$$

with high probability. Although this result is a corollary of our general theorem, 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 of these connections. ♣

13.2.3 Bounds for nonparametric problems

Let us now illustrate the use of our techniques for some nonparametric problems.

Example 13.10 (Bounds for Lipschitz functions) Consider the class of functions

$$\mathscr{F}_{\text{Lip}}(L) := \{f\colon [0,1] \to \mathbb{R} \mid f(0) = 0,\ f \text{ is } L\text{-Lipschitz}\}. \tag{13.32}$$

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 prediction error associated with nonparametric least squares over this function class.

Noting the inclusion

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

it suffices to upper bound the metric entropy of $\mathscr{F}_{\text{Lip}}(2L)$. Based on our discussion from Example 5.10, the metric entropy of this class in the supremum norm scales as $\log N_\infty(\epsilon; \mathscr{F}_{\text{Lip}}(2L)) \simeq (L/\epsilon)$. Consequently, we have

$$\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log N_n(t;\ \mathbb{B}_n(\delta; \mathscr{F}_{\text{Lip}}(2L)))}\, dt \precsim \int_0^{\delta} \sqrt{\log N_\infty(t;\ \mathscr{F}_{\text{Lip}}(2L))}\, dt$$

$$\precsim \frac{1}{\sqrt{n}} \int_0^{\delta} (L/t)^{\frac{1}{2}}\, dt$$

$$\precsim \frac{1}{\sqrt{n}} \sqrt{L\delta},$$

where $\precsim$ denotes an inequality holding apart from constants not dependent on the triplet (δ, L, n). Thus, it suffices to choose $\delta_n > 0$ such that $\frac{\sqrt{L\delta_n}}{\sqrt{n}} \precsim \frac{\delta_n^2}{\sigma}$, or equivalently $\delta_n^2 \simeq (\frac{L\sigma^2}{n})^{\frac{2}{3}}$.

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Putting together the pieces, Corollary 13.7 implies that the error in the nonparametric least-squares estimate satisfies the bound

$$\|\widehat{f} - f^*\|_n^2 \precsim \Big(\frac{L\sigma^2}{n}\Big)^{2/3} \tag{13.33}$$

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

Example 13.11 (Bounds for convex regression) As a continuation of the previous example, let us consider the class of *convex* 1-Lipschitz functions, namely

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

As discussed in Example 13.4, computation of the nonparametric 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 $\mathscr{F}_{\text{conv}}$, when measured in the infinity norm, satisfies the upper bound

$$\log N(\epsilon; \mathscr{F}_{\text{conv}}, \|\cdot\|_\infty) \precsim \Big(\frac{1}{\epsilon}\Big)^{1/2} \tag{13.34}$$

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 prediction error. In particular, calculations similar to those in the previous example show that the conditions of Corollary 13.7 hold for $\delta_n^2 \simeq (\frac{\sigma^2}{n})^{\frac{4}{5}}$, and so we are guaranteed that

$$\|\widehat{f} - f^*\|_n^2 \precsim \Big(\frac{\sigma^2}{n}\Big)^{4/5} \tag{13.35}$$

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

Note that our error bound (13.35) for convex Lipschitz functions is substantially faster than our earlier bound (13.33) for Lipschitz functions *without a convexity constraint*—in particular, the respective rates are $n^{-4/5}$ versus $n^{-2/3}$. In Chapter 15, we show that both of these rates are minimax-optimal, meaning that, apart from constant factors, they cannot be improved substantially. Thus, we see that the additional constraint of convexity is significant from a statistical point of view. In fact, as we explore in Exercise 13.8, in terms of their estimation error, convex Lipschitz functions behave exactly like the class of all twice-differentiable functions with bounded second derivative, so that the convexity constraint amounts to imposing an extra degree of smoothness. ♣

13.2.4 Proof of Theorem 13.5

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Establishing a basic inequality

Recall the basic inequality (13.18) established in our earlier discussion. In terms of the shorthand notation $\hat{\Delta} = \widehat{f} - f^*$, it can be written as

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

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

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 is useful for subsequent arguments. Let $\mathscr{H}$ be an arbitrary star-shaped function class, and let $\delta_n > 0$ satisfy the inequality $\frac{\mathcal{G}_n(\delta;\mathscr{H})}{\delta} \leq \frac{\delta}{2\sigma}$. For a given scalar $u \geq \delta_n$, define the event

$$\mathcal{A}(u) := \left\{\exists\, g \in \mathscr{H} \cap \{\|g\|_n \geq u\} \;\mid\; \left|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(x_i)\right| \geq 2\|g\|_n u\right\}. \tag{13.37}$$

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

Lemma 13.12 *For all $u \geq \delta_n$, we have*

$$\mathbb{P}[\mathcal{A}(u)] \leq e^{-\frac{nu^2}{2\sigma^2}}. \tag{13.38}$$

Let us prove the main result by exploiting this lemma, in particular with the settings $\mathscr{H} = \mathscr{F}^*$ and $u = \sqrt{t\delta_n}$ for some $t \geq \delta_n$, 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 $\hat{\Delta} \in \mathscr{F}^*$ 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

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

Consequently, the basic inequality (13.36) implies that $\|\hat{\Delta}\|_n^2 \leq 4\|\hat{\Delta}\|_n\,\sqrt{t\delta_n}$, or equivalently that $\|\hat{\Delta}\|_n^2 \leq 16t\delta_n$, a bound that holds with probability at least $1 - e^{-\frac{nt\delta_n}{2\sigma^2}}$.

In order to complete the proof of Theorem 13.5, it remains to prove Lemma 13.12.

Proof of Lemma 13.12

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

such that

$$\Big|\frac{\sigma}{n}\sum_{i=1}^{n} w_i g(x_i)\Big| \geq 2\|g\|_n u. \tag{13.39}$$

Defining the function $\widetilde{g} := \frac{u}{\|g\|_n} g$, we observe that $\|\widetilde{g}\|_n = u$. Since $g \in \mathscr{H}$ and $\frac{u}{\|g\|_n} \in (0, 1]$, the star-shaped assumption implies that $\widetilde{g} \in \mathscr{H}$. Consequently, we have shown that if there exists a function g satisfying the inequality (13.39), which occurs whenever the event $\mathcal{A}(u)$ is true, then there exists a function $\widetilde{g} \in \mathscr{H}$ with $\|\widetilde{g}\|_n = u$ such that

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

We thus conclude that

$$\mathbb{P}[\mathcal{A}(u)] \leq \mathbb{P}[Z_n(u) \geq 2u^2], \quad \text{where} \quad Z_n(u) := \sup_{\substack{\widetilde{g}\in\mathscr{H} \\ \|\widetilde{g}\|_n \leq u}} |\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{g}(x_i)|. \tag{13.40}$$

