14

Localization and uniform laws

As discussed previously in Chapter 4, uniform laws of large numbers concern the deviations between sample and population averages, when measured in a uniform sense over a given function class. The classical forms of uniform laws are asymptotic in nature, guaranteeing that the deviations converge to zero in probability or almost surely. The more modern approach is to provide non-asymptotic guarantees that hold for all sample sizes, and provide sharp rates of convergence. In order to achieve the latter goal, an important step is to localize the deviations to a small neighborhood of the origin. We have already encountered a form of localization in our discussion of nonparametric regression from Chapter 13. In this chapter, we turn to a more in-depth study of this technique and its use in establishing sharp uniform laws for various types of processes.

14.1 Population and empirical L^2-norms

We begin our exploration with a detailed study of the relation between the population and empirical L^2-norms. Given a function $f : \mathcal{X} \to \mathbb{R}$ and a probability distribution $\mathbb{P}$ over $\mathcal{X}$, the usual $L^2(\mathbb{P})$-norm is given by

$$\|f\|_{L^2(\mathbb{P})}^2 := \int_{\mathcal{X}} f^2(x)\mathbb{P}(dx) = \mathbb{E}[f^2(X)], \tag{14.1}$$

and we say that $f \in L^2(\mathbb{P})$ whenever this norm is finite. When the probability distribution $\mathbb{P}$ is clear from the context, we adopt $\|f\|_2$ as a convenient shorthand for $\|f\|_{L^2(\mathbb{P})}$.

Given a set of n samples $\{x_i\}_{i=1}^n := \{x_1, x_2, \ldots, x_n\}$, each drawn i.i.d. according to $\mathbb{P}$, consider the empirical distribution

$$\mathbb{P}_n(x) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(x)$$

that places mass $1/n$ at each sample. It induces the *empirical L^2-norm*

$$\|f\|_{L^2(\mathbb{P}_n)}^2 := \frac{1}{n}\sum_{i=1}^n f^2(x_i) = \int_{\mathcal{X}} f^2(x)\mathbb{P}_n(dx). \tag{14.2}$$

Again, to lighten notation, when the underlying empirical distribution $\mathbb{P}_n$ is clear from context, we adopt the convenient shorthand $\|f\|_n$ for $\|f\|_{L^2(\mathbb{P}_n)}$.

In our analysis of nonparametric least squares from Chapter 13, we provided bounds on

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

the $L^2(\mathbb{P}_n)$-error in which the samples $\{x_i\}_{i=1}^n$ were viewed as fixed. By contrast, throughout this chapter, we view the samples as being random variables, so that the empirical norm is itself a random variable. Since each $x_i \sim \mathbb{P}$, the linearity of expectation guarantees that

$$\mathbb{E}[\|f\|_n^2] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f^2(x_i)\right] = \|f\|_2^2 \qquad \text{for any function } f \in L^2(\mathbb{P}).$$

Consequently, under relatively mild conditions on the random variable $f(x)$, the law of large numbers implies that $\|f\|_n^2$ converges to $\|f\|_2^2$. Such a limit theorem has its usual non-asymptotic analogs: for instance, if the function f is uniformly bounded, that is, if

$$\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)| \leq b \qquad \text{for some } b < \infty,$$

then Hoeffding's inequality (cf. Proposition 2.5 and equation (2.11)) implies that

$$\mathbb{P}\left[\left|\|f\|_n^2 - \|f\|_2^2\right| \geq t\right] \leq 2e^{-\frac{nt^2}{2b^4}}.$$

As in Chapter 4, our interest is in extending this type of tail bound—valid for a single function f—to a result that applies uniformly to all functions in a certain function class $\mathscr{F}$. Our analysis in this chapter, however, will be more refined: by using localized forms of complexity, we obtain optimal bounds.

14.1.1 A uniform law with localization

We begin by stating a theorem that controls the deviations in the random variable $\left|\|f\|_n - \|f\|_2\right|$, when measured in a uniform sense over a function class $\mathscr{F}$. We then illustrate some consequences of this result in application to nonparametric regression.

As with our earlier results on nonparametric least squares from Chapter 13, our result is stated in terms of a localized form of Rademacher complexity. For the current purposes, it is convenient to define the complexity at the population level. For a given radius $\delta > 0$ and function class $\mathscr{F}$, consider the *localized population Rademacher complexity*

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}) = \mathbb{E}_{\varepsilon, x}\left[\sup_{\substack{f \in \mathscr{F} \\ \|f\|_2 \leq \delta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right], \tag{14.3}$$

where $\{x_i\}_{i=1}^n$ are i.i.d. samples from some underlying distribution $\mathbb{P}$, and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher variables taking values in $\{-1, +1\}$ equiprobably, independent of the sequence $\{x_i\}_{i=1}^n$.

In the following result, we assume that $\mathscr{F}$ is star-shaped around the origin, meaning that, for any $f \in \mathscr{F}$ and scalar $\alpha \in [0, 1]$, the function αf also belongs to $\mathscr{F}$. In addition, we require the function class to be b-uniformly bounded, meaning that there is a constant $b < \infty$ such that $\|f\|_\infty \leq b$ for all $f \in \mathscr{F}$.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Theorem 14.1 *Given a star-shaped and b-uniformly bounded function class $\mathscr{F}$, let δ_n be any positive solution of the inequality*

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}) \leq \frac{\delta^2}{b}. \tag{14.4}$$

Then for any $t \geq \delta_n$, we have

$$\left| \|f\|_n^2 - \|f\|_2^2 \right| \leq \frac{1}{2}\|f\|_2^2 + \frac{t^2}{2} \qquad \text{for all } f \in \mathscr{F} \tag{14.5a}$$

with probability at least $1 - c_1 e^{-c_2 \frac{nt^2}{b^2}}$. If in addition $n\delta_n^2 \geq \frac{2}{c_2}\log(4\log(1/\delta_n))$, then

$$\left| \|f\|_n - \|f\|_2 \right| \leq c_0\, \delta_n \qquad \text{for all } f \in \mathscr{F} \tag{14.5b}$$

with probability at least $1 - c_1' e^{-c_2' \frac{n\delta_n^2}{b^2}}$.

It is worth noting that a similar result holds in terms of the localized *empirical Rademacher complexity*, namely the data-dependent quantity

$$\widehat{\mathcal{R}}_n(\delta) \equiv \widehat{\mathcal{R}}_n(\delta; \mathscr{F}) := \mathbb{E}_\varepsilon\Big[\sup_{\substack{f \in \mathscr{F} \\ \|f\|_n \leq \delta}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i) \Big|\Big], \tag{14.6}$$

and any positive solution $\hat{\delta}_n$ to the inequality

$$\widehat{\mathcal{R}}_n(\delta) \leq \frac{\delta^2}{b}. \tag{14.7}$$

Since the Rademacher complexity $\widehat{\mathcal{R}}_n$ depends on the data, this critical radius $\hat{\delta}_n$ is a random quantity, but it is closely related to the deterministic radius δ_n defined in terms of the population Rademacher complexity (14.3). More precisely, let δ_n and $\hat{\delta}_n$ denote the smallest positive solutions to inequalities (14.4) and (14.7), respectively. Then there are universal constants $c < 1 < C$ such that, with probability at least $1 - c_1 e^{-c_2 \frac{n\delta_n^2}{b}}$, we are guaranteed that $\hat{\delta}_n \in [c\delta_n,\ C\delta_n]$, and hence

$$\left| \|f\|_n - \|f\|_2 \right| \leq \frac{c_0}{c}\,\hat{\delta}_n \qquad \text{for all } f \in \mathscr{F}. \tag{14.8}$$

See Proposition 14.25 in the Appendix (Section 14.5) for the details and proof.

Theorem 14.1 is best understood by considering some concrete examples.

Example 14.2 (Bounds for quadratic functions) For a given coefficient vector $\theta \in \mathbb{R}^3$, define the quadratic function $f_\theta(x) := \theta_0 + \theta_1 x + \theta_2 x^2$, and let us consider the set of all bounded quadratic functions over the unit interval $[-1, 1]$, that is, the function class

$$\mathcal{P}_2 := \{f_\theta \text{ for some } \theta \in \mathbb{R}^3 \text{ such that } \max_{x \in [-1,1]} |f_\theta(x)| \leq 1\}. \tag{14.9}$$

Suppose that we are interested in relating the population and empirical L^2-norms uniformly over this family, when the samples are drawn from the uniform distribution over $[-1, 1]$.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

We begin by exploring a naive approach, one that ignores localization and hence leads to a sub-optimal rate. From our results on VC dimension in Chapter 4—in particular, see Proposition 4.20—it is straightforward to see that $\mathcal{P}_2$ has VC dimension at most 3. In conjunction with the boundedness of the function class, Lemma 4.14 guarantees that for any $\delta > 0$, we have

$$\mathbb{E}_\varepsilon\Big[\sup_{\substack{f_\theta \in \mathcal{P}_2 \\ \|f_\theta\|_2 \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big] \overset{\text{(i)}}{\leq} 2\sqrt{\frac{3\log(n+1)}{n}} \leq 4\sqrt{\frac{\log(n+1)}{n}} \tag{14.10}$$

for any set of samples $\{x_i\}_{i=1}^n$. As we will see, this upper bound is actually rather loose for small values of δ, since inequality (i) makes no use of the localization condition $\|f_\theta\|_2 \leq \delta$.

Based on the naive upper bound (14.10), we can conclude that there is a constant c_0 such that inequality (14.4) is satisfied with $\delta_n = c_0(\frac{\log(n+1)}{n})^{1/4}$. Thus, for any $t \geq c_0(\frac{\log(n+1)}{n})^{1/4}$, Theorem 14.1 guarantees that

$$\Big|\|f\|_n^2 - \|f\|_2^2\Big| \leq \tfrac{1}{2}\|f\|_2^2 + t^2 \qquad \text{for all } f \in \mathcal{P}_2 \tag{14.11}$$

with probability at least $1 - c_1 e^{-c_2 t^2}$. This bound establishes that $\|f\|_2^2$ and $\|f\|_n^2$ are of the same order for all functions with norm $\|f\|_2 \geq c_0(\frac{\log(n+1)}{n})^{1/4}$, but this order of fluctuation is sub-optimal. As we explore in Exercise 14.3, an entropy integral approach can be used to remove the superfluous logarithm from this result, but the slow $n^{-1/4}$ rate remains.

Let us now see how localization can be exploited to yield the optimal scaling $n^{-1/2}$. In order to do so, it is convenient to re-parameterize our quadratic functions in terms of an orthonormal basis of $L^2[-1, 1]$. In particular, the first three functions in the Legendre basis take the form

$$\phi_0(x) = \frac{1}{\sqrt{2}}, \quad \phi_1(x) = \sqrt{\frac{3}{2}}\, x \quad \text{and} \quad \phi_2(x) = \sqrt{\frac{5}{8}}\,(3x^2 - 1).$$

By construction, these functions are orthonormal in $L^2[-1, 1]$, meaning that the inner product $\langle \phi_j, \phi_k \rangle_{L^2[-1,1]} := \int_{-1}^1 \phi_j(x)\phi_k(x)\,dx$ is equal to one if $j = k$, and zero otherwise. Using these basis functions, any polynomial function in $\mathcal{P}_2$ then has an expansion of the form $f_\gamma(x) = \gamma_0\phi_0(x) + \gamma_1\phi_1(x) + \gamma_2\phi_2(x)$, where $\|f_\gamma\|_2 = \|\gamma\|_2$ by construction. Given a set of n samples, let us define an $n \times 3$ matrix $\mathbf{M}$ with entries $M_{ij} = \phi_j(x_i)$. In terms of this matrix, we then have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\substack{f_\gamma \in \mathcal{P}_2 \\ \|f_\gamma\|_2 \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f_\gamma(x_i)\Big|\Big] &\leq \mathbb{E}\Big[\sup_{\|\gamma\|_2 \leq \delta} \Big|\frac{1}{n}\varepsilon^{\mathrm{T}}\mathbf{M}\gamma\Big|\Big] \\
&\overset{\text{(i)}}{\leq} \frac{\delta}{n}\,\mathbb{E}\Big[\|\varepsilon^{\mathrm{T}}\mathbf{M}\|_2\Big] \\
&\overset{\text{(ii)}}{\leq} \frac{\delta}{n}\,\sqrt{\mathbb{E}[\|\varepsilon^{\mathrm{T}}\mathbf{M}\|_2^2]},
\end{aligned}$$

where step (i) follows from the Cauchy–Schwarz inequality, and step (ii) follows from Jensen's inequality and concavity of the square-root function. Now since the Rademacher variables are independent, we have

$$\mathbb{E}_\varepsilon[\|\varepsilon^{\mathrm{T}}\mathbf{M}\|_2^2] = \operatorname{trace}(\mathbf{M}\mathbf{M}^{\mathrm{T}}) = \operatorname{trace}(\mathbf{M}^{\mathrm{T}}\mathbf{M}).$$

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

By the orthonormality of the basis $\{\phi_0, \phi_1, \phi_2\}$, we have $\mathbb{E}_x[\text{trace}(\mathbf{M}^T\mathbf{M})] = 3n$. Putting together the pieces yields the upper bound

$$\mathbb{E}\Big[\sup_{\substack{f_\gamma \in \mathcal{P}_2 \\ \|f_\gamma\|_2 \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f_\gamma(x_i)\Big|\Big] \leq \frac{\sqrt{3}\,\delta}{\sqrt{n}}.$$

Based on this bound, we see that there is a universal constant c such that inequality (14.4) is satisfied with $\delta_n = \frac{c}{\sqrt{n}}$. Applying Theorem 4.10 then guarantees that for any $t \geq \frac{c}{\sqrt{n}}$, we have

$$\Big|\|f\|_n^2 - \|f\|_2^2\Big| \leq \frac{\|f\|_2^2}{2} + \frac{1}{2}t^2 \qquad \text{for all } f \in \mathcal{P}_2, \tag{14.12}$$

a bound that holds with probability at least $1 - c_1 e^{-c_2 n t^2}$. Unlike the earlier bound (14.11), this result has exploited the localization and thereby increased the rate from the slow one of $(\frac{\log n}{n})^{1/4}$ to the optimal one of $(\frac{1}{n})^{1/2}$. ♣

Whereas the previous example concerned a parametric class of functions, Theorem 14.1 also applies to nonparametric function classes. Since metric entropy has been computed for many such classes, it provides one direct route for obtaining upper bounds on the solutions of inequalities (14.4) or (14.7). One such avenue is summarized in the following:

Corollary 14.3 *Let $N_n(t; \mathbb{B}_n(\delta; \mathscr{F}))$ denote the t-covering number of the set $\mathbb{B}_n(\delta; \mathscr{F}) = \{f \in \mathscr{F} \mid \|f\|_n \leq \delta\}$ in the empirical $L^2(\mathbb{P}_n)$-norm. Then the empirical version of critical inequality* (14.7) *is satisfied for any $\delta > 0$ such that*

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2b}}^{\delta} \sqrt{\log N_n(t; \mathbb{B}_n(\delta; \mathscr{F}))}\, dt \leq \frac{\delta^2}{b}. \tag{14.13}$$

The proof of this result is essentially identical to the proof of Corollary 13.7, so that we leave the details to the reader.

In order to make use of Corollary 14.3, we need to control the covering number N_n in the empirical $L^2(\mathbb{P}_n)$-norm. One approach is based on observing that the covering number N_n can always bounded by the covering number $N_{\sup}$ in the supremum norm $\|\cdot\|_\infty$. Let us illustrate this approach with an example.

Example 14.4 (Bounds for convex Lipschitz functions) Recall from Example 13.11 the class of convex 1-Lipschitz functions

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

From known results, the metric entropy of this function class in the sup-norm is upper bounded as $\log N_{\sup}(t; \mathscr{F}_{\text{conv}}) \precsim t^{-1/2}$ for all $t > 0$ sufficiently small (see the bibliographic

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

section for details). Thus, in order to apply Corollary 14.3, it suffices to find $\delta > 0$ such that

$$\frac{1}{\sqrt{n}} \int_0^{\delta} (1/t)^{1/4} \, dt = \frac{1}{\sqrt{n}} \, \frac{4}{3} \, \delta^{3/4} \; \precsim \; \delta^2.$$

Setting $\delta = c\, n^{-2/5}$ for a sufficiently large constant $c > 0$ is suitable, and applying Theorem 14.1 with this choice yields

$$\left| \|f\|_2 - \|f\|_n \right| \leq c' \; n^{-2/5} \qquad \text{for all } f \in \mathscr{F}_{\text{conv}}([0,1]; 1)$$

with probability greater than $1 - c_1 e^{-c_2 n^{1/5}}$. ♣

In the exercises at the end of this chapter, we explore various other results that can be derived using Corollary 14.3.

14.1.2 Specialization to kernel classes

As discussed in Chapter 12, reproducing kernel Hilbert spaces (RKHSs) have a number of attractive computational properties in application to nonparametric estimation. In this section, we discuss the specialization of Theorem 14.1 to the case of a function class $\mathscr{F}$ that corresponds to the unit ball of an RKHS.

Recall that any RKHS is specified by a symmetric, positive semidefinite kernel function $\mathcal{K}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Under mild conditions, Mercer's theorem (as stated previously in Theorem 12.20) ensures that $\mathcal{K}$ has a countable collection of non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$. The following corollary shows that the population form of the localized Rademacher complexity for an RKHS is determined by the decay rate of these eigenvalues, and similarly, the empirical version is determined by the eigenvalues of the empirical kernel matrix.

Corollary 14.5 *Let $\mathscr{F} = \{f \in \mathbb{H} \mid \|f\|_{\mathbb{H}} \leq 1\}$ be the unit ball of an RKHS with eigenvalues $(\mu_j)_{j=1}^{\infty}$. Then the localized population Rademacher complexity* (14.3) *is upper bounded as*

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}) \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\}}. \tag{14.14a}$$

Similarly, letting $(\widehat{\mu}_j)_{j=1}^{n}$ denote the eigenvalues of the renormalized kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$, the localized empirical Rademacher complexity (14.6) *is upper bounded as*

$$\widehat{\mathcal{R}}_n(\delta; \mathscr{F}) \leq \sqrt{\frac{2}{n}} \sqrt{\sum_{j=1}^{n} \min\{\widehat{\mu}_j, \delta^2\}}. \tag{14.14b}$$

Given knowledge of the eigenvalues of the kernel (operator or matrix), these upper bounds on the localized Rademacher complexities allow us to specify values δ_n that satisfy the inequalities (14.4) and (14.7), in the population and empirical cases, respectively. Lemma 13.22

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

from Chapter 13 provides an upper bound on the empirical Gaussian complexity for a kernel class, which yields the claim (14.14b). The proof of inequality (14.14a) is based on techniques similar to the proof of Lemma 13.22; we work through the details in Exercise 14.4.

Let us illustrate the use of Corollary 14.5 with some examples.

Example 14.6 (Bounds for first-order Sobolev space) Consider the first-order Sobolev space

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

Recall from Example 12.16 that it is a reproducing kernel Hilbert space with kernel function $\mathcal{K}(x, z) = \min\{x, z\}$. From the result of Exercise 12.14, the unit ball $\{f \in \mathbb{H}^1[0,1] \mid \|f\|_{\mathbb{H}} \leq 1\}$ is uniformly bounded with $b = 1$, so that Corollary 14.5 may be applied. Moreover, from Example 12.23, the eigenvalues of this kernel function are given by $\mu_j = (\frac{2}{(2j-1)\pi})^2$ for $j = 1, 2, \ldots$. Using calculations analogous to those from Example 13.20, it can be shown that

$$\frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{\infty} \min\{\delta^2, \mu_j\}} \leq c' \sqrt{\frac{\delta}{n}}$$

for some universal constant $c' > 0$. Consequently, Corollary 14.5 implies that the critical inequality (14.4) is satisfied for $\delta_n = cn^{-1/3}$. Applying Theorem 14.1, we conclude that

$$\sup_{\|f\|_{\mathbb{H}^1[0,1]} \leq 1} \left| \|f\|_2 - \|f\|_n \right| \leq c_0 n^{-1/3}$$

with probability greater than $1 - c_1 e^{-c_2 n^{1/3}}$. ♣

Example 14.7 (Bounds for Gaussian kernels) Consider the RKHS generated by the Gaussian kernel $\mathcal{K}(x, z) = e^{-\frac{1}{2}(x-z)^2}$ defined on the unit square $[-1, 1] \times [-1, 1]$. As discussed in Example 13.21, there are universal constants (c_0, c_1) such that the eigenvalues of the associated kernel operator satisfy a bound of the form

$$\mu_j \leq c_0 \, e^{-c_1 j \log j} \qquad \text{for } j = 1, 2, \ldots.$$

Following the same line of calculation as in Example 13.21, it is straightforward to show that inequality (14.14a) is satisfied by $\delta_n = c_0 \sqrt{\frac{\log(n+1)}{n}}$ for a sufficiently large but universal constant c_0. Consequently, Theorem 14.1 implies that, for the unit ball of the Gaussian kernel RKHS, we have

$$\sup_{\|f\|_{\mathbb{H}} \leq 1} \left| \|f\|_2 - \|f\|_n \right| \leq c_0 \sqrt{\frac{\log(n+1)}{n}}$$

with probability greater than $1 - 2e^{-c_1 \log(n+1)}$. By comparison to the parametric function class discussed in Example 14.2, we see that the unit ball of a Gaussian kernel RKHS obeys a uniform law with a similar rate. This fact illustrates that the unit ball of the Gaussian kernel RKHS—even though nonparametric in nature—is still relatively small. ♣

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

14.1.3 Proof of Theorem 14.1

Let us now return to prove Theorem 14.1. By a rescaling argument, it suffices to consider the case $b = 1$. Moreover, it is convenient to redefine δ_n as a positive solution to the inequality

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}) \leq \frac{\delta^2}{16}. \tag{14.15}$$

This new δ_n is simply a rescaled version of the original one, and we shall use it to prove a version of the theorem with $c_0 = 1$.

With these simplifications, our proof is based on the family of random variables

$$Z_n(r) := \sup_{f \in \mathbb{B}_2(r; \mathscr{F})} \left| \|f\|_2^2 - \|f\|_n^2 \right|, \qquad \text{where } \mathbb{B}_2(r; \mathscr{F}) = \{f \in \mathscr{F} \mid \|f\|_2 \leq r\}, \tag{14.16}$$

indexed by $r \in (0, 1]$. We let $\mathcal{E}_0$ and $\mathcal{E}_1$, respectively, denote the events that inequality (14.5a) or inequality (14.5b) are violated. We also define the auxiliary events $\mathcal{A}_0(r) := \{Z_n(r) \geq r^2/2\}$, and

$$\mathcal{A}_1 := \left\{ Z_n(\|f\|_2) \geq \delta_n \|f\|_2 \text{ for some } f \in \mathscr{F} \text{ with } \|f\|_2 \geq \delta_n \right\}.$$

The following lemma shows that it suffices to control these two auxiliary events:

Lemma 14.8 *For any star-shaped function class, we have*

$$\mathcal{E}_0 \overset{\text{(i)}}{\subseteq} \mathcal{A}_0(t) \quad \text{and} \quad \mathcal{E}_1 \overset{\text{(ii)}}{\subseteq} \mathcal{A}_0(\delta_n) \cup \mathcal{A}_1. \tag{14.17}$$

Proof Beginning with the inclusion (i), we divide the analysis into two cases. First, suppose that there exists some function with norm $\|f\|_2 \leq t$ that violates inequality (14.5a). For this function, we must have $\left| \|f\|_n^2 - \|f\|_2^2 \right| > \frac{t^2}{2}$, showing that $Z_n(t) > \frac{t^2}{2}$ so that $\mathcal{A}_0(t)$ must hold. Otherwise, suppose that the inequality (14.5a) is violated by some function with $\|f\|_2 > t$. Any such function satisfies the inequality $\left| \|f\|_2^2 - \|f\|_n^2 \right| > \|f\|_2^2/2$. We may then define the rescaled function $\widetilde{f} = \frac{t}{\|f\|_2} f$; by construction, it has $\|\widetilde{f}\|_2 = t$, and also belongs to $\mathscr{F}$ due to the star-shaped condition. Hence, reasoning as before, we find that $\mathcal{A}_0(t)$ must also hold in this case.

Turning to the inclusion (ii), it is equivalent to show that $\mathcal{A}_0^c(\delta_n) \cap \mathcal{A}_1^c \subseteq \mathcal{E}_1^c$. We split the analysis into two cases:

Case 1: Consider a function $f \in \mathscr{F}$ with $\|f\|_2 \leq \delta_n$. Then on the complement of $\mathcal{A}_0(\delta_n)$, either we have $\|f\|_n \leq \delta_n$, in which case $\left| \|f\|_n - \|f\|_2 \right| \leq \delta_n$, or we have $\|f\|_n \geq \delta_n$, in which case

$$\left| \|f\|_n - \|f\|_2 \right| = \frac{\left| \|f\|_2^2 - \|f\|_n^2 \right|}{\|f\|_n + \|f\|_2} \leq \frac{\delta_n^2}{\delta_n} = \delta_n.$$

Case 2: Next consider a function $f \in \mathscr{F}$ with $\|f\|_2 > \delta_n$. In this case, on the complement

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

of $\mathcal{A}_1$, we have

$$\big|\|f\|_n - \|f\|_2\big| = \frac{\big|\|f\|_n^2 - \|f\|_2^2\big|}{\|f\|_n + \|f\|_2} \leq \frac{\|f\|_2\, \delta_n}{\|f\|_n + \|f\|_2} \leq \delta_n,$$

which completes the proof. □

In order to control the events $\mathcal{A}_0(r)$ and $\mathcal{A}_1$, we need to control the tail behavior of the random variable $Z_n(r)$.

Lemma 14.9 *For all $r, s \geq \delta_n$, we have*

$$\mathbb{P}\Big[Z_n(r) \geq \frac{r\delta_n}{4} + \frac{s^2}{4}\Big] \leq 2e^{-c_2 n \min\{\frac{s^4}{r^2}, s^2\}}. \tag{14.18}$$

Setting both r and s equal to $t \geq \delta_n$ in Lemma 14.9 yields the bound $\mathbb{P}[\mathcal{A}_0(t)] \leq 2e^{-c_2 n t^2}$. Using inclusion (i) in Lemma 14.8, this completes the proof of inequality (14.5a).

Let us now prove Lemma 14.9.