Since the noise variables $w_i \sim \mathcal{N}(0, 1)$ are i.i.d., the variable $\frac{\sigma}{n}\sum_{i=1}^{n} w_i \widetilde{g}(x_i)$ is zero-mean and Gaussian for each fixed $\widetilde{g}$. Therefore, the variable $Z_n(u)$ corresponds to the supremum of a Gaussian process. If we view this supremum as a function of the standard Gaussian vector $(w_1, \ldots, w_n)$, then it can be verified that the associated Lipschitz constant is at most $\frac{\sigma u}{\sqrt{n}}$. Consequently, Theorem 2.26 guarantees the tail bound $\mathbb{P}[Z_n(u) \geq \mathbb{E}[Z_n(u)] + s] \leq e^{-\frac{ns^2}{2u^2\sigma^2}}$, valid for any $s > 0$. Setting $s = u^2$ yields

$$\mathbb{P}[Z_n(u) \geq \mathbb{E}[Z_n(u)] + u^2] \leq e^{-\frac{nu^2}{2\sigma^2}}. \tag{13.41}$$

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 13.6, the function $v \mapsto \frac{\mathcal{G}_n(v)}{v}$ is non-decreasing, and since $u \geq \delta_n$ by assumption, we have

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

where step (i) uses the critical condition (13.17). Putting together the pieces, we have shown that $\mathbb{E}[Z_n(u)] \leq u\delta_n$. Combined with the tail bound (13.41), we obtain

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

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

13.3 Oracle inequalities

In our analysis thus far, we have assumed that the regression function f^* belongs to the function class $\mathscr{F}$ over which the constrained least-squares estimator (13.7) is defined. In practice, this assumption might be violated, but it is nonetheless of interest to obtain bounds on the performance of the nonparametric least-squares estimator. In such settings, we expect its performance to involve both the *estimation error* that arises in Theorem 13.5, and some additional form of *approximation error*, arising from the fact that $f^* \notin \mathscr{F}$.

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

A natural way in which to measure approximation error is in terms of the best approximation to f^* using functions from $\mathscr{F}$. In the setting of interest in this chapter, the error in this best approximation is given by $\inf_{f \in \mathscr{F}} \|f - f^*\|_n^2$. Note that this error can only be achieved by an "oracle" that has direct access to the samples $\{f^*(x_i)\}_{i=1}^n$. 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 13.5. 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 $\partial\mathscr{F} = \{f_1 - f_2 \mid f_1, f_2 \in \mathscr{F}\}$. We assume that this set is star-shaped; if not, it should be replaced by its star hull in the results to follow.

Theorem 13.13 *Let δ_n be any positive solution to the inequality*

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

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

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

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

Remarks: Note that the guarantee (13.42b) is actually a family of bounds, one for each $f \in \mathscr{F}$. When $f^* \in \mathscr{F}$, then we can set $f = f^*$, so that the bound (13.42b) reduces to asserting that $\|\widehat{f} - f^*\|_n^2 \precsim t\delta_n$ with high probability, where δ_n satisfies our previous critical inequality (13.17). Thus, up to constant factors, we recover Theorem 13.5 as a special case of Theorem 13.13. In the more general setting when $f^* \notin \mathscr{F}$, setting $t = \delta_n$ and taking the infimum over $f \in \mathscr{F}$ yields an upper bound of the form

$$\|\widehat{f} - f^*\|_n^2 \precsim \inf_{f \in \mathscr{F}} \|f - f^*\|_n^2 + \delta_n^2. \tag{13.43a}$$

Similarly, by integrating the tail bound, we are guaranteed that

$$\mathbb{E}\Big[\|\widehat{f} - f^*\|_n^2\Big] \precsim \inf_{f \in \mathscr{F}} \|f - f^*\|_n^2 + \delta_n^2 + \frac{\sigma^2}{n}. \tag{13.43b}$$

These forms of the bound clarify the terminology *oracle inequality*: more precisely, the quantity $\inf_{f \in \mathscr{F}} \|f - f^*\|_n^2$ is the error achievable only by an oracle that has access to uncorrupted samples of the function f^*. The bound (13.43a) guarantees that the least-squares estimate $\widehat{f}$ has prediction error that is at most a constant multiple of the oracle error, plus a term proportional to δ_n^2. The term $\inf_{f \in \mathscr{F}} \|f - f^*\|_n^2$ can be viewed a form of *approximation error* that decreases as the function class $\mathscr{F}$ grows, whereas the term δ_n^2 is the *estimation error* that increases as $\mathscr{F}$ becomes more complex. This upper bound can thus be used to choose $\mathscr{F}$ 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.

13.3.1 Some examples of oracle inequalities

Theorem 13.13 as well as oracle inequality (13.43a) are best understood by applying them to derive explicit rates for some particular examples.

Example 13.14 (Orthogonal series expansion) Let $(\phi_m)_{m=1}^\infty$ be an orthonormal basis of $L^2(\mathbb{P})$, and for each integer $T = 1, 2, \ldots$, consider the function class

$$\mathscr{F}_{\text{ortho}}(1;T) := \Big\{f = \sum_{m=1}^{T} \beta_m \phi_m \mid \sum_{m=1}^{T} \beta_m^2 \leq 1\Big\}, \tag{13.44}$$

and let $\widehat{f}$ be the constrained least-squares estimate over this class. Its computation is straightforward: it reduces to a version of linear ridge regression (see Exercise 13.10).

Let us consider the guarantees of Theorem 13.13 for $\widehat{f}$ as an estimate of some 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 have $f^* = \sum_{m=1}^\infty \theta_m^* \phi_m$ for some coefficient sequence $(\theta_m^*)_{m=1}^\infty$. Moreover, by Parseval's theorem, we have the equivalence $\|f^*\|_2^2 = \sum_{m=1}^\infty (\theta_m^*)^2 \leq 1$, and a straightforward calculation yields that

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

Moreover, this infimum is achieved by the truncated function $\widetilde{f}_T = \sum_{m=1}^T \theta_m^* \phi_m$; see Exercise 13.10 for more details.

On the other hand, since the estimator over $\mathscr{F}_{\text{ortho}}(1;T)$ corresponds to a form of ridge regression in dimension T, the calculations from Example 13.8 imply that the critical equation (13.42a) is satisfied by $\delta_n^2 \simeq \sigma^2 \frac{T}{n}$. Setting $f = \widetilde{f}_T$ in the oracle inequality (13.43b) and then taking expectations over the covariates $\mathbf{X} = \{x_i\}_{i=1}^n$ yields that the least-squares estimate $\widehat{f}$ over $\mathscr{F}_{\text{ortho}}(1;T)$ satisfies the bound

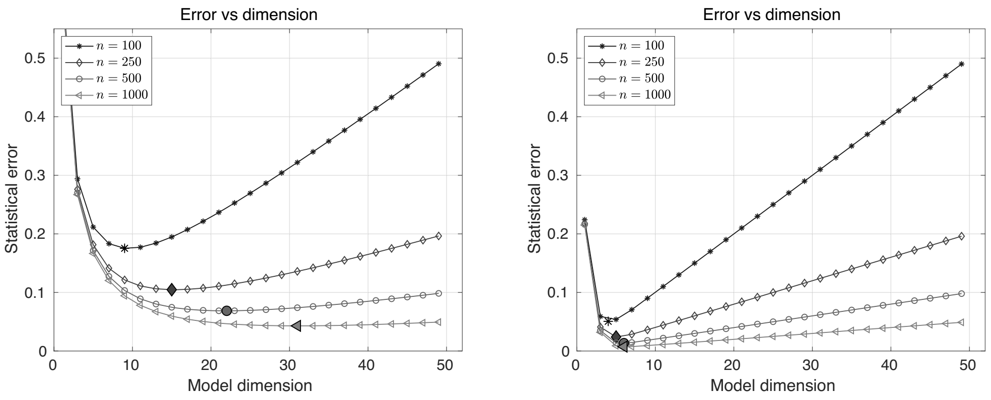
$$\mathbb{E}_{\mathbf{X},w}\big[\|\widehat{f} - f^*\|_n^2\big] \precsim \sum_{m=T+1}^{\infty} (\theta_m^*)^2 + \sigma^2 \frac{T}{n}. \tag{13.45}$$

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

The optimal choice of T will depend on the rate at which the basis coefficients $(\theta_m^*)_{m=1}^\infty$ decay to zero. For example, suppose that they exhibit a polynomial decay, say $|\theta_m^*| \leq Cm^{-\alpha}$ for some $\alpha > 1/2$. In Example 13.15 to follow, we provide a concrete instance of such polynomial decay using Fourier coefficients and α-times-differentiable functions. Figure 13.4(a) shows a plot of the upper bound (13.45) as a function of T, with one curve for each of 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 tradeoff 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. ♣

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press



(a) (b)

Figure 13.4 Plot of upper bound (13.45) 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 $n \in \{100, 250, 500, 1000\}$. (a) Polynomial decaying coefficients $|\theta^*_m| \leq m^{-1}$. (b) Exponential decaying coefficients $|\theta^*_m| \leq e^{-m/2}$.

Example 13.15 (Fourier bases and differentiable functions) Define the constant function $\phi_0(x) = 1$ for all $x \in [0, 1]$, and the sinusoidal functions

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

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

$$f^* = \theta^*_0 + \sum_{m=1}^{\infty} \{\theta^*_m \phi_m + \widetilde{\theta}^*_m \widetilde{\phi}_m\}.$$

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

$$\mathscr{G}(1; M) = \Big\{\beta_0 + \sum_{m=1}^{M} (\beta_m \phi_m + \widetilde{\beta}_m \widetilde{\phi}_m) \mid \beta_0^2 + \sum_{m=1}^{M} (\beta_m^2 + \widetilde{\beta}_m^2) \leq 1\Big\}. \tag{13.46}$$

Note that this is simply a re-indexing of a function class $\mathscr{F}_{\text{ortho}}(1; T)$ of the form (13.44) with $T = 2M + 1$.

Now suppose that for some integer $\alpha \geq 1$, the target function f^* is α-times differentiable, and suppose that $\int_0^1 [(f^*)^{(\alpha)}(x)]^2\, dx \leq R$ for some radius R. It can be verified that there is a constant c such that $(\beta^*_m)^2 + (\widetilde{\beta}^*_m)^2 \leq \frac{c}{m^{2\alpha}}$ for all $m \geq 1$, and, moreover, we can find a function $f \in \mathscr{G}(1; M)$ such that

$$\|f - f^*\|_2^2 \leq \frac{c' R}{M^{2\alpha}}. \tag{13.47}$$

See Exercise 13.11 for details on these properties.

Putting together the pieces, the bound (13.45) combined with the approximation-theoretic guarantee (13.47) implies that the least-squares estimate $\widehat{f}_M$ over $\mathscr{G}(1; M)$ satisfies the bound

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

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

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

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. ♣

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

Example 13.16 (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 $s \in \{1, 2, \ldots, d\}$, consider the class of all linear regression functions based on s-sparse vectors—namely, the class

$$\mathscr{F}_{\text{spar}}(s) := \{f_\theta \mid \theta \in \mathbb{R}^d, \; \|\theta\|_0 \leq s\},$$

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

Disregarding computational considerations, a natural estimator is given by

$$\widehat{\theta} \in \arg \min_{\theta \in \mathscr{F}_{\text{spar}}(s)} \|y - \mathbf{X}\theta\|_n^2, \tag{13.48}$$

corresponding to performing least squares over the set of all regression vectors with at most s non-zero coefficients. As a corollary of Theorem 13.13, we claim that the $L^2(\mathbb{P}_n)$-error of this estimator is upper bounded as

$$\|f_{\widehat{\theta}} - f_{\theta^*}\|_n^2 \precsim \inf_{\theta \in \mathscr{F}_{\text{spar}}(s)} \|f_\theta - f_{\theta^*}\|_n^2 + \underbrace{\sigma^2 \frac{s \log(\frac{ed}{s})}{n}}_{\delta_n^2} \tag{13.49}$$

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

In order to derive this result as a corollary of Theorem 13.13, we need to compute the local Gaussian complexity (13.42a) for our function class. Making note of the inclusion $\partial\mathscr{F}_{\text{spar}}(s) \subset \mathscr{F}_{\text{spar}}(2s)$, we have $\mathcal{G}_n(\delta; \partial\mathscr{F}_{\text{spar}}(s)) \leq \mathcal{G}_n(\delta; \mathscr{F}_{\text{spar}}(2s))$. Now let $S \subset \{1, 2, \ldots, d\}$ be an arbitrary $2s$-sized subset of indices, and let $\mathbf{X}_S \in \mathbb{R}^{n \times 2s}$ denote the submatrix with

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

columns indexed by S. We can then write

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

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 2.26 implies the tail bound

$$\mathbb{P}[Z_n(S) \ge \mathbb{E}[Z_n(S)] + t\delta] \le e^{-\frac{nt^2}{2}} \qquad \text{for all } t > 0. \tag{13.50}$$

We now upper bound the expectation. Consider the singular value decomposition $\mathbf{X}_S = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, 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 $\|\mathbf{X}_S \theta_S\|_2 = \|\mathbf{D}\mathbf{V}^{\mathrm{T}} \theta_S\|_2$, we arrive at the upper bound