Proof Beginning with the expectation, we have

$$\mathbb{E}[Z_n(r)] \overset{\text{(i)}}{\leq} 2\mathbb{E}\Big[\sup_{f \in \mathbb{B}_2(r;\mathscr{F})} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f^2(x_i)\Big|\Big] \overset{\text{(ii)}}{\leq} 4\mathbb{E}\Big[\sup_{f \in \mathbb{B}_2(r;\mathscr{F})} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big] = 4\mathcal{R}_n(r),$$

where step (i) uses a standard symmetrization argument (in particular, see the proof of Theorem 4.10 in Chapter 4); and step (ii) follows from the boundedness assumption ($\|f\|_\infty \leq 1$ uniformly for all $f \in \mathscr{F}$) and the Ledoux–Talagrand contraction inequality (5.61) from Chapter 5. Given our star-shaped condition on the function class, Lemma 13.6 guarantees that the function $r \mapsto \mathcal{R}_n(r)/r$ is non-increasing on the interval $(0, \infty)$. Consequently, for any $r \geq \delta_n$, we have

$$\frac{\mathcal{R}_n(r)}{r} \overset{\text{(iii)}}{\leq} \frac{\mathcal{R}_n(\delta_n)}{\delta_n} \overset{\text{(iv)}}{\leq} \frac{\delta_n}{16}, \tag{14.19}$$

where step (iii) follows from the non-increasing property, and step (iv) follows from our definition of δ_n. Putting together the pieces, we find that the expectation is upper bounded as $\mathbb{E}[Z_n(r)] \leq \frac{r\delta_n}{4}$.

Next we establish a tail bound above the expectation using Talagrand's inequality from Theorem 3.27. Let f be an arbitrary member of $\mathbb{B}_2(r; \mathscr{F})$. Since $\|f\|_\infty \leq 1$ for all $f \in \mathscr{F}$, the recentered functions $g = f^2 - \mathbb{E}[f^2(X)]$ are bounded as $\|g\|_\infty \leq 1$, and moreover

$$\operatorname{var}(g) \leq \mathbb{E}[f^4] \leq \mathbb{E}[f^2] \leq r^2,$$

using the fact that $f \in \mathbb{B}_2(r; \mathscr{F})$. Consequently, by applying Talagrand's concentration inequality (3.83), we find that there is a universal constant c such that

$$\mathbb{P}\Big[Z_n(r) \geq \mathbb{E}[Z(r)] + \frac{s^2}{4}\Big] \leq 2\,\exp\left(-\frac{ns^4}{c\,(r^2 + r\delta_n + s^2)}\right) \leq e^{-c_2 n \min\{\frac{s^4}{r^2}, s^2\}},$$

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

where the final step uses the fact that $r \geq \delta_n$. □

It remains to use Lemmas 14.8 and 14.9 to establish inequality (14.5b). By combining inclusion (ii) in Lemma 14.8 with the union bound, it suffices to bound the sum $\mathbb{P}[\mathcal{A}_0(\delta_n)] + \mathbb{P}[\mathcal{A}_1]$. Setting $r = s = \delta_n$ in the bound (14.18) yields the bound $\mathbb{P}[\mathcal{A}_0(\delta_n)] \leq e^{-c_1 n\delta_n^2}$, whereas setting $s^2 = r\delta_n$ yields the bound

$$\mathbb{P}[Z_n(r) \geq \frac{r\delta_n}{2}] \leq 2e^{-c_2 n\delta_n^2}. \tag{14.20}$$

Given this bound, one is tempted to "complete" the proof by setting $r = \|f\|_2$, and applying the tail bound (14.20) to the variable $Z_n(\|f\|_2)$. The delicacy here is that the tail bound (14.20) applies only to a *deterministic* radius r, as opposed to the random[1] radius $\|f\|_2$. This difficulty can be addressed by using a so-called "peeling" argument. For $m = 1, 2, \ldots$, define the events

$$\mathcal{S}_m := \{f \in \mathscr{F} \mid 2^{m-1}\delta_n \leq \|f\|_2 \leq 2^m \delta_n\}.$$

Since $\|f\|_2 \leq \|f\|_\infty \leq 1$ by assumption, any function $\mathscr{F} \cap \{\|f\|_2 \geq \delta_n\}$ belongs to some $\mathcal{S}_m$ for $m \in \{1, 2, \ldots, M\}$, where $M \leq 4\log(1/\delta_n)$.

By the union bound, we have $\mathbb{P}(\mathcal{A}_1) \leq \sum_{m=1}^M \mathbb{P}(\mathcal{A}_1 \cap \mathcal{S}_m)$. Now if the event $\mathcal{A}_1 \cap \mathcal{S}_m$ occurs, then there is a function f with $\|f\|_2 \leq r_m := 2^m\delta_n$ such that

$$\left|\|f\|_n^2 - \|f\|_2^2\right| \geq \|f\|_2\, \delta_n \geq \tfrac{1}{2} r_m \delta_n.$$

Consequently, we have $\mathbb{P}[\mathcal{S}_m \cap \mathcal{E}_1] \leq \mathbb{P}[Z(r_m) \geq \frac{1}{2} r_m \delta_n] \leq e^{-c_2 n\delta_n^2}$, and putting together the pieces yields

$$\mathbb{P}[\mathcal{A}_1] \leq \sum_{m=1}^M e^{-n\delta_n^2/16} \leq e^{-c_2 n\delta_n^2 + \log M} \leq e^{-\frac{c_2 n\delta_n^2}{2}},$$

where the final step follows from the assumed inequality $\frac{c_2}{2} n\delta_n^2 \geq \log(4\log(1/\delta_n))$. □

14.2 A one-sided uniform law

A potentially limiting aspect of Theorem 14.1 is that it requires the underlying function class to be b-uniformly bounded. To a certain extent, this condition can be relaxed by instead imposing tail conditions of the sub-Gaussian or sub-exponential type. See the bibliographic discussion for references to results of this type.

However, in many applications—including the problem of nonparametric least squares from Chapter 13—it is the *lower bound* on $\|f\|_n^2$ that is of primary interest. As discussed in Chapter 2, for ordinary scalar random variables, such one-sided tail bounds can often be obtained under much milder conditions than their corresponding two-sided analogs. Concretely, in the current context, for any fixed function $f \in \mathscr{F}$, applying the lower tail bound (2.23) to the i.i.d. sequence $\{f(x_i)\}_{i=1}^n$ yields the guarantee

$$\mathbb{P}[\|f\|_n^2 \leq \|f\|_2^2 - t] \leq e^{-\frac{nt^2}{2\,\mathbb{E}[f^4(x)]}}. \tag{14.21}$$

[1] It is random because the norm of the function f that violates the bound is a random variable.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Consequently, whenever the fourth moment can be controlled by some multiple of the second moment, then we can obtain non-trivial lower tail bounds.

Our goal in this section is to derive lower tail bounds of this type that hold uniformly over a given function class. Let us state more precisely the type of fourth-moment control that is required. In particular, suppose that there exists a constant C such that

$$\mathbb{E}[f^4(x)] \leq C^2\, \mathbb{E}[f^2(x)] \qquad \text{for all } f \in \mathscr{F} \text{ with } \|f\|_2 \leq 1. \tag{14.22a}$$

When does a bound of this type hold? It is certainly implied by the global condition

$$\mathbb{E}[f^4(x)] \leq C^2 (\mathbb{E}[f^2(x)])^2 \qquad \text{for all } f \in \mathscr{F}. \tag{14.22b}$$

However, as illustrated in Example 14.11 below, there are other function classes for which the milder condition (14.22a) can hold while the stronger condition (14.22b) fails.

Let us illustrate these fourth-moment conditions with some examples.

Example 14.10 (Linear functions and random matrices) For a given vector $\theta \in \mathbb{R}^d$, define the linear function $f_\theta(x) = \langle x, \theta \rangle$, and consider the class of all linear functions $\mathscr{F}_{\text{lin}} = \{f_\theta \mid \theta \in \mathbb{R}^d\}$. As discussed in more detail in Example 14.13 to follow shortly, uniform laws for $\|f\|_n^2$ over such a function class are closely related to random matrix theory. Note that the linear function class $\mathscr{F}_{\text{lin}}$ is never uniformly bounded in a meaningful way. Nonetheless, it is still possible for the strong moment condition (14.22b) to hold under certain conditions on the zero-mean random vector x.

For instance, suppose that for each $\theta \in \mathbb{R}^d$, the random variable $f_\theta(x) = \langle x, \theta \rangle$ is Gaussian. In this case, using the standard formula (2.54) for the moments of a Gaussian random vector, we have $\mathbb{E}[f_\theta^4(x)] = 3(\mathbb{E}[f_\theta^2(x)])^2$, showing that condition (14.22b) holds uniformly with $C^2 = 3$. Note that C does not depend on the variance of $f_\theta(x)$, which can be arbitrarily large. Exercise 14.6 provides some examples of non-Gaussian variables for which the fourth-moment condition (14.22b) holds in application to linear functions. ♣

Example 14.11 (Additive nonparametric models) Given a univariate function class $\mathscr{G}$, consider the class of functions on $\mathbb{R}^d$ given by

$$\mathscr{F}_{\text{add}} = \{f : \mathbb{R}^d \to \mathbb{R} \mid f = \sum_{j=1}^{d} g_j \quad \text{for some } g_j \in \mathscr{G}\}. \tag{14.23}$$

The problem of estimating a function of this type is known as *additive regression*, and it provides one avenue for escaping the curse of dimension; see the bibliographic section for further discussion.

Suppose that the univariate function class $\mathscr{G}$ is uniformly bounded, say $\|g_j\|_\infty \leq b$ for all $g_j \in \mathscr{G}$, and consider a distribution over $x \in \mathbb{R}^d$ under which each $g_j(x_j)$ is a zero-mean random variable. (This latter assumption can always be ensured by a recentering step.) Assume moreover that the design vector $x \in \mathbb{R}^d$ has four-way independent components—that is, for any distinct quadruple (j, k, ℓ, m), the random variables (x_j, x_k, x_ℓ, x_m) are jointly independent. For a given $\delta \in (0, 1]$, consider a function $f = \sum_{j=1}^d g_j \in \mathscr{F}$ such that $\mathbb{E}[f^2(x)] = \delta^2$, or

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

equivalently, using our independence conditions, such that

$$\mathbb{E}[f^2(x)] = \sum_{j=1}^{d} \|g_j\|_2^2 = \delta^2.$$

For any such function, the fourth moment can be bounded as

$$\mathbb{E}[f^4(x)] = \mathbb{E}\Big[\Big(\sum_{j=1}^{d} g_j(x_j)\Big)^4\Big] = \sum_{j=1}^{d} \mathbb{E}[g_j^4(x_j)] + 6 \sum_{j \neq k} \mathbb{E}[g_j^2(x_j)]\mathbb{E}[g_k^2(x_k)]$$
$$\leq \sum_{j=1}^{d} \mathbb{E}[g_j^4(x_j)] + 6\delta^4,$$

where we have used the zero-mean property, and the four-way independence of the coordinates. Since $\|g_j\|_\infty \leq b$ for each $g_j \in \mathscr{G}$, we have $\mathbb{E}[g_j^4(x_j)] \leq b^2 \mathbb{E}[g_j^2(x_j)]$, and putting together the pieces yields

$$\mathbb{E}[f^4(x)] \leq b^2\delta^2 + 6\delta^4 \leq (b^2 + 6)\,\delta^2,$$

where the final step uses the fact that $\delta \leq 1$ by assumption. Consequently, for any $\delta \in (0, 1]$, the weaker condition (14.22a) holds with $C^2 = b^2 + 6$. ♣

Having seen some examples of function classes that satisfy the moment conditions (14.22a) and/or (14.22b), let us now state a one-sided uniform law. Recalling that $\bar{\mathcal{R}}_n$ denotes the population Rademacher complexity, consider the usual type of inequality

$$\frac{\bar{\mathcal{R}}_n(\delta; \mathscr{F})}{\delta} \leq \frac{\delta}{128C}, \tag{14.24}$$

where the constant C appears in the fourth-moment condition (14.22a). Our statement also involves the convenient shorthand $\mathbb{B}_2(\delta) := \{f \in \mathscr{F} \mid \|f\|_2 \leq \delta\}$.

Theorem 14.12 *Consider a star-shaped class $\mathscr{F}$ of functions, each zero-mean under $\mathbb{P}$, and such that the fourth-moment condition* (14.22a) *holds uniformly over $\mathscr{F}$, and suppose that the sample size n is large enough to ensure that there is a solution $\delta_n \leq 1$ to the inequality* (14.24)*. Then for any $\delta \in [\delta_n, 1]$, we have*

$$\|f\|_n^2 \geq \tfrac{1}{2}\|f\|_2^2 \qquad \text{for all } f \in \mathscr{F} \setminus \mathbb{B}_2(\delta) \tag{14.25}$$

with probability at least $1 - e^{-c_1 \frac{n\delta^2}{C^2}}$.

Remark: The set $\mathscr{F} \setminus \mathbb{B}_2(\delta)$ can be replaced with $\mathscr{F}$ whenever the set $\mathscr{F} \cap \mathbb{B}_2(\delta)$ is cone-like—that is, whenever any non-zero function $f \in \mathbb{B}_2(\delta) \cap \mathscr{F}$ can be rescaled by $\alpha := \delta/\|f\|_2 \geq 1$, thereby yielding a new function $g := \alpha f$ that *remains within* $\mathscr{F}$.