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

Since $w \sim \mathcal{N}(0, \mathbf{I}_n)$ and the matrix $\mathbf{U}$ has orthonormal columns, we have $\mathbf{U}^{\mathrm{T}} w \sim \mathcal{N}(0, \mathbf{I}_{2s})$, and therefore $\mathbb{E}\|\mathbf{U}^{\mathrm{T}} w\|_2 \le \sqrt{2s}$. Combining this upper bound with the earlier tail bound (13.50), an application of the union bound yields

$$\mathbb{P}\Bigg[\max_{|S|=2s} Z_n(S) \ge \delta\Big(\sqrt{\frac{2s}{n}} + t\Big)\Bigg] \le \binom{d}{2s} e^{-\frac{nt^2}{2}}, \qquad \text{valid for all } t \ge 0.$$

By integrating this tail bound, we find that

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

so that the critical inequality (13.17) is satisfied for $\delta_n^2 \simeq \sigma^2 \frac{s \log(ed/s)}{n}$, as claimed. ♣

13.3.2 Proof of Theorem 13.13

We now turn to the proof of our oracle inequality; it is a relatively straightforward extension of the proof of Theorem 13.5. Given an arbitrary $\widetilde{f} \in \mathscr{F}$, since it is feasible and $\widehat{f}$ is optimal, we have

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - \widehat{f}(x_i))^2 \le \frac{1}{2n} \sum_{i=1}^{n} (y_i - \widetilde{f}(x_i))^2.$$

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

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

where we have defined $\widehat{\Delta} := \widehat{f} - f^*$ and $\widetilde{\Delta} = \widehat{f} - \widetilde{f}$.

It remains to analyze the term on the right-hand side involving $\widetilde{\Delta}$. We break our analysis

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

into two cases.

Case 1: First suppose that $\|\widetilde{\Delta}\|_n \leq \sqrt{t\delta_n}$. We then have

$$\begin{aligned}\|\widehat{\Delta}\|_n^2 = \|\widehat{f} - f^*\|_n^2 &= \|(\widetilde{f} - f^*) + \widetilde{\Delta}\|_n^2 \\ &\overset{\text{(i)}}{\leq} \Big\{\|\widetilde{f} - f^*\|_n + \sqrt{t\delta_n}\Big\}^2 \\ &\overset{\text{(ii)}}{\leq} (1 + 2\beta)\|\widetilde{f} - f^*\|_n^2 + (1 + \frac{2}{\beta})t\delta_n,\end{aligned}$$

where step (i) follows from the triangle inequality, and step (ii) is valid for any $\beta > 0$, using the Fenchel–Young inequality. Now setting $\beta = \frac{\gamma}{1-\gamma}$ for some $\gamma \in (0, 1)$, observe that $1 + 2\beta = \frac{1+\gamma}{1-\gamma}$, and $1 + \frac{2}{\beta} = \frac{2-\gamma}{\gamma} \leq \frac{2}{\gamma(1-\gamma)}$, so that the stated claim (13.42b) follows.

Case 2: Otherwise, we may assume that $\|\widetilde{\Delta}\|_n > \sqrt{t\delta_n}$. Noting that the function $\widetilde{\Delta}$ belongs to the difference class $\partial\mathscr{F} := \mathscr{F} - \mathscr{F}$, we then apply Lemma 13.12 with $u = \sqrt{t\delta_n}$ and $\mathscr{H} = \partial\mathscr{F}$. Doing so yields that

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

Combining with the basic inequality (13.51), we find that, with probability at least $1-2e^{-\frac{nt\delta_n}{2\sigma^2}}$, the squared error is bounded as

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

where the second step follows from the triangle inequality. Applying the Fenchel–Young inequality with parameter $\beta > 0$, we find that

$$4\sqrt{t\delta_n}\,\|\widehat{\Delta}\|_n \leq 4\beta\|\widehat{\Delta}\|_n^2 + \frac{4}{\beta}t\delta_n$$

and

$$4\sqrt{t\delta_n}\|\widetilde{f} - f^*\|_n \leq 4\beta\|\widetilde{f} - f^*\|_n^2 + \frac{4}{\beta}t\delta_n.$$

Combining the pieces yields

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

For all $\beta \in (0, 1/4)$, rearranging yields the bound

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

Setting $\gamma = 4\beta$ yields the claim.

13.4 Regularized estimators

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 analogs.

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

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

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 $\mathscr{F}$.

13.4.1 Oracle inequalities for regularized estimators

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

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

where $w_i \sim \mathcal{N}(0,1)$ are i.i.d. variates. When the function class $\mathscr{F}$ and rescaled ball $\mathbb{B}_{\partial\mathscr{F}}(3) = \{g \in \partial\mathscr{F} \mid \|g\|_{\mathscr{F}} \leq 3\}$ are clear from the context, we adopt $\mathcal{G}_n(\delta)$ as a convenient shorthand. For a user-defined radius $R > 0$, we let $\delta_n > 0$ be any number satisfying the inequality

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

Theorem 13.17 *Given the previously described observation model and a convex function class $\mathscr{F}$, suppose that we solve the convex program* (13.52) *with some regularization parameter $\lambda_n \geq 2\delta_n^2$. Then there are universal positive constants (c_j, c'_j) such that*

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

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

$$\mathbb{E}\|\widehat{f} - f^*\|_n^2 \leq c'_0 \inf_{\|f\|_{\mathscr{F}} \leq R} \|f - f^*\|_n^2 + c'_1\, R^2\, \{\delta_n^2 + \lambda_n\}. \tag{13.55b}$$

We return to prove this claim in Section 13.4.4.

13.4.2 Consequences for kernel ridge regression

Recall from Chapter 12 our discussion of the kernel ridge regression estimate (12.28). There we showed that this KRR estimate has attractive computational properties, in that it only requires computing the empirical kernel matrix, and then solving a linear system (see Proposition 12.33). Here we turn to the complementary question of understanding its statistical behavior. Since it is a special case of the general estimator (13.52), Theorem 13.17 can be used to derive upper bounds on the prediction error. Interestingly, these bounds have a very intuitive interpretation, one involving the eigenvalues of the empirical kernel matrix.

From our earlier definition, the (rescaled) empirical kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite, 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 $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n \geq 0$. The following corollary of Theorem 13.17 provides bounds on the performance of the kernel ridge regression estimate in terms of these eigenvalues:

Corollary 13.18 *For the KRR estimate* (12.28)*, the bounds of Theorem 13.17 hold for any $\delta_n > 0$ satisfying the inequality*

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

We provide the proof in Section 13.4.3. Before doing so, let us examine the implications of Corollary 13.18 for some specific choices of kernels.