In order to illustrate Theorem 14.12, let us revisit our earlier examples.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Example 14.13 (Linear functions and random matrices, *continued*) Recall the linear function class $\mathscr{F}_{\text{lin}}$ introduced previously in Example 14.10. Uniform laws over this function class are closely related to earlier results on non-asymptotic random matrix theory from Chapter 6. In particular, supposing that the design vector x has a zero-mean distribution with covariance matrix $\mathbf{\Sigma}$, the function $f_\theta(x) = \langle x, \theta \rangle$ has $L^2(\mathbb{P})$-norm

$$\|f_\theta\|_2^2 = \theta^{\mathrm{T}} \mathbb{E}[xx^{\mathrm{T}}]\theta = \|\sqrt{\mathbf{\Sigma}}\theta\|_2^2 \qquad \text{for each } f_\theta \in \mathscr{F}. \tag{14.26}$$

On the other hand, given a set of n samples $\{x_i\}_{i=1}^n$, we have

$$\|f_\theta\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle x_i, \theta \rangle^2 = \frac{1}{n}\|\mathbf{X}\theta\|_2^2, \tag{14.27}$$

where the design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ has the vector x_i^{T} as its ith row. Consequently, in application to this function class, Theorem 14.12 provides a uniform lower bound on the quadratic forms $\frac{1}{n}\|\mathbf{X}\theta\|_2^2$: in particular, as long as the sample size n is large enough to ensure that $\delta_n \leq 1$, we have

$$\frac{1}{n}\|\mathbf{X}\theta\|_2^2 \geq \frac{1}{2}\|\sqrt{\mathbf{\Sigma}}\theta\|_2^2 \qquad \text{for all } \theta \in \mathbb{R}^d. \tag{14.28}$$

As one concrete example, suppose that the covariate vector x follows a $\mathcal{N}(0, \mathbf{\Sigma})$ distribution. For any $\theta \in \mathbb{S}^{d-1}$, the random variable $\langle x, \theta \rangle$ is sub-Gaussian with parameter at most $|||\sqrt{\mathbf{\Sigma}}|||_2$, but this quantity could be very large, and potentially growing with the dimension d. However, as discussed in Example 14.10, the strong moment condition (14.22b) always holds with $C^2 = 3$, regardless of the size of $|||\sqrt{\mathbf{\Sigma}}|||_2$. In order to apply Theorem 14.12, we need to determine a positive solution δ_n to the inequality (14.24). Writing each $x = \sqrt{\mathbf{\Sigma}}w$, where $w \sim \mathcal{N}(0, \mathbf{\Sigma})$, note that we have $\|f_\theta(x)\|_2 = \|\sqrt{\mathbf{\Sigma}}\theta\|_2$. Consequently, by definition of the local Rademacher complexity, we have

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}_{\text{lin}}) = \mathbb{E}\Big[\sup_{\substack{\theta \in \mathbb{R}^d \\ \|\sqrt{\mathbf{\Sigma}}\theta\|_2 \leq \delta}} \Big|\Big\langle \frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i, \ \sqrt{\mathbf{\Sigma}}\theta \Big\rangle\Big|\Big] = \delta\, \mathbb{E}\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i\|_2.$$

Note that the random variables $\{\varepsilon_i w_i\}_{i=1}^n$ are i.i.d. and standard Gaussian (since the symmetrization by independent Rademacher variables has no effect). Consequently, previous results from Chapter 2 guarantee that $\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i\|_2 \leq \sqrt{\frac{d}{n}}$. Putting together the pieces, we conclude that $\delta_n^2 \precsim \frac{d}{n}$. Therefore, for this particular ensemble, Theorem 14.12 implies that, as long as $n \succsim d$, then

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq \frac{1}{2}\|\sqrt{\mathbf{\Sigma}}\theta\|_2^2 \qquad \text{for all } \theta \in \mathbb{R}^d \tag{14.29}$$

with high probability. The key part of this lower bound is that the maximum eigenvalue $|||\sqrt{\mathbf{\Sigma}}|||_2$ never enters the result.

As another concrete example, the four-way independent and B-bounded random variables described in Exercise 14.6 also satisfy the moment condition (14.22b) with $C^2 = B + 6$. A similar calculation then shows that, with high probability, this ensemble also satisfies a lower bound of the form (14.29) where $\mathbf{\Sigma} = \mathbf{I}_d$. Note that these random variables need not be sub-Gaussian—in fact, the condition does not even require the existence of moments larger than four. ♣

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

In Exercise 14.7, we illustrate the use of Theorem 14.12 for controlling the restricted eigenvalues (RE) of some random matrix ensembles.

Let us now return to a nonparametric example:

Example 14.14 (Additive nonparametric models, *continued*) In this example, we return to the class $\mathscr{F}_{\text{add}}$ of additive nonparametric models previously introduced in Example 14.11. We let ε_n be the critical radius for the univariate function class $\mathscr{G}$ in the definition (14.23); thus, the scalar ε_n satisfies an inequality of the form $\bar{\mathcal{R}}_n(\varepsilon; \mathscr{F}) \precsim \varepsilon^2$. In Exercise 14.8, we prove that the critical radius δ_n for the d-dimensional additive family $\mathscr{F}_{\text{add}}$ satisfies the upper bound $\delta_n \precsim \sqrt{d}\,\varepsilon_n$. Consequently, Theorem 14.12 guarantees that

$$\|f\|_n^2 \geq \tfrac{1}{2}\|f\|_2^2 \qquad \text{for all } f \in \mathscr{F}_{\text{add}} \text{ with } \|f\|_2 \geq c_0\sqrt{d}\,\varepsilon_n \tag{14.30}$$

with probability at least $1 - e^{-c_1 nd\varepsilon_n^2}$.

As a concrete example, suppose that the univariate function class $\mathscr{G}$ is given by a first-order Sobolev space; for such a family, the univariate rate scales as $\varepsilon_n^2 \asymp n^{-2/3}$ (see Example 13.20 for details). For this particular class of additive models, with probability at least $1 - e^{-c_1 dn^{1/3}}$, we are guaranteed that

$$\underbrace{\|\sum_{j=1}^d g_j\|_n^2}_{\|f\|_n^2} \geq \frac{1}{2}\underbrace{\sum_{j=1}^d \|g_j\|_2^2}_{\|f\|_2^2} \tag{14.31}$$

uniformly over all functions of the form $f = \sum_{j=1}^d g_j$ with $\|f\|_2 \succsim \sqrt{d}n^{-1/3}$. ♣

14.2.1 Consequences for nonparametric least squares

Theorem 14.12, in conjunction with our earlier results from Chapter 13, has some immediate corollaries for nonparametric least squares. Recall the standard model for nonparametric regression, in which we observe noisy samples of the form $y_i = f^*(x_i) + \sigma w_i$, where $f^* \in \mathscr{F}$ is the unknown regression function. Our corollary involves the local complexity of the shifted function class $\mathscr{F}^* = \mathscr{F} - f^*$.

We let δ_n and ε_n (respectively) be any positive solutions to the inequalities

$$\frac{\bar{\mathcal{R}}_n(\delta; \mathscr{F})}{\delta} \overset{\text{(i)}}{\leq} \frac{\delta}{128C} \quad \text{and} \quad \frac{\mathcal{G}_n(\varepsilon; \mathscr{F}^*)}{\varepsilon} \overset{\text{(ii)}}{\leq} \frac{\varepsilon}{2\sigma}, \tag{14.32}$$

where the localized Gaussian complexity $\mathcal{G}_n(\varepsilon; \mathscr{F}^*)$ was defined in equation (13.16), prior to the statement of Theorem 13.5. To be clear, the quantity ε_n is a random variable, since it depends on the covariates $\{x_i\}_{i=1}^n$, which are modeled as random in this chapter.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Corollary 14.15 *Under the conditions of Theorems 13.5 and 14.12, there are universal positive constants (c_0, c_1, c_2) such that the nonparametric least-squares estimate $\widehat{f}$ satisfies*

$$\mathbb{P}_{w,x}\Big[\|\widehat{f} - f^*\|_2^2 \geq c_0\,(\varepsilon_n^2 + \delta_n^2)\Big] \leq c_1 e^{-c_2 \frac{n\delta_n^2}{\sigma^2 + C^2}}. \tag{14.33}$$

Proof We split the argument into two cases:

Case 1: Suppose that $\delta_n \geq \varepsilon_n$. We are then guaranteed that δ_n is a solution to inequality (ii) in equation (14.32). Consequently, we may apply Theorem 13.5 with $t = \delta_n$ to find that

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

On the other hand, Theorem 14.12 implies that

$$\mathbb{P}_{x,w}\Big[\|\widehat{f} - f^*\|_2^2 \geq 2\delta_n^2 + 2\|\widehat{f} - f^*\|_n^2\Big] \leq e^{-c_2 \frac{n\delta_n^2}{C^2}}.$$

Putting together the pieces yields that

$$\mathbb{P}_{x,w}\Big[\|\widehat{f} - f^*\|_2^2 \geq c_0\delta_n^2\Big] \leq c_1 e^{-c_2 \frac{n\delta_n^2}{\sigma^2 + C^2}},$$

which implies the claim.

Case 2: Otherwise, we may assume that the event $\mathcal{A} := \{\delta_n < \varepsilon_n\}$ holds. Note that this event depends on the random covariates $\{x_i\}_{i=1}^n$ via the random quantity ε_n. It suffices to bound the probability of the event $\mathcal{E} \cap \mathcal{A}$, where

$$\mathcal{E} := \Big\{\|\widehat{f} - f^*\|_2^2 \geq 16\varepsilon_n^2 + 2\delta_n^2\Big\}.$$

In order to do so, we introduce a third event, namely $\mathcal{B} := \{\|\widehat{f} - f^*\|_n^2 \leq 8\varepsilon_n^2\}$, and make note of the upper bound

$$\mathbb{P}[\mathcal{E} \cap \mathcal{A}] \leq \mathbb{P}[\mathcal{E} \cap \mathcal{B}] + \mathbb{P}[\mathcal{A} \cap \mathcal{B}^c].$$

On one hand, we have

$$\mathbb{P}[\mathcal{E} \cap \mathcal{B}] \leq \mathbb{P}\Big[\|\widehat{f} - f^*\|_2^2 \geq 2\|\widehat{f} - f^*\|_n^2 + 2\delta_n^2\Big] \leq e^{-c_2 \frac{n\delta_n^2}{C^2}},$$

where the final inequality follows from Theorem 14.12.

On the other hand, let $\mathbb{I}[\mathcal{A}]$ be a zero–one indicator for the event $\mathcal{A} := \{\delta_n < \varepsilon_n\}$. Then applying Theorem 13.5 with $t = \varepsilon_n$ yields

$$\mathbb{P}[\mathcal{A} \cap \mathcal{B}^c] \leq \mathbb{E}_x\Big[e^{-\frac{n\varepsilon_n^2}{2\sigma^2}}\,\mathbb{I}[\mathcal{A}]\Big] \leq e^{-\frac{n\delta_n^2}{2\sigma^2}}.$$

Putting together the pieces yields the claim. □

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

14.2.2 Proof of Theorem 14.12

Let us now turn to the proof of Theorem 14.12. We first claim that it suffices to consider functions belonging to the boundary of the δ-ball—namely, the set $\partial\mathbb{B}_2(\delta) = \{f \in \mathscr{F} \mid \|f\|_2 = \delta\}$. Indeed, suppose that the inequality (14.25) is violated for some $g \in \mathscr{F}$ with $\|g\|_2 > \delta$. By the star-shaped condition, the function $f := \frac{\delta}{\|g\|_2} g$ belongs to $\mathscr{F}$ and has norm $\|f\|_2 = \delta$. Finally, by rescaling, the inequality $\|g\|_n^2 < \frac{1}{2}\|g\|_2^2$ is equivalent to $\|f\|_n^2 < \frac{1}{2}\|f\|_2^2$.

For any function $f \in \partial\mathbb{B}_2(\delta)$, it is equivalent to show that

$$\|f\|_n^2 \geq \frac{3}{4}\|f\|_2^2 - \frac{\delta^2}{4}. \tag{14.34}$$

In order to prove this bound, we make use of a truncation argument. For a level $\tau > 0$ to be chosen, consider the truncated quadratic

$$\varphi_\tau(u) := \begin{cases} u^2 & \text{if } |u| \leq \tau, \\ \tau^2 & \text{otherwise,} \end{cases} \tag{14.35}$$

and define $f_\tau(x) = \operatorname{sign}(f(x))\sqrt{\varphi_\tau(f(x))}$. By construction, for any $f \in \partial\mathbb{B}_2(\delta)$, we have $\|f\|_n^2 \geq \|f_\tau\|_n^2$, and hence

$$\|f\|_n^2 \geq \|f_\tau\|_2^2 - \sup_{f \in \partial\mathbb{B}_2(\delta)} \Big| \|f_\tau\|_n^2 - \|f_\tau\|_2^2 \Big|. \tag{14.36}$$

The remainder of the proof consists of showing that a suitable choice of truncation level τ ensures that

$$\|f_\tau\|_2^2 \geq \tfrac{3}{4}\|f\|_2^2 \qquad \text{for all } f \in \partial\mathbb{B}_2(\delta) \tag{14.37a}$$

and

$$\mathbb{P}[Z_n \geq \tfrac{1}{4}\delta^2] \leq c_1 e^{-c_2 n\delta^2} \qquad \text{where } Z_n := \sup_{f \in \partial\mathbb{B}_2(\delta)} \Big| \|f_\tau\|_n^2 - \|f_\tau\|_2^2 \Big|. \tag{14.37b}$$

These two bounds in conjunction imply that the lower bound (14.34) holds with probability at least $1 - c_1 e^{-c_2 n\delta^2}$, uniformly all f with $\|f\|_2 = \delta$.

Proof of claim (14.37a)*:* Letting $\mathbb{I}[|f(x)| \geq \tau]$ be a zero–one indicator for the event $|f(x)| \geq \tau$, we have

$$\|f\|_2^2 - \|f_\tau\|_2^2 \leq \mathbb{E}\Big[f^2(x)\, \mathbb{I}[|f(x)| \geq \tau]\Big] \leq \sqrt{\mathbb{E}[f^4(x)]}\ \sqrt{\mathbb{P}[|f(x)| \geq \tau]},$$

where the last step uses the Cauchy–Schwarz inequality. Combining the moment bound (14.22a) with Markov's inequality yields

$$\|f\|_2^2 - \|f_\tau\|_2^2 \leq C\,\|f\|_2\ \sqrt{\frac{\mathbb{E}[f^4(x)]}{\tau^4}} \leq C^2\ \frac{\|f\|_2^2}{\tau^2},$$

where the final inequality uses the moment bound (14.22a) again. Setting $\tau^2 = 4C^2$ yields the bound $\|f\|_2^2 - \|f_\tau\|_2^2 \leq \frac{1}{4}\|f\|_2^2$, which is equivalent to the claim (14.37a).