Example 13.19 (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 $\min\{n, m\}$. Therefore, for any sample size n larger than 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 (13.56) is satisfied for all $\delta \gtrsim \frac{\sigma}{R}\sqrt{\frac{m}{n}}$, so that the KRR estimate satisfies the bound

$$\|\widehat{f} - f^*\|_n^2 \precsim \inf_{\|f\|_{\mathbb{H}} \leq R} \|f - f^*\|_n^2 + \sigma^2 \frac{m}{n},$$

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, 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 13.20 (First-order Sobolev space) Previously, we introduced the kernel function $\mathcal{K}(x, z) = \min\{x, z\}$ defined on the unit square $[0, 1] \times [0, 1]$. As discussed in Example 12.16, the associated RKHS corresponds to a first-order Sobolev space

$$\mathbb{H}^1[0,1] := \Big\{f\colon [0,1] \to \mathbb{R} \mid f(0) = 0,\ \text{and } f \text{ is abs. cts. with } f' \in L^2[0,1]\Big\}.$$

As shown in Example 12.23, the kernel integral operator associated with this space has the eigendecomposition

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

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.[3]

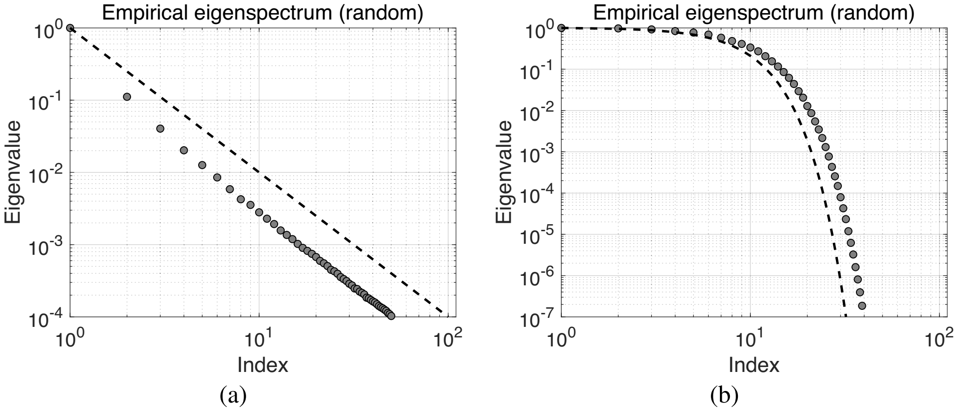


Figure 13.5 Log–log behavior of the eigenspectrum of the empirical kernel matrix based on $n = 2000$ samples drawn i.i.d. from the uniform distribution over the interval $\mathcal{X}$ for two different kernel functions. The plotted circles correspond to empirical eigenvalues, whereas the dashed line 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 $\mathcal{X} = [-1, 1]$.

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

[3] In particular, Proposition 14.25 shows that the critical radii computed using the population and empirical kernel eigenvalues are equivalent up to constant factors.

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

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

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

Consequently, the critical inequality (13.56) is satisfied by $\delta_n^{3/2} \simeq \frac{\sigma}{R\sqrt{n}}$, or equivalently $\delta_n^2 \simeq (\frac{\sigma^2}{R^2}\ \frac{1}{n})^{2/3}$. Putting together the pieces, Corollary 13.18 implies that the KRR estimate will satisfy the upper bound

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

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 13.21 (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 12.25, the eigenvalues of the associated kernel operator scale as $\mu_j \simeq e^{-cj \log j}$ as $j \to +\infty$. Accordingly, let us adopt the heuristic that the empirical eigenvalues satisfy a bound of the form $\hat{\mu}_j \leq c_0 e^{-c_1 j \log j}$. 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 14.25 in Chapter 14 allows us to make a rigorous argument that reaches the conclusion 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 nonparametric regression over the Gaussian kernel class satisfies the bound

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

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 13.20. 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. ♣

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

13.4.3 Proof of Corollary 13.18

The proof of this corollary is based on a bound on the local Gaussian complexity (13.53) of the unit ball of an RKHS. Since it is of independent interest, let us state it as a separate result:

Lemma 13.22 *Consider an RKHS with kernel function $\mathcal{K}$. For a given set of design points $\{x_i\}_{i=1}^n$, let $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n \geq 0$ be the eigenvalues of the normalized kernel matrix* **K** *with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$. Then for all $\delta > 0$, we have*

$$\mathbb{E}\left[\sup_{\substack{\|f\|_{\mathbb{H}} \leq 1 \\ \|f\|_n \leq \delta}} \left|\frac{1}{n}\sum_{i=1}^n w_i f(x_i)\right|\right] \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^n \min\{\delta^2, \hat{\mu}_j\}}, \tag{13.57}$$

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

$$g(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i), \tag{13.58}$$

some vector of coefficients $\alpha \in \mathbb{R}^n$. Indeed, as argued in our proof of Proposition 12.33, 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 (13.58). 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 **K**, the constraint $\|g\|_n \leq \delta$ is equivalent to $\|\mathbf{K}\alpha\|_2 \leq \delta$, whereas the inequality $\|g\|_{\mathbb{H}}^2 \leq 1$ corresponds to $\|g\|_{\mathbb{H}}^2 = \alpha^{\mathrm{T}}\mathbf{K}\alpha \leq 1$. Thus, we can write the local Gaussian complexity as an optimization problem in the vector $\alpha \in \mathbb{R}^n$ with a linear cost function and quadratic constraints—namely,

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

Since the kernel matrix **K** is symmetric and positive semidefinite, it has an eigendecomposition[4] of the form $\mathbf{K} = \mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{U}$, where **U** is orthogonal and $\mathbf{\Lambda}$ is diagonal with entries $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n > 0$. If we then define the transformed vector $\beta = \mathbf{K}\alpha$, we find (following some algebra) that the complexity can be written as

$$\mathcal{G}_n(\delta) = \frac{1}{\sqrt{n}}\mathbb{E}_w[\sup_{\beta \in \mathcal{D}} |w^{\mathrm{T}}\beta|], \quad \text{where} \quad \mathcal{D} := \{\beta \in \mathbb{R}^n \mid \|\beta\|_2^2 \leq \delta^2, \sum_{j=1}^n \frac{\beta_j^2}{\hat{\mu}_j} \leq 1\}$$

[4] In this argument, so as to avoid potential division by zero, we assume that **K** has strictly positive eigenvalues; otherwise, we can simply repeat the argument given here while restricting the relevant summations to positive eigenvalues.

is the intersection of two ellipses. Now define the ellipse

$$\mathcal{E} := \Big\{\beta \in \mathbb{R}^n \mid \sum_{j=1}^n \eta_j \beta_j^2 \leq 2\Big\}, \qquad \text{where } \eta_j = \max\{\delta^{-2}, \hat{\mu}_j^{-1}\}.$$

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

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

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

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

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 $(\eta_j)^{-1} = (\max\{\delta^{-2}, \hat{\mu}_j^{-1}\})^{-1} = \min\{\delta^2, \hat{\mu}_j\}$ yields the claim. □

13.4.4 Proof of Theorem 13.17

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 $\inf_{\|f\|_{\mathscr{F}}\leq 1} \|f - f^*\|_n^2$. Our final mean-squared error then should be multiplied by R^2 so as to obtain a result for the original problem.

In order to keep the notation streamlined, we introduce the shorthand $\tilde{\sigma} = \sigma/R$. Let $\widetilde{f}$ be any element of $\mathscr{F}$ such that $\|\widetilde{f}\|_{\mathscr{F}} \leq 1$. At the end of the proof, we optimize this choice. Since $\widehat{f}$ and $\widetilde{f}$ are optimal and feasible (respectively) for the program (13.52), we have

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

Defining the errors $\widehat{\Delta} = \widehat{f} - f^*$ and $\widetilde{\Delta} = \widehat{f} - \widetilde{f}$ and recalling that $y_i = f^*(x_i) + \tilde{\sigma} w_i$, performing some algebra yields the *modified basic inequality*

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

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

Since $\|\widetilde{f}\|_{\mathscr{F}} \leq 1$ by assumption, we certainly have the possibly weaker bound

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \frac{1}{2}\|\widetilde{f} - f^*\|_n^2 + \frac{\tilde{\sigma}}{n}\Big|\sum_{i=1}^n w_i\widetilde{\Delta}(x_i)\Big| + \lambda_n. \tag{13.60}$$

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Consequently, if $\|\widetilde{\Delta}\|_n \leq \sqrt{t\delta_n}$, we can then follow the same argument as in the proof of Theorem 13.13, thereby establishing the bound (along with the extra term λ_n from our modified basic inequality).

Otherwise, we may assume that $\|\widetilde{\Delta}\|_n > \sqrt{t\delta_n}$, and we do so throughout the remainder of the proof. We now split the argument into two cases.

Case 1: First, suppose that $\|\widehat{f}\|_{\mathscr{F}} \leq 2$. The bound $\|\widetilde{f}\|_{\mathscr{F}} \leq 1$ together with the inequality $\|\widehat{f}\|_{\mathscr{F}} \leq 2$ implies that $\|\widetilde{\Delta}\|_{\mathscr{F}} \leq 3$. Consequently, by applying Lemma 13.12 over the set of functions $\{g \in \partial\mathscr{F} \mid \|g\|_{\mathscr{F}} \leq 3\}$, we conclude that

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

By the triangle inequality, we have

$$\begin{aligned} 2\sqrt{t\delta_n}\|\widetilde{\Delta}\|_n &\leq 2\sqrt{t\delta_n}\|\widehat{\Delta}\|_n + 2\sqrt{t\delta_n}\|\widetilde{f} - f^*\|_n \\ &\leq 2\sqrt{t\delta_n}\|\widehat{\Delta}\|_n + 2t\delta_n + \frac{\|\widetilde{f} - f^*\|_n^2}{2}, \end{aligned} \tag{13.61}$$

where the second step uses the Fenchel–Young inequality. Substituting these upper bounds into the basic inequality (13.60), we find that

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

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

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

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

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

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

where we again use the bound $\|\widetilde{f}\|_{\mathscr{F}} \leq 1$ in the final step.

Substituting this upper bound into our modified basic inequality (13.59) yields the upper bound

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

Our next step is to upper bound the stochastic component in the inequality (13.62).

Lemma 13.23 *There are universal positive constants (c_1, c_2) such that, with probability greater than $1 - c_1 e^{-\frac{n\delta_n^2}{c_2\tilde{\sigma}^2}}$, we have*

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

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

We now complete the proof of the theorem using this lemma. We begin by observing that, since $\|\widetilde{f}\|_{\mathscr{F}} \leq 1$ and $\|\widehat{f}\|_{\mathscr{F}} > 2$, the triangle inequality implies that $\|\widetilde{\Delta}\|_{\mathscr{F}} \geq \|\widehat{f}\|_{\mathscr{F}} - \|\widetilde{f}\|_{\mathscr{F}} > 1$, so that Lemma 13.23 may be applied. Substituting the upper bound (13.63) into the inequality (13.62) yields

$$\begin{aligned}\frac{1}{2}\|\widehat{\Delta}\|_n^2 &\leq \frac{1}{2}\|\widetilde{f} - f^*\|_n^2 + 2\delta_n\|\widetilde{\Delta}\|_n + \{2\delta_n^2 - \lambda_n\}\|\widetilde{\Delta}\|_{\mathscr{F}} + 2\lambda_n + \frac{\|\widetilde{\Delta}\|_n^2}{16}\\ &\leq \frac{1}{2}\|\widetilde{f} - f^*\|_n^2 + 2\delta_n\|\widetilde{\Delta}\|_n + 2\lambda_n + \frac{\|\widetilde{\Delta}\|_n^2}{16},\end{aligned} \tag{13.64}$$

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 $\widetilde{\Delta}$ into quantities involving $\widehat{\Delta}$: in particular, by the triangle inequality, we have $\|\widetilde{\Delta}\|_n \leq \|\widetilde{f} - f^*\|_n + \|\widehat{\Delta}\|_n$. Thus, we have

$$2\delta_n\|\widetilde{\Delta}\|_n \leq 2\delta_n\|\widetilde{f} - f^*\|_n + 2\delta_n\|\widehat{\Delta}\|_n, \tag{13.65a}$$

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

$$\frac{\|\widetilde{\Delta}\|_n^2}{16} \leq \frac{1}{8}\Big\{\|\widetilde{f} - f^*\|_n^2 + \|\widehat{\Delta}\|_n^2\Big\}. \tag{13.65b}$$

Substituting inequalities (13.65a) and (13.65b) into the earlier bound (13.64) and performing some algebra yields

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

The claim (13.55a) follows by applying the quadratic formula to this inequality.