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Proof of claim (14.37b)*:* Beginning with the expectation, a standard symmetrization argument (see Proposition 4.11) guarantees that

$$\mathbb{E}_x[Z_n] \leq 2\,\mathbb{E}_{x,\varepsilon}\Big[\sup_{f\in\mathbb{B}_2(\delta;\mathscr{F})}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f_\tau^2(x_i)\Big|\Big].$$

Our truncation procedure ensures that $f_\tau^2(x) = \varphi_\tau(f(x))$, where φ_τ is a Lipschitz function with constant $L = 2\tau$. Consequently, the Ledoux–Talagrand contraction inequality (5.61) guarantees that

$$\mathbb{E}_x[Z_n] \leq 8\tau\mathbb{E}_{x,\varepsilon}\Big[\sup_{f\in\mathbb{B}_2(\delta;\mathscr{F})}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|\Big] \leq 8\tau\bar{\mathcal{R}}_n(\delta;\mathscr{F}) \leq 8\tau\,\frac{\delta^2}{128C},$$

where the final step uses the assumed inequality $\bar{\mathcal{R}}_n(\delta;\mathscr{F}) \leq \frac{\delta^2}{128C}$. Our previous choice $\tau = 2C$ ensures that $\mathbb{E}_x[Z_n] \leq \frac{1}{8}\delta^2$.

Next we prove an upper tail bound on the random variable Z_n, in particular using Talagrand's theorem for empirical processes (Theorem 3.27). By construction, we have $\|f_\tau^2\|_\infty \leq \tau^2 = 4C^2$, and

$$\operatorname{var}(f_\tau^2(x)) \leq \mathbb{E}[f_\tau^4(x)] \leq \tau^2\|f\|_2^2 = 4C^2\,\delta^2.$$

Consequently, Talagrand's inequality (3.83) implies that

$$\mathbb{P}[Z_n \geq \mathbb{E}[Z_n] + u] \leq c_1\exp\left(-\frac{c_2 nu^2}{C\delta^2 + C^2 u}\right). \tag{14.38}$$

Since $\mathbb{E}[Z_n] \leq \frac{\delta^2}{8}$, the claim (14.37b) follows by setting $u = \frac{\delta^2}{8}$.

14.3 A uniform law for Lipschitz cost functions

Up to this point, we have considered uniform laws for the difference between the empirical squared norm $\|f\|_n^2$ and its expectation $\|f\|_2^2$. As formalized in Corollary 14.15, such results are useful, for example, in deriving bounds on the $L^2(\mathbb{P})$-error of the nonparametric least-squares estimator. In this section, we turn to a more general class of prediction problems, and a type of uniform law that is useful for many of them.

14.3.1 General prediction problems

A general prediction problem can be specified in terms of a space $\mathcal{X}$ of covariates or predictors, and a space $\mathcal{Y}$ of response variables. A predictor is a function f that maps a covariate $x \in \mathcal{X}$ to a prediction $\widehat{y} = f(x) \in \widetilde{\mathcal{Y}}$. Here the space $\widetilde{\mathcal{Y}}$ may be either the same as the response space $\mathcal{Y}$, or a superset thereof. The goodness of a predictor f is measured in terms of a cost function $\mathcal{L}\colon \widetilde{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, whose value $\mathcal{L}(\widehat{y}, y)$ corresponds to the cost of predicting $\widehat{y} \in \widetilde{\mathcal{Y}}$ when the underlying true response is some $y \in \mathcal{Y}$. Given a collection of n samples $\{(x_i, y_i)\}_{i=1}^n$, a natural way in which to determine a predictor is by minimizing the empirical cost

$$\mathbb{P}_n(\mathcal{L}(f(x), y)) := \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f(x_i), y_i). \tag{14.39}$$

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Although the estimator $\widehat{f}$ is obtained by minimizing the empirical cost (14.39), our ultimate goal is in assessing its quality when measured in terms of the population cost function

$$\mathbb{P}(\mathcal{L}(f(x), y)) := \mathbb{E}_{x,y}[\mathcal{L}(f(x), y)], \tag{14.40}$$

and our goal is thus to understand when a minimizer of the empirical cost (14.39) is a near-minimizer of the population cost.

As discussed previously in Chapter 4, this question can be addressed by deriving a suitable type of uniform law of large numbers. More precisely, for each $f \in \mathscr{F}$, let us define the function $\mathcal{L}_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ via $\mathcal{L}_f(x, y) = \mathcal{L}(f(x), y)$, and let us write

$$\mathbb{P}_n(\mathcal{L}_f) = \mathbb{P}_n(\mathcal{L}(f(x), y)) \quad \text{and} \quad \bar{\mathcal{L}}_f := \mathbb{P}(\mathcal{L}_f) = \mathbb{P}(\mathcal{L}(f(x), y)).$$

In terms of this convenient shorthand, our question can be understand as deriving a Glivenko–Cantelli law for the so-called *cost class* $\{\mathcal{L}_f \mid f \in \mathscr{F}\}$.

Throughout this section, we study prediction problems for which $\mathcal{Y}$ is some subset of the real line $\mathbb{R}$. For a given constant $L > 0$, we say that the cost function $\mathcal{L}$ is *L-Lipschitz in its first argument* if

$$|\mathcal{L}(z, y) - \mathcal{L}(\widetilde{z}, y)| \leq L|z - \widetilde{z}| \tag{14.41}$$

for all pairs $z, \widetilde{z} \in \widetilde{\mathcal{Y}}$ and $y \in \mathcal{Y}$. We say that the population cost function $f \mapsto \mathbb{P}(\mathcal{L}_f)$ is γ-*strongly convex* with respect to the $L^2(\mathbb{P})$-norm at f^* if there is some $\gamma > 0$ such that

$$\mathbb{P}\Big[\underbrace{\mathcal{L}_f}_{\mathcal{L}(f(x),y)} - \underbrace{\mathcal{L}_f^*}_{\mathcal{L}(f^*(x),y)} - \underbrace{\frac{\partial \mathcal{L}}{\partial z}\Big|_{f^*}}_{\frac{\partial \mathcal{L}}{\partial z}(f^*(x),y)} \underbrace{(f - f^*)}_{f(x)-f^*(x)}\Big] \geq \frac{\gamma}{2} \|f - f^*\|_2^2 \tag{14.42}$$

for all $f \in \mathscr{F}$. Note that it is sufficient (but not necessary) for the function $z \mapsto \mathcal{L}(z, y)$ to be γ-strongly convex in a pointwise sense for each $y \in \mathcal{Y}$. Let us illustrate these conditions with some examples.

Example 14.16 (Least-squares regression) In a standard regression problem, the response space $\mathcal{Y}$ is the real line or some subset thereof, and our goal is to estimate a regression function $x \mapsto f(x) \in \mathbb{R}$. In Chapter 13, we studied methods for nonparametric regression based on the least-squares cost $\mathcal{L}(z, y) = \frac{1}{2}(y-z)^2$. This cost function is *not* globally Lipschitz in general; however, it does become Lipschitz in certain special cases. For instance, consider the standard observation model $y = f^*(x) + \varepsilon$ in the special case of bounded noise—say $|\varepsilon| \leq c$ for some constant c. If we perform nonparametric regression over a b-uniformly bounded function class $\mathscr{F}$, then for all $f, g \in \mathscr{F}$, we have

$$\begin{aligned} \big|\mathcal{L}(f(x), y) - \mathcal{L}(g(x), y)\big| &= \tfrac{1}{2}\big|(y - f(x))^2 - (y - g(x))^2\big| \\ &\leq \tfrac{1}{2}\big|f^2(x) - g^2(x)\big| + |y|\,|f(x) - g(x)| \\ &\leq (b + (b + c))|f(x) - g(x)|, \end{aligned}$$

so that the least squares satisfies the Lipschitz condition (14.41) with $L = 2b + c$. Of course, this example is rather artificial since it excludes any types of non-bounded noise variables ε, including the canonical case of Gaussian noise.

In terms of strong convexity, note that, for any $y \in \mathbb{R}$, the function $z \mapsto \frac{1}{2}(y-z)^2$ is strongly

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

convex with parameter $\gamma = 1$, so that $f \mapsto \mathcal{L}_f$ satisfies the strong convexity condition (14.42) with $\gamma = 1$. ♣

Example 14.17 (Robust forms of regression) A concern with the use of the squared cost function in regression is its potential lack of robustness: if even a very small subset of observations are corrupted, then they can have an extremely large effect on the resulting solution. With this concern in mind, it is interesting to consider a more general family of cost functions, say of the form

$$\mathcal{L}(z, y) = \Psi(y - z), \tag{14.43}$$

where $\Psi : \mathbb{R} \to [0, \infty]$ is a function that is a symmetric around zero with $\Psi(0) = 0$, and almost everywhere differentiable with $\|\Psi'\|_\infty \leq L$. Note that the least-squares cost fails to satisfy the required derivative bound, so it does *not* fall within this class.

Examples of cost functions in the family (14.43) include the ℓ_1-norm $\Psi_{\ell_1}(u) = |u|$, as well as Huber's robust function

$$\Psi_{\text{Huber}}(u) = \begin{cases} \dfrac{u^2}{2} & \text{if } |u| \leq \tau, \\ \tau u - \dfrac{\tau^2}{2} & \text{otherwise,} \end{cases} \tag{14.44}$$

where $\tau > 0$ is a parameter to be specified. The Huber cost function offers some sort of compromise between the least-squares cost and the ℓ_1-norm cost function.

By construction, the function Ψ_{ℓ_1} is almost everywhere differentiable with $\|\Psi'_{\ell_1}\|_\infty \leq 1$, whereas the Huber cost function is everywhere differentiable with $\|\Psi'_{\text{Huber}}\|_\infty \leq \tau$. Consequently, the ℓ_1-norm and Huber cost functions satisfy the Lipschitz condition (14.41) with parameters $L = 1$ and $L = \tau$, respectively. Moreover, since the Huber cost function is locally equivalent to the least-squares cost, the induced cost function (14.43) is locally strongly convex under fairly mild tail conditions on the random variable $y - f(x)$. ♣

Example 14.18 (Logistic regression) The goal of binary classification is to predict a label $y \in \{-1, +1\}$ on the basis of a covariate vector $x \in \mathcal{X}$. Suppose that we model the conditional distribution of the label $y \in \{-1, +1\}$ as

$$\mathbb{P}_f(y \mid x) = \frac{1}{1 + e^{-2yf(x)}}, \tag{14.45}$$

where $f : \mathcal{X} \to \mathbb{R}$ is the discriminant function to be estimated. The method of maximum likelihood then corresponds to minimizing the cost function

$$\mathcal{L}_f(x, y) := \mathcal{L}(f(x), y) = \log\left(1 + e^{-2yf(x)}\right). \tag{14.46}$$

It is easy to see that the function $\mathcal{L}$ is 1-Lipschitz in its first argument. Moreover, at the population level, we have

$$\mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*}) = \mathbb{E}_{x,y}\Big[\log \frac{1 + e^{-2f(x)y}}{1 + e^{-2f^*(x)y}}\Big] = \mathbb{E}_x[D(\mathbb{P}_{f^*}(\cdot \mid x) \,\|\, \mathbb{P}_f(\cdot \mid x))],$$

corresponding to the expected value of the Kullback–Leibler divergence between the two conditional distributions indexed by f^* and f. Under relatively mild conditions on the behavior of the random variable $f(x)$ as f ranges over $\mathscr{F}$, this cost function will be γ-strongly convex. ♣

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

Example 14.19 (Support vector machines and hinge cost) Support vector machines are another method for binary classification, again based on estimating discriminant functions $f: \mathcal{X} \to \mathbb{R}$. In their most popular instantiation, the discriminant functions are assumed to belong to some reproducing kernel Hilbert space $\mathbb{H}$, equipped with the norm $\|\cdot\|_{\mathbb{H}}$. The support vector machine is based on the *hinge cost function*

$$\mathcal{L}(f(x), y) = \max\{0, 1 - yf(x)\}, \tag{14.47}$$

which is 1-Lipschitz by inspection. Again, the strong convexity properties of the population cost $f \mapsto \mathbb{P}(\mathcal{L}_f)$ depend on the distribution of the covariates x, and the function class $\mathscr{F}$ over which we optimize.