It remains to prove Lemma 13.23. We claim that it suffices to prove the bound (13.63) for functions $g \in \partial\mathscr{F}$ such that $\|g\|_{\mathscr{F}} = 1$. Indeed, suppose that it holds for all such functions, and that we are given a function Δ with $\|\Delta\|_{\mathscr{F}} > 1$. By assumption, we can apply the inequality (13.63) to the new function $g := \Delta/\|\Delta\|_{\mathscr{F}}$, which belongs to $\partial\mathscr{F}$ by the star-shaped assumption. Applying the bound (13.63) to g and then multiplying both sides by $\|\Delta\|_{\mathscr{F}}$, we obtain

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

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

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

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

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

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

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

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

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

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

where inequality (i) follows from Lemma 13.6, 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 (13.66) yields

$$\mathbb{P}\left[Z_n(t) \geq t\delta_n + \frac{t^2}{32}\right] \leq e^{-c_2\frac{nt^2}{\tilde{\sigma}^2}} \qquad \text{for each } t > \delta_n. \tag{13.67b}$$

We are now equipped to complete the proof by a "peeling" argument. Let $\mathcal{E}$ denote the event that the bound (13.63) is violated for some function $g \in \partial\mathscr{F}$ with $\|g\|_{\mathscr{F}} = 1$. 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 $\|g\|_{\mathscr{F}} = 1$. For $m = 0, 1, 2, \ldots$, define $t_m = 2^m\delta_n$. We then have the decomposition $\mathcal{E} = \mathcal{E}(0, t_0) \cup \left(\bigcup_{m=0}^{\infty} \mathcal{E}(t_m, t_{m+1})\right)$ and hence, by the union bound,

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

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

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

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

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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

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

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

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

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

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

13.5 Bibliographic details and background

Nonparametric regression is a classical problem in statistics with a lengthy and rich history. Although this chapter is limited to the method of nonparametric least squares, there are a variety of other cost functions that can be used for regression, which might be preferable for reasons of robustness. The techniques described this chapter are relevant for analyzing any such M-estimator—that is, any method based on minimizing or maximizing some criterion of fit. In addition, nonparametric regression can be tackled via methods that are not most naturally viewed as M-estimators, including orthogonal function expansions, local polynomial representations, kernel density estimators, nearest-neighbor methods and scatterplot smoothing methods, among others. We refer the reader to the books (Gyorfi et al., 2002; Härdle et al., 2004; Wasserman, 2006; Eggermont and LaRiccia, 2007; Tsybakov, 2009) and references therein for further background on these and other methods.

An extremely important idea in this chapter was the use of localized forms of Gaussian or Rademacher complexity, as opposed to the global forms studied in Chapter 4. These localized complexity measures are needed in order to obtain optimal rates for nonparametric estimation problems. 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 function complexities of the form given in Corollary 13.7 are used extensively by van de Geer (2000), whereas other authors have studied localized forms of the Rademacher and Gaussian complexities (Koltchinskii, 2001, 2006; Bartlett et al., 2005). The bound on the localized Rademacher complexity of reproducing kernel Hilbert spaces, as stated in Lemma 13.22, is due to Mendelson (2002); see also the paper by Bartlett and Mendelson (2002) for related results. The peeling technique used in the proof of Lemma 13.23 is widely used in empirical process theory (Alexander, 1987; van de Geer, 2000).

The ridge regression estimator from Examples 13.1 and 13.8 was introduced by Hoerl and Kennard (1970). The Lasso estimator from Example 13.1 is treated in detail in Chapter 7. The cubic spline estimator from Example 13.2, as well as the kernel ridge regression estimator from Example 13.3, are standard methods; see Chapter 12 as well as the books (Wahba, 1990; Gu, 2002) for more details. The ℓ_q-ball constrained estimators from Examples 13.1 and 13.9 were analyzed by Raskutti et al. (2011), who also used information-theoretic methods, 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

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

Pajor (1988), as well as Guédon and Litvak (2000); see also the arguments given by Raskutti et al. (2011) for details on the specific claims given here.

The problems of convex and/or monotonic regression from Example 13.4 are particular examples of what is known as shape-constrained estimation. It has been the focus of classical work (Hildreth, 1954; Brunk, 1955, 1970; Hanson and Pledger, 1976), as well as much recent and on-going work (e.g., Balabdaoui et al., 2009; Cule et al., 2010; Dümbgen et al., 2011; Seijo and Sen, 2011; Chatterjee et al., 2015), especially in the multivariate setting. The books (Rockafellar, 1970; Hiriart-Urruty and Lemaréchal, 1993; Borwein and Lewis, 1999; Bertsekas, 2003; Boyd and Vandenberghe, 2004) contain further information on subgradients and other aspects of convex analysis. The bound (13.34) on the sup-norm (L_∞) metric entropy for bounded convex Lipschitz functions is due to Bronshtein (1976); see also Section 8.4 of Dudley (1999) for more details. On the other hand, the class of all convex functions $f\colon [0,1] \to [0,1]$ without any Lipschitz constraint is *not* totally bounded in the sup-norm metric; see Exercise 5.1 for details. Guntuboyina and Sen (2013) provide bounds on the entropy in the L_p-metrics over the range $p \in [1,\infty)$ for convex functions without the Lipschitz condition.

Stone (1985) introduced the class of additive nonparametric regression models discussed in Exercise 13.9, and subsequent work has explored many extensions and variants of these models (e.g., Hastie and Tibshirani, 1986; Buja et al., 1989; Meier et al., 2009; Ravikumar et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012). Exercise 13.9 in this chapter and Exercise 14.8 in Chapter 14 explore some properties of the standard additive model.

13.6 Exercises

Exercise 13.1 (Characterization of the Bayes least-squares estimate)

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

(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.)

(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).

Exercise 13.2 (Prediction error in linear regression) Recall the linear regression model from Example 13.8 with fixed design. Show via a direct argument that

$$\mathbb{E}[\|f_{\widehat{\theta}} - f_{\theta^*}\|_n^2] \le \sigma^2 \frac{\text{rank}(\mathbf{X})}{n},$$

valid for any observation noise that is zero-mean with variance σ^2.

Exercise 13.3 (Cubic smoothing splines) Recall the cubic spline estimate (13.10) from Example 13.2, as well as the kernel function $\mathcal{K}(x, z) = \int_0^1 (x-y)_+ \, (z-y)_+ \, dy$ from Example 12.29.

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

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

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

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

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

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

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

Exercise 13.4 (Star-shaped sets and convexity) In this exercise, we explore some properties of star-shaped sets.

(a) Show that a set C is star-shaped around one of its points x^* if and only if the point $\alpha x + (1-\alpha)x^*$ belongs to C for any $x \in C$ and any $\alpha \in [0, 1]$.

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

Exercise 13.5 (Lower bounds on the critical inequality) Consider the critical inequality (13.17) in the case $f^* = 0$, so that $\mathscr{F}^* = \mathscr{F}$.

(a) Show that the critical inequality (13.17) is always satisfied for $\delta^2 = 4\sigma^2$.

(b) Suppose that a convex function class $\mathscr{F}$ contains the constant function $f \equiv 1$. Show that any $\delta \in (0, 1]$ satisfying the critical inequality (13.17) must be lower bounded as $\delta^2 \geq \min\{1, \frac{8}{\pi}\frac{\sigma^2}{n}\}$.

Exercise 13.6 (Local Gaussian complexity and adaptivity) This exercise illustrates how, even for a fixed base function class, the local Gaussian complexity $\mathcal{G}_n(\delta; \mathscr{F}^*)$ 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 $\mathscr{F}_{\ell_1}(1) = \{f_\theta \mid \|\theta\|_1 \leq 1\}$. Suppose that we observe samples of the form

$$y_i = f_{\theta^*}(e_i) + \frac{\sigma}{\sqrt{n}} w_i \;=\; \theta_i^* + \frac{\sigma}{\sqrt{n}} w_i,$$

where $w_i \sim \mathcal{N}(0, 1)$ is an i.i.d. noise sequence. Let us analyze the performance of the ℓ_1-constrained least-squares estimator

$$\widehat{\theta} = \arg \min_{f_\theta \in \mathscr{F}_{\ell_1}(1)} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f_\theta(e_i))^2 \right\} \;=\; \arg \min_{\substack{\theta\in\mathbb{R}^d \\ \|\theta\|_1 \leq 1}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \theta_i)^2 \right\}.$$

(a) For any $f_{\theta^*} \in \mathscr{F}_{\ell_1}(1)$, show that $\mathcal{G}_n(\delta; \mathscr{F}^*_{\ell_1}(1)) \leq c_1 \sqrt{\frac{\log n}{n}}$ for some universal constant c_1, and hence that $\|\widehat{\theta} - \theta^*\|_2^2 \leq c_1' \sigma \sqrt{\frac{\log n}{n}}$ with high probability.

(b) Now consider some f_{θ^*} with $\theta^* \in \{e_1, \ldots, e_n\}$—that is, one of the canonical basis vectors. Show that there is a universal constant c_2 such that the local Gaussian complexity is bounded as $\mathcal{G}_n(\delta; \mathscr{F}^*_{\ell_1}(1)) \leq c_2\, \delta \frac{\sqrt{\log n}}{n}$, and hence that $\|\widehat{\theta} - \theta^*\|_2^2 \leq c_2' \frac{\sigma^2 \log n}{n}$ with high probability.

Exercise 13.7 (Rates for polynomial regression) Consider the class of all $(m-1)$-degree polynomials

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

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

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

Exercise 13.8 (Rates for twice-differentiable functions) Consider the function class $\mathscr{F}$ of functions $f \colon [0,1] \to \mathbb{R}$ 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

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

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

Exercise 13.9 (Rates for additive nonparametric models) Given a convex and symmetric class $\mathscr{G}$ of univariate functions $g \colon \mathbb{R} \to \mathbb{R}$ equipped with a norm $\|\cdot\|_{\mathscr{G}}$, consider the class of additive functions over $\mathbb{R}^d$, namely

$$\mathscr{F}_{\text{add}} = \{f \colon \mathbb{R}^d \to \mathbb{R} \mid f = \sum_{j=1}^d g_j \quad \text{for some } g_j \in \mathscr{G} \text{ with } \|g_j\|_{\mathscr{G}} \leq 1\}. \tag{13.70}$$

Suppose that we have n i.i.d. samples of the form $y_i = f^*(x_i) + \sigma w_i$, where each $x_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$, $w_i \sim \mathcal{N}(0,1)$, and $f^* := \sum_{j=1}^d g_j^*$ is some function in $\mathscr{F}_{\text{add}}$, and that we estimate f^* by the constrained least-squares estimate

$$\widehat{f} := \arg\min_{f \in \mathscr{F}_{\text{add}}} \Big\{\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2\Big\}.$$

For each $j = 1, \ldots, d$, define the jth-coordinate Gaussian complexity

$$\mathcal{G}_{n,j}(\delta; 2\mathscr{G}) = \mathbb{E}\Big[\sup_{\substack{\|g_j\|_{\mathscr{G}} \leq 2 \\ \|g_j\|_n \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^n w_i g_j(x_{ij})\Big|\Big],$$

and let $\delta_{n,j} > 0$ be the smallest positive solution to the inequality $\frac{\mathcal{G}_{n,j}(\delta; 2\mathscr{G})}{\delta} \leq \frac{\delta}{2\sigma}$.

(a) Defining $\delta_{n,\max} = \max_{j=1,\ldots,d} \delta_{n,j}$, show that, for each $t \geq \delta_{n,\max}$, we have

$$\frac{\sigma}{n}\Big|\sum_{i=1}^n w_i \widehat{\Delta}(x_i)\Big| \leq d t \delta_{n,\max} + 2\sqrt{t\delta_{n,\max}}\Big(\sum_{j=1}^d \|\widehat{\Delta}_j\|_n\Big)$$

with probability at least $1 - c_1 d e^{-c_2 n t \delta_{n,\max}}$. (Note that $\widehat{f} = \sum_{j=1}^d \widehat{g}_j$ for some $\widehat{g}_j \in \mathscr{G}$, so that the function $\widehat{\Delta}_j = \widehat{g}_j - g_j^*$ corresponds to the error in coordinate j, and $\widehat{\Delta} := \sum_{j=1}^d \widehat{\Delta}_j$ is the full error function.)

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press

(b) Suppose that there is a universal constant $K \geq 1$ such that

$$\sqrt{\sum_{j=1}^{n} \|g_j\|_n^2} \leq \sqrt{K}\, \| \sum_{j=1}^{d} g_j \|_n \qquad \text{for all } g_j \in \mathscr{G}.$$

Use this bound and part (a) to show that $\|\widehat{f} - f^*\|_n^2] \leq \; c_3\, K\, d\, \delta_{n,\max}^2$ with high probability.

Exercise 13.10 (Orthogonal series expansions) Recall the function class $\mathscr{F}_{\text{ortho}}(1; T)$ from Example 13.14 defined by orthogonal series expansion with T coefficients.

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

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

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

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

Exercise 13.11 (Differentiable functions and Fourier coefficients) For a given integer $\alpha \geq 1$ and radius $R > 0$, consider the class of functions $\mathscr{F}_\alpha(R) \subset L^2[0, 1]$ such that:

- The function f is α-times differentiable, with $\int_0^1 (f^{(\alpha)}(x))^2\, dx \leq R$.
- It and its derivatives satisfy the boundary conditions $f^{(j)}(0) = f^{(j)}(1) = 0$ for all $j = 0, 1, \ldots, \alpha$.

(a) For a function $f \in \mathscr{F}_\alpha(R) \cap \{\|f\|_2 \leq 1\}$, let $\{\beta_0, (\beta_m, \beta_m)_{m=1}^{\infty}\}$ be its Fourier coefficients as previously defined in Example 13.15. Show that there is a constant c such that $\beta_m^2 + \widetilde{\beta}_m^2 \leq \frac{cR}{m^{2\alpha}}$ for all $m \geq 1$.

(b) Verify the approximation-theoretic guarantee (13.47).

https://doi.org/10.1017/9781108627771.013 Published online by Cambridge University Press