Given a set of n samples $\{(x_i, y_i)\}_{i=1}^n$, a common choice is to minimize the empirical risk

$$\mathbb{P}_n(\mathcal{L}(f(x), y)) = \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - y_i f(x_i)\}$$

over a ball $\|f\|_{\mathbb{H}} \leq R$ in some reproducing kernel Hilbert space. As explored in Exercise 12.20, this optimization problem can be reformulated as a quadratic program in n dimensions, and so can be solved easily. ♣

14.3.2 Uniform law for Lipschitz cost functions

With these examples as underlying motivation, let us now turn to stating a general uniform law for Lipschitz cost functions. Let $f^* \in \mathscr{F}$ minimize the population cost function $f \mapsto \mathbb{P}(\mathcal{L}_f)$, and consider the shifted function class.

$$\mathscr{F}^* := \{f - f^* \mid f \in \mathscr{F}\}. \tag{14.48}$$

Our uniform law involves the population version of the localized Rademacher complexity

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}^*) := \mathbb{E}_{x,\varepsilon}\left[\sup_{\substack{g \in \mathscr{F}^* \\ \|g\|_2 \leq \delta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, g(x_i)\right|\right]. \tag{14.49}$$

Theorem 14.20 (Uniform law for Lipschitz cost functions) *Given a uniformly* 1*-bounded function class $\mathscr{F}$ that is star-shaped around the population minimizer f^*, let $\delta_n^2 \geq \frac{c}{n}$ be any solution to the inequality*

$$\bar{\mathcal{R}}_n(\delta; \mathscr{F}^*) \leq \delta^2. \tag{14.50}$$

(a) *Suppose that the cost function is L-Lipschitz in its first argument. Then we have*

$$\sup_{f \in \mathscr{F}} \frac{\left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right|}{\|f - f^*\|_2 + \delta_n} \leq 10L\delta_n \tag{14.51}$$

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

(b) *Suppose that the cost function is L-Lipschitz and γ-strongly convex. Then for any*

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

function $\widehat{f} \in \mathscr{F}$ such that $\mathbb{P}_n(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 0$, we have

$$\|\widehat{f} - f^*\|_2 \leq \left(\frac{20L}{\gamma} + 1\right)\delta_n \tag{14.52a}$$

and

$$\mathbb{P}(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 10L\left(\frac{20L}{\gamma} + 2\right)\delta_n^2, \tag{14.52b}$$

where both inequalities hold with the same probability as in part (a).

Under certain additional conditions on the function class, part (a) can be used to guarantee consistency of a procedure that chooses $\widehat{f} \in \mathscr{F}$ to minimize the empirical cost $f \mapsto \mathbb{P}_n(\mathcal{L}_f)$ over $\mathscr{F}$. In particular, since $f^* \in \mathscr{F}$ by definition, this procedure ensures that $\mathbb{P}_n(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 0$. Consequently, for any function class $\mathscr{F}$ with[2] $\|\cdot\|_2$-diameter at most D, the inequality (14.51) implies that

$$\mathbb{P}(\mathcal{L}_{\widehat{f}}) \leq \mathbb{P}(\mathcal{L}_{f^*}) + 10L\delta_n\{2D + \delta_n\} \tag{14.53}$$

with high probability. Thus, the bound (14.53) implies the consistency of the empirical cost minimization procedure in the following sense: up to a term of order δ_n, the value $\mathbb{P}(\mathcal{L}_{\widehat{f}})$ is as small as the optimum $\mathbb{P}(\mathcal{L}_{f^*}) = \min_{f \in \mathscr{F}} \mathbb{P}(\mathcal{L}_f)$.

Proof of Theorem 14.20

The proof is based on an analysis of the family of random variables

$$Z_n(r) = \sup_{\|f - f^*\|_2 \leq r} \left|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})\right|,$$

where $r > 0$ is a radius to be varied. The following lemma provides suitable control on the upper tails of these random variables:

Lemma 14.21 *For each $r \geq \delta_n$, the variable $Z_n(r)$ satisfies the tail bound*

$$\mathbb{P}[Z_n(r) \geq 8L\, r\delta_n + u] \leq c_1 \exp\left(-\frac{c_2 n u^2}{L^2 r^2 + Lu}\right). \tag{14.54}$$

Deferring the proof of this intermediate claim for the moment, let us use it to complete the proof of Theorem 14.20; the proof itself is similar to that of Theorem 14.1. Define the events $\mathcal{E}_0 := \{Z_n(\delta_n) \geq 9L\delta_n^2\}$, and

$$\mathcal{E}_1 := \{\exists\, f \in \mathscr{F} \mid |\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})| \geq 10L\delta_n \|f - f^*\|_2 \text{ and } \|f - f^*\|_2 \geq \delta_n\}.$$

If there is some function $f \in \mathscr{F}$ that violates the bound (14.51), then at least one of the events $\mathcal{E}_0$ or $\mathcal{E}_1$ must occur. Applying Lemma 14.21 with $u = L\delta_n^2$ guarantees that $\mathbb{P}[\mathcal{E}_0] \leq c_1 e^{-c_2 n \delta_n^2}$. Moreover, using the same peeling argument as in Theorem 14.1, we find

[2] A function class $\mathscr{F}$ has $\|\cdot\|_2$-diameter at most D if $\|f\|_2 \leq D$ for all $f \in \mathscr{F}$. In this case, we have $\|\widehat{f} - f^*\|_2 \leq 2D$.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

that $\mathbb{P}[\mathcal{E}_1] \leq c_1 e^{-c_2' n \delta_n^2}$, valid for all $\delta_n^2 \geq \frac{c}{n}$. Putting together the pieces completes the proof of the claim (14.51) in part (a).

Let us now prove the claims in part (b). By examining the proof of part (a), we see that it actually implies that either $\|\widehat{f} - f^*\|_2 \leq \delta_n$, or

$$\left|\mathbb{P}_n(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*})\right| \leq 10 L \delta_n \|\widehat{f} - f^*\|_2.$$

Since $\mathbb{P}_n(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 0$ by assumption, we see that any minimizer must satisfy either the bound $\|\widehat{f} - f^*\|_2 \leq \delta_n$, or the bound $\mathbb{P}(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 10 L \delta_n \|\widehat{f} - f^*\|_2$. On one hand, if the former inequality holds, then so does inequality (14.52a). On the other hand, if the latter inequality holds, then, combined with the strong convexity condition (14.42), we obtain $\|\widehat{f} - f^*\|_2 \leq \frac{10L}{\gamma}$, which also implies inequality (14.52a).

In order to establish the bound (14.52b), we make use of inequality (14.52a) within the original inequality (14.51); we then perform some algebra, recalling that $\widehat{f}$ satisfies the inequality $\mathbb{P}_n(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 0$.

It remains to prove Lemma 14.21. By a rescaling argument, we may assume that $b = 1$. In order to bound the upper tail of $Z_n(r)$, we need to control the differences $\mathcal{L}_f - \mathcal{L}_{f^*}$ uniformly over all functions $f \in \mathscr{F}$ such that $\|f - f^*\|_2 \leq r$. By the Lipschitz condition on the cost function and the boundedness of the functions f, we have $|\mathcal{L}_f - \mathcal{L}_{f^*}|_\infty \leq L\|f - f^*\|_\infty \leq 2L$. Moreover, we have

$$\operatorname{var}(\mathcal{L}_f - \mathcal{L}_{f^*}) \leq \mathbb{P}[(\mathcal{L}_f - \mathcal{L}_{f^*})^2] \overset{\text{(i)}}{\leq} L^2 \|f - f^*\|_2^2 \overset{\text{(ii)}}{\leq} L^2 r^2,$$

where inequality (i) follows from the Lipschitz condition on the cost function, and inequality (ii) follows since $\|f - f^*\|_2 \leq r$. Consequently, by Talagrand's concentration theorem for empirical processes (Theorem 3.27), we have

$$\mathbb{P}\Big[Z_n(r) \geq 2\mathbb{E}[Z_n(r)] + u\Big] \leq c_1 \exp\left\{-\frac{c_2 n u^2}{L^2 r^2 + Lu}\right\}. \tag{14.55}$$

It remains to upper bound the expectation: in particular, we have

$$\begin{aligned}
\mathbb{E}[Z_n(r)] &\overset{\text{(i)}}{\leq} 2\mathbb{E}\Big[\sup_{\|f - f^*\|_2 \leq r} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \Big\{\mathcal{L}(f(x_i), y_i) - \mathcal{L}(f^*(x_i), y_i)\Big\}\Big|\Big] \\
&\overset{\text{(ii)}}{\leq} 4L\, \mathbb{E}\Big[\sup_{\|f - f^*\|_2 \leq r} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (f(x_i) - f^*(x_i))\Big|\Big] \\
&= 4L \bar{\mathcal{R}}_n(r; \mathscr{F}^*) \\
&\overset{\text{(iii)}}{\leq} 4L r \delta_n, \qquad \text{valid for all } r \geq \delta_n,
\end{aligned}$$

where step (i) follows from a symmetrization argument; step (ii) follows from the L-Lipschitz condition on the first argument of the cost function, and the Ledoux–Talagrand contraction inequality (5.61); and step (iii) uses the fact that the function $r \mapsto \frac{\bar{\mathcal{R}}_n(r; \mathscr{F}^*)}{r}$ is non-increasing, and our choice of δ_n. Combined with the tail bound (14.55), the proof of Lemma 14.21 is complete.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

14.4 Some consequences for nonparametric density estimation

The results and techniques developed thus far have some useful applications to the problem of nonparametric density estimation. The problem is easy to state: given a collection of i.i.d. samples $\{x_i\}_{i=1}^n$, assumed to have been drawn from an unknown distribution with density f^*, how do we estimate the unknown density? The density estimation problem has been the subject of intensive study, and there are many methods for tackling it. In this section, we restrict our attention to two simple methods that are easily analyzed using the results from this and preceding chapters.

14.4.1 Density estimation via the nonparametric maximum likelihood estimate

Perhaps the most easily conceived method for density estimation is via a nonparametric analog of maximum likelihood. In particular, suppose that we fix some base class of densities $\mathscr{F}$, and then maximize the likelihood of the observed samples over this class. Doing so leads to a constrained form of the nonparametric maximum likelihood estimate (MLE)—namely

$$\widehat{f} \in \arg\min_{f \in \mathscr{F}} \mathbb{P}_n(-\log f(x)) = \arg\min_{f \in \mathscr{F}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log f(x_i) \right\}. \tag{14.56}$$

To be clear, the class of densities $\mathscr{F}$ must be suitably restricted for this estimator to be well defined, which we assume to be the case for the present discussion. (See Exercise 14.9 for an example in which the nonparametric MLE $\widehat{f}$ fails to exist.) As an alternative to constraining the estimate, it also possible to define a regularized form of the nonparametric MLE.

In order to illustrate the use of some bounds from this chapter, let us analyze the estimator (14.56) in the simple case when the true density f^* is assumed to belong to $\mathscr{F}$. Given an understanding of this case, it is relatively straightforward to derive a more general result, in which the error is bounded by a combination of estimation error and approximation error terms, with the latter being non-zero when $f^* \notin \mathscr{F}$.

For reasons to be clarified, it is convenient to measure the error in terms of the squared *Hellinger distance*. For densities f and g with respect to a base measure μ, it is given by

$$H^2(f \,\|\, g) := \frac{1}{2} \int_{\mathcal{X}} (\sqrt{f} - \sqrt{g})^2 \, d\mu. \tag{14.57a}$$

As we explore in Exercise 14.10, a useful connection here is that the Kullback–Leibler (KL) divergence is lower bounded by (a multiple of) the squared Hellinger distance—viz.

$$D(f \,\|\, g) \geq 2H^2(f \,\|\, g). \tag{14.57b}$$

Up to a constant pre-factor, the squared Hellinger distance is equivalent to the $L^2(\mu)$-norm difference of the square-root densities. For this reason, the square-root function class $\mathscr{G} = \{g = \sqrt{f} \text{ for some } f \in \mathscr{F}\}$ plays an important role in our analysis, as does the shifted square-root function class $\mathscr{G}^* := \mathscr{G} - \sqrt{f^*}$.

In the relatively simple result to be given here, we assume that there are positive constants (b, v) such that the square-root density class $\mathscr{G}$ is $\sqrt{b}$-uniformly bounded, and star-shaped around $\sqrt{f^*}$, and moreover that the unknown density $f^* \in \mathscr{F}$ is uniformly lower bounded

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

as

$$f^*(x) \geq \nu > 0 \qquad \text{for all } x \in \mathcal{X}.$$

In terms of the population Rademacher complexity $\bar{\mathcal{R}}_n$, our result involves the critical inequality

$$\bar{\mathcal{R}}_n(\delta; \mathscr{G}^*) \leq \frac{\delta^2}{\sqrt{b+\nu}}. \tag{14.58}$$

With this set-up, we have the following guarantee:

Corollary 14.22 *Given a class of densities satisfying the previous conditions, let δ_n be any solution to the critical inequality* (14.58) *such that $\delta_n^2 \geq (1 + \frac{b}{\nu})\,\frac{1}{n}$. Then the nonparametric density estimate $\widehat{f}$ satisfies the Hellinger bound*

$$H^2(\widehat{f} \,\|\, f^*) \leq c_0\, \delta_n^2 \tag{14.59}$$

with probability greater than $1 - c_1 e^{-c_2 \frac{\nu}{b+\nu} n\delta_n^2}$.

Proof Our proof is based on applying Theorem 14.20(b) to the transformed function class

$$\mathcal{H} = \left\{ \sqrt{\frac{f+f^*}{2f^*}} \;\middle|\; f \in \mathscr{F} \right\}$$

equipped with the cost functions $\mathcal{L}_h(x) = -\log h(x)$. Since $\mathscr{F}$ is b-uniformly bounded and $f^*(x) \geq \nu$ for all $x \in \mathcal{X}$, for any $h \in \mathcal{H}$, we have

$$\|h\|_\infty = \left\| \sqrt{\frac{f+f^*}{2f^*}} \right\|_\infty \leq \sqrt{\frac{1}{2}\Big(\frac{b}{\nu} + 1\Big)} = \frac{1}{\sqrt{2\nu}}\ \sqrt{b+\nu}.$$

Moreover, for any $h \in \mathcal{H}$, we have $h(x) \geq 1/\sqrt{2}$ for all $x \in \mathcal{X}$ and whence the mean value theorem applied to the logarithm, combined with the triangle inequality, implies that

$$\big|\mathcal{L}_h(x) - \mathcal{L}_{\widetilde{h}}(x)\big| \leq \sqrt{2}\,|h(x) - \widetilde{h}(x)| \qquad \text{for all } x \in \mathcal{X}, \text{ and } h, \widetilde{h} \in \mathcal{H},$$

showing that the logarithmic cost function is L-Lipschitz with $L = \sqrt{2}$. Finally, by construction, for any $h \in \mathcal{H}$ and with $h^* := \frac{f^*+f^*}{2f^*} = 1$, we have

$$\|h - h^*\|_2^2 = \mathbb{E}_{f^*}\left[\Big\{\big(\frac{f+f^*}{2f^*}\big)^{\frac{1}{2}} - 1\Big\}^2\right] = 2H^2\left(\frac{f+f^*}{2} \,\|\, f^*\right).$$

Therefore, the lower bound (14.57b) on the squared Hellinger distance in terms of the KL divergence is equivalent to asserting that $\mathbb{P}(\mathcal{L}_h - \mathcal{L}_h^*) \geq \|h - h^*\|_2^2$, meaning that the cost function is 2-strongly convex around h^*. Consequently, the claim (14.59) follows via an application of Theorem 14.20(b). □

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

14.4.2 Density estimation via projections

Another very simple method for density estimation is via projection onto a function class $\mathscr{F}$. Concretely, again given n samples $\{x_i\}_{i=1}^n$, assumed to have been drawn from an unknown density f^* on a space $\mathcal{X}$, consider the projection-based estimator

$$\widehat{f} \in \arg\min_{f \in \mathscr{F}} \left\{ \frac{1}{2}\|f\|_2^2 - \mathbb{P}_n(f) \right\} = \arg\min_{f \in \mathscr{F}} \left\{ \frac{1}{2}\|f\|_2^2 - \frac{1}{n}\sum_{i=1}^n f(x_i) \right\}. \tag{14.60}$$

For many choices of the underlying function class $\mathscr{F}$, this estimator can be computed in closed form. Let us consider some examples to illustrate.

Example 14.23 (Density estimation via series expansion) This is a follow-up on Example 13.14, where we considered the use of series expansion for regression. Here we consider the use of such expansions for density estimation—say, for concreteness, of univariate densities supported on $[0, 1]$. For a given integer $T \geq 1$, consider a collection of functions $\{\phi_m\}_{m=1}^T$, taken to be orthogonal in $L^2[0, 1]$, and consider the linear function class

$$\mathscr{F}_{\text{ortho}}(T) := \Big\{ f = \sum_{m=1}^T \beta_m \phi_m \mid \beta \in \mathbb{R}^T, \ \beta_1 = 1 \Big\}. \tag{14.61}$$

As one concrete example, we might define the indicator functions

$$\phi_m(x) = \begin{cases} 1 & \text{if } x \in (m-1, m]/T, \\ 0 & \text{otherwise.} \end{cases} \tag{14.62}$$

With this choice, an expansion of the form $f = \sum_{m=1}^T \beta_m \phi_m(T)$ yields a piecewise constant function that is non-negative and integrates to 1. When used for density estimation, it is known as a *histogram estimate*, and is perhaps the simplest type of density estimate.

Another example is given by truncating the Fourier basis previously described in Example 13.15. In this case, since the first function $\phi_1(x) = 1$ for all $x \in [0, 1]$ and the remaining functions are orthogonal, we are guaranteed that the function expansion integrates to one. The resulting density estimate is known as a *projected Fourier-series estimate*. A minor point is that, since the sinusoidal functions are not non-negative, it is possible that the projected Fourier-series density estimate could take negative values; this concern could be alleviated by projecting the function values back onto the orthant.

For the function class $\mathscr{F}_{\text{ortho}}(T)$, the density estimate (14.60) is straightforward to compute: some calculation shows that

$$\widehat{f_T} = \sum_{m=1}^T \widehat{\beta}_m \phi_m, \qquad \text{where } \widehat{\beta}_m = \frac{1}{n}\sum_{i=1}^n \phi_m(x_i). \tag{14.63}$$

For example, when using the histogram basis (14.62), the coefficient $\widehat{\beta}_m$ corresponds to the fraction of samples that fall into the interval $(m-1, m]/T$. When using a Fourier basis expansion, the estimate $\widehat{\beta}_m$ corresponds to an empirical Fourier-series coefficient. In either case, the estimate $\widehat{f_T}$ is easy to compute.

Figure 14.1 shows plots of histogram estimates of a Gaussian density $N(1/2, (0.15)^2)$, with the plots in Figure 14.1(a) and (b) corresponding to sample sizes $n = 100$ and $n = 2000$,

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

respectively. In addition to the true density in light gray, each plot shows the histogram estimate for $T \in \{5, 20\}$. By construction, each histogram estimate is piecewise constant, and the parameter T determines the length of the pieces, and hence how quickly the estimate varies. For sample size $n = 100$, the estimate with $T = 20$ illustrates the phenomenon of overfitting, whereas for $n = 2000$, the estimate with $T = 5$ leads to oversmoothing.

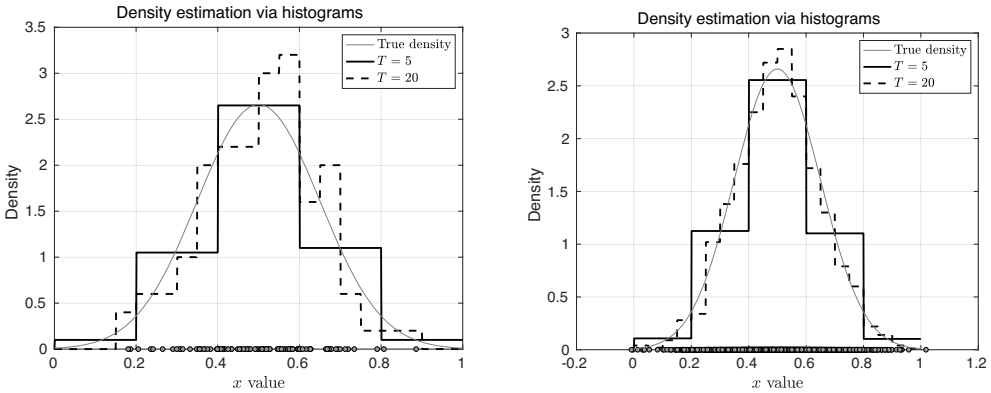


Figure 14.1 Plots of the behavior of the histogram density estimate. Each plot shows the true function (in this case, a Gaussian distribution $\mathcal{N}(1/2, (0.15)^2)$) in light gray and two density estimates using $T = 5$ bins (solid line) and $T = 20$ bins (dashed line). (a) Estimates based on $n = 100$ samples. (b) Estimates based on $n = 2000$ samples.

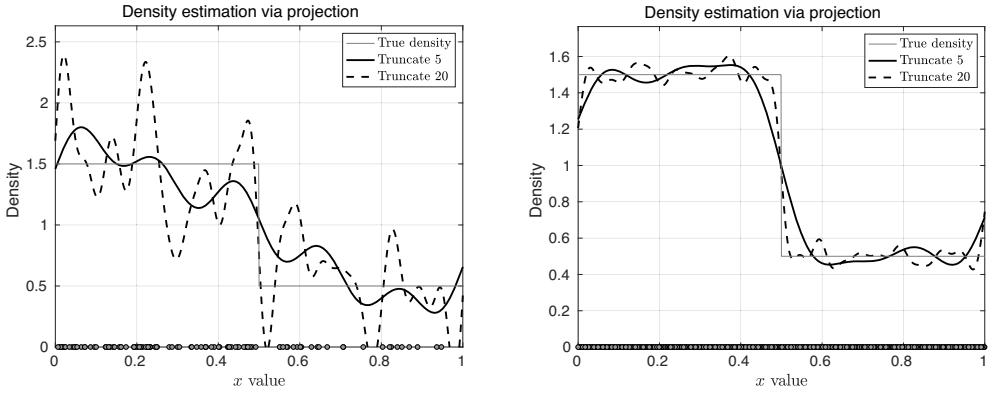
Figure 14.2 shows some plots of the Fourier-series estimator for estimating the density

$$f^*(x) = \begin{cases} 3/2 & \text{for } x \in [0, 1/2], \\ 1/2 & \text{for } x \in (1/2, 1]. \end{cases} \tag{14.64}$$

As in Figure 14.1, the plots in Figure 14.2(a) and (b) are for sample sizes $n = 100$ and $n = 2000$, respectively, with the true density f^* shown in a gray line. The solid and dashed lines show the truncated Fourier-series estimator with $T = 5$ and $T = 20$ coefficients, respectively. Again, we see overfitting by the estimator with $T = 20$ coefficients when the sample size is small ($n = 100$). For the larger sample size ($n = 2000$), the estimator with $T = 20$ is more accurate than the $T = 5$ estimator, which suffers from oversmoothing. ♣

Having considered some examples of the density estimate (14.60), let us now state a theoretical guarantee on its behavior. As with our earlier results, this guarantee applies to the estimate based on a star-shaped class of densities $\mathscr{F}$, which we assume to be uniformly bounded by some b. Recalling that $\bar{\mathcal{R}}_n$ denotes the (localized) Rademacher complexity, we let $\delta_n > 0$ be any positive solution to the inequality $\bar{\mathcal{R}}_n(\delta; \mathscr{F}) \leq \frac{\delta^2}{b}$.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press



(a) (b)

Figure 14.2 Plots of the behavior of the orthogonal series density estimate (14.63) using Fourier series as the orthonormal basis. Each plot shows the true function f^* from equation (14.64) in light gray, and two density estimates for $T = 5$ (solid line) and $T = 20$ (dashed line). (a) Estimates based on $n = 100$ samples. (b) Estimates based on $n = 2000$ samples.

Corollary 14.24 *There are universal constants c_j, $j = 0, 1, 2, 3$, such that for any density f^* uniformly bounded by b, the density estimate* (14.60) *satisfies the oracle inequality*

$$\|\widehat{f} - f^*\|_2^2 \le c_0 \inf_{f \in \mathcal{F}} \|f - f^*\|_2^2 + c_1 \delta_n^2 \tag{14.65}$$

with probability at least $1 - c_2 e^{-c_3 n \delta_n^2}$.

The proof of this result is very similar to our oracle inequality for nonparametric regression (Theorem 13.13). Accordingly, we leave the details as an exercise for the reader.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

14.5 Appendix: Population and empirical Rademacher complexities

Let $\delta_n > 0$ and $\hat{\delta}_n > 0$ be the smallest positive solutions to the inequalities $\bar{\mathcal{R}}_n(\delta_n) \leq \delta_n^2$ and $\widehat{\mathcal{R}}_n(\hat{\delta}_n) \leq \hat{\delta}_n^2$, respectively. Note that these inequalities correspond to our previous definitions (14.4) and (14.7), with $b = 1$. (The general case $b \neq 1$ can be recovered by a rescaling argument.) In this appendix, we show that these quantities satisfy a useful sandwich relation:

Proposition 14.25 *For any* 1*-bounded and star-shaped function class* $\mathscr{F}$*, the population and empirical radii satisfy the sandwich relation*

$$\frac{\delta_n}{4} \overset{\text{(i)}}{\leq} \hat{\delta}_n \overset{\text{(ii)}}{\leq} 3\delta_n, \tag{14.66}$$

with probability at least $1 - c_1 e^{-c_2 n \delta_n^2}$.

Proof For each $t > 0$, let us define the random variable

$$\bar{Z}_n(t) := \mathbb{E}_\epsilon\Big[\sup_{\substack{f \in \mathscr{F} \\ \|f\|_2 \leq t}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big],$$

so that $\bar{\mathcal{R}}_n(t) = \mathbb{E}_x[\bar{Z}_n(t)]$ by construction. Define the events

$$\mathcal{E}_0(t) := \Big\{\big|\bar{Z}_n(t) - \bar{\mathcal{R}}_n(t)\big| \leq \frac{\delta_n t}{8}\Big\} \quad \text{and} \quad \mathcal{E}_1 := \Big\{\sup_{f \in \mathscr{F}} \frac{\big|\|f\|_n^2 - \|f\|_2^2\big|}{\|f\|_2^2 + \delta_n^2} \leq \frac{1}{2}\Big\}.$$

Note that, conditioned on $\mathcal{E}_1$, we have

$$\|f\|_n \leq \sqrt{\tfrac{3}{2}\|f\|_2^2 + \tfrac{1}{2}\delta_n^2} \leq 2\|f\|_2 + \delta_n \tag{14.67a}$$

and

$$\|f\|_2 \leq \sqrt{2\|f\|_n^2 + \delta_n^2} \leq 2\|f\|_n + \delta_n, \tag{14.67b}$$

where both inequalities hold for all $f \in \mathscr{F}$. Consequently, conditioned on $\mathcal{E}_1$, we have

$$\bar{Z}_n(t) \leq \mathbb{E}_\epsilon\left[\sup_{\substack{f \in \mathscr{F} \\ \|f\|_n \leq 2t + \delta_n}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\right] = \widehat{\mathcal{R}}_n(2t + \delta_n) \tag{14.68a}$$

and

$$\widehat{\mathcal{R}}_n(t) \leq \bar{Z}_n(2t + \delta_n). \tag{14.68b}$$

Equipped with these inequalities, we now proceed to prove our claims.

Proof of upper bound (ii) in (14.66)*:* Conditioned on the events $\mathcal{E}_0(7\delta_n)$ and $\mathcal{E}_1$, we have

$$\widehat{\mathcal{R}}_n(3\delta_n) \overset{\text{(i)}}{\leq} \bar{Z}_n(7\delta_n) \overset{\text{(ii)}}{\leq} \mathcal{R}_n(7\delta_n) + \tfrac{7}{8}\delta_n^2,$$

where step (i) follows from inequality (14.68b) with $t = 3\delta_n$, and step (ii) follows from

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

$\mathcal{E}_0(7\delta_n)$. Since $7\delta_n \geq \delta_n$, the argument used to establish the bound (14.19) guarantees that $\mathcal{R}_n(7\delta_n) \leq 7\delta_n^2$. Putting together the pieces, we have proved that

$$\widehat{\mathcal{R}}_n(3\delta_n) \leq 8\delta_n^2 < (3\delta_n)^2.$$

By definition, the quantity $\hat{\delta}_n$ is the smallest positive number satisfying this inequality, so that we conclude that $\hat{\delta}_n \leq 3\delta_n$, as claimed.

Proof of lower bound (i) in (14.66)*:* Conditioning on the events $\mathcal{E}_0(\delta_n)$ and $\mathcal{E}_1$, we have

$$\delta_n^2 = \bar{\mathcal{R}}_n(\delta_n) \overset{\text{(i)}}{\leq} \bar{Z}_n(\delta_n) + \tfrac{1}{8}\delta_n^2 \overset{\text{(ii)}}{\leq} \widehat{\mathcal{R}}_n(3\delta_n) + \tfrac{1}{8}\delta_n^2 \overset{\text{(iii)}}{\leq} 3\delta_n\hat{\delta}_n + \tfrac{1}{8}\delta_n^2,$$

where step (i) follows $\mathcal{E}_0(\delta_n)$, step (ii) follows from inequality (14.68a) with $t = \delta_n$, and step (iii) follows from the same argument leading to equation (14.19). Rearranging yields that $\frac{7}{8}\delta_n^2 \leq 3\delta_n\hat{\delta}_n$, which implies that $\hat{\delta}_n \geq \delta_n/4$.

Bounding the probabilities of $\mathcal{E}_0(t)$ *and* $\mathcal{E}_1$*:* On one hand, Theorem 14.1 implies that $\mathbb{P}[\mathcal{E}_1^c] \leq c_1 e^{-c_2 n\delta_n^2}$.

Otherwise, we need to bound the probability $\mathbb{P}[\mathcal{E}_0^c(\alpha\delta_n)]$ for an arbitrary constant $\alpha \geq 1$. In particular, our proof requires control for the choices $\alpha = 1$ and $\alpha = 7$. From theorem 16 of Bousquet et al. (2003), we have

$$\mathbb{P}[\mathcal{E}_0^c(\alpha\delta_n)] = \mathbb{P}\Big[\big|\bar{Z}_n(\alpha\delta_n) - \bar{\mathcal{R}}_n(\alpha\delta_n)\big| \geq \frac{\alpha\delta_n^2}{8}\Big] \leq 2\exp\Big(-\frac{1}{64}\,\frac{n\alpha\delta_n^4}{2\bar{\mathcal{R}}_n(\alpha\delta_n) + \frac{\alpha\delta_n^2}{12}}\Big).$$

For any $\alpha \geq 1$, we have $\bar{\mathcal{R}}_n(\alpha\delta_n) \geq \bar{\mathcal{R}}_n(\delta_n) = \delta_n^2$, whence $\mathbb{P}[\mathcal{E}_0^c(\alpha\delta_n)] \leq 2e^{-c_2 n\delta_n^2}$. □

14.6 Bibliographic details and background

The localized forms of the Rademacher and Gaussian complexities used in this chapter are standard objects in mathematical statistics (Koltchinskii, 2001, 2006; Bartlett et al., 2005). Localized entropy integrals, such as the one underlying Corollary 14.3, were introduced by van de Geer (2000). The two-sided results given in Section 14.1 are based on b-uniform boundedness conditions on the functions. This assumption, common in much of non-asymptotic empirical process theory, allows for the use of standard concentration inequalities for empirical processes (e.g., Theorem 3.27) and the Ledoux–Talagrand contraction inequality (5.61). For certain classes of unbounded functions, two-sided bounds can also be obtained based on sub-Gaussian and/or sub-exponential tail conditions; for instance, see the papers (Mendelson et al., 2007; Adamczak, 2008; Adamczak et al., 2010; Mendelson, 2010) for results of this type. One-sided uniform laws related to Theorem 14.12 have been proved by various authors (Raskutti et al., 2012; Oliveira, 2013; Mendelson, 2015). The proof given here is based on a truncation argument.

Results on the localized Rademacher complexities, as stated in Corollary 14.5, can be found in Mendelson (2002). The class of additive regression models from Example 14.11 were introduced by Stone (1985), and have been studied in great depth (e.g., Hastie and Tibshirani, 1986; Buja et al., 1989). An interesting extension is the class of sparse additive models, in which the function f is restricted to have a decomposition using at most $s \ll d$

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

univariate functions; such models have been the focus of more recent study (e.g., Meier et al., 2009; Ravikumar et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012).

The support vector machine from Example 14.19 is a popular method for classification introduced by Boser et al. (1992); see the book by Steinwart and Christmann (2008) for further details. The problem of density estimation treated briefly in Section 14.4 has been the subject of intensive study; we refer the reader to the books (Devroye and Györfi, 1986; Silverman, 1986; Scott, 1992; Eggermont and LaRiccia, 2001) and references therein for more details. Good and Gaskins (1971) proposed a roughness-penalized form of the nonparametric maximum likelihood estimate; see Geman and Hwang (1982) and Silverman (1982) for analysis of this and some related estimators. We analyzed the constrained form of the nonparametric MLE under the simplifying assumption that the true density f^* belongs to the density class $\mathscr{F}$. In practice, this assumption may not be satisfied, and there would be an additional form of approximation error in the analysis, as in the oracle inequalities discussed in Chapter 13.

14.7 Exercises

Exercise 14.1 (Bounding the Lipschitz constant) In the setting of Proposition 14.25, show that $\mathbb{E}\big[\sup_{\|f\|_2 \leq t} \|f\|_n\big] \leq \sqrt{5}t$ for all $t \geq \delta_n$.

Exercise 14.2 (Properties of local Rademacher complexity) Recall the localized Rademacher complexity

$$\bar{\mathcal{R}}_n(\delta) := \mathbb{E}_{x,\varepsilon}\Big[\sup_{\substack{f \in \mathscr{F} \\ \|f\|_2 \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big],$$

and let δ_n be the smallest positive solution to the inequality $\bar{\mathcal{R}}_n(\delta) \leq \delta^2$. Assume that function class $\mathscr{F}$ is star-shaped around the origin (so that $f \in \mathscr{F}$ implies $\alpha f \in \mathscr{F}$ for all $\alpha \in [0,1]$).

(a) Show that $\bar{\mathcal{R}}_n(s) \leq \max\{\delta_n^2, s\delta_n\}$. (*Hint:* Lemma 13.6 could be useful.)

(b) For some constant $C \geq 1$, let $t_n > 0$ be the small positive solution to the inequality $\bar{\mathcal{R}}_n(t) \leq Ct^2$. Show that $t_n \leq \frac{\delta_n}{\sqrt{C}}$. (*Hint*: Part (a) could be useful.)

Exercise 14.3 (Sharper rates via entropy integrals) In the setting of Example 14.2, show that there is a universal constant c' such that

$$\mathbb{E}_\varepsilon\Big[\sup_{\substack{f_\theta \in \mathcal{P}_2 \\ \|f_\theta\|_2 \leq \delta}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big] \leq c' \sqrt{\frac{1}{n}}.$$

Exercise 14.4 (Uniform laws for kernel classes) In this exercise, we work through the proof of the bound (14.14a) from Corollary 14.5.

(a) Letting $(\phi_j)_{j=1}^\infty$ be the eigenfunctions of the kernel operator, show that

$$\sup_{\substack{\|f\|_{\mathbb{H}} \leq 1 \\ \|f\|_2 \leq \delta}} \Big|\sum_{i=1}^n \varepsilon_i f(x_i)\Big| = \sup_{\theta \in \mathbb{K}} \Big|\sum_{j=1}^\infty \theta_j z_j\Big|,$$

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

where $z_j := \sum_{i=1}^n \varepsilon_i \phi_j(x_i)$ and

$$\mathcal{D} := \Big\{(\theta)_{j=1}^\infty \mid \sum_{j=1}^\infty \theta_j^2 \le \delta,\ \sum_{j=1}^\infty \frac{\theta_j^2}{\mu_j} \le 1\Big\}.$$

(b) Defining the sequence $\eta_j = \min\{\delta^2, \mu_j\}$ for $j = 1, 2, \ldots$, show that $\mathcal{D}$ is contained within the ellipse $\mathcal{E} := \{(\theta)_{j=1}^\infty \mid \sum_{j=1}^\infty \theta_j^2/\eta_j \le 2\}$.

(c) Use parts (a) and (b) to show that

$$\mathbb{E}_{\varepsilon,x}\Big[\sup_{\substack{\|f\|_{\mathbb{H}} \le 1 \\ \|f\|_2 \le \delta}} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\Big|\Big] \le \sqrt{\frac{2}{n}}\sqrt{\sum_{j=1}^\infty \min\{\delta^2, \mu_j\}}.$$

Exercise 14.5 (Empirical approximations of kernel integral operators) Let $\mathcal{K}$ be a PSD kernel function satisfying the conditions of Mercer's theorem (Theorem 12.20), and define the associated representer $R_x(\cdot) = \mathcal{K}(\cdot, x)$. Letting $\mathbb{H}$ be the associated reproducing kernel Hilbert space, consider the integral operator $T_{\mathcal{K}}$ as defined in equation (12.11a).

(a) Letting $\{x_i\}_{i=1}^n$ denote i.i.d. samples from $\mathbb{P}$, define the random linear operator $\widehat{T}_{\mathcal{K}}: \mathbb{H} \to \mathbb{H}$ via

$$f \mapsto \widehat{T}_{\mathcal{K}}(f) := \frac{1}{n}\sum_{i=1}^n [R_{x_i} \otimes R_{x_i}](f) \ = \ \frac{1}{n}\sum_{i=1}^n f(x_i) R_{x_i}.$$

Show that $\mathbb{E}[\widehat{T}_{\mathcal{K}}] = T_{\mathcal{K}}$.

(b) Use techniques from this chapter to bound the operator norm

$$|||\widehat{T}_{\mathcal{K}} - T_{\mathcal{K}}|||_{\mathbb{H}} := \sup_{\|f\|_{\mathbb{H}} \le 1} \|(\widehat{T}_{\mathcal{K}} - T_{\mathcal{K}})(f)\|_{\mathbb{H}}.$$

(c) Letting ϕ_j denote the jth eigenfunction of $T_{\mathcal{K}}$, with associated eigenvalue $\mu_j > 0$, show that

$$\|\widehat{T}_{\mathcal{K}}(\phi_j) - \mu_j \phi_j\|_{\mathbb{H}} \le \frac{|||\widehat{T}_{\mathcal{K}} - T_{\mathcal{K}}|||_{\mathbb{H}}}{\mu_j}.$$

Exercise 14.6 (Linear functions and four-way independence) Recall the class $\mathscr{F}_{\text{lin}}$ of linear functions from Example 14.10. Consider a random vector $x \in \mathbb{R}^d$ with four-way independent components—i.e., the variables (x_j, x_k, x_ℓ, x_m) are independent for all distinct quadruples of indices. Assume, moreover, that each component has mean zero and variance one, and that $\mathbb{E}[x_j^4] \le B$. Show that the strong moment condition (14.22b) is satisfied with $C = B + 6$.

Exercise 14.7 (Uniform laws and sparse eigenvalues) In this exercise, we explore the use of Theorem 14.12 for bounding sparse restricted eigenvalues (see Chapter 7). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a random matrix with i.i.d. $\mathcal{N}(0, \mathbf{\Sigma})$ rows. For a given parameter $s > 0$, define the function class $\mathscr{F}_{\text{spcone}} = \{f_\theta \mid \|\theta\|_1 \le \sqrt{s}\|\theta\|_2\}$, where $f_\theta(x) = \langle x, \theta \rangle$. Letting $\rho^2(\mathbf{\Sigma})$ denote the maximal diagonal entry of $\mathbf{\Sigma}$, show that, as long as

$$n > c_0 \frac{\rho^2(\mathbf{\Sigma})}{\gamma_{\min}(\mathbf{\Sigma})}\, s \log(\frac{ed}{s})$$

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press

for a sufficiently large constant c, then we are guaranteed that

$$\underbrace{\|f_\theta\|_n^2}_{\|\mathbf{X}\theta\|_2^2/n} \geq \frac{1}{2} \underbrace{\|f_\theta\|_2^2}_{\|\sqrt{\mathbf{\Sigma}}\theta\|_2^2} \qquad \text{for all } f_\theta \in \mathscr{F}_{\text{spcone}}$$

with probability at least $1 - e^{-c_1 n}$. Thus, we have proved a somewhat sharper version of Theorem 7.16. (*Hint:* Exercise 7.15 could be useful to you.)

Exercise 14.8 (Estimation of nonparametric additive models) Recall from Example 14.11 the class $\mathscr{F}_{\text{add}}$ of additive models formed by some base class $\mathscr{G}$ that is convex and 1-uniformly bounded ($\|g\|_\infty \leq 1$ for all $g \in \mathscr{G}$). Let δ_n be the smallest positive solution to the inequality $\bar{\mathcal{R}}_n(\delta; \mathscr{F}) \leq \delta^2$. Letting ϵ_n be the smallest positive solution to the inequality $\bar{\mathcal{R}}_n(\epsilon; \mathscr{G}) \leq \epsilon^2$, show that $\delta_n^2 \precsim d\,\epsilon_n^2$.

Exercise 14.9 (Nonparametric maximum likelihood) Consider the nonparametric density estimate (14.56) over the class of all differentiable densities. Show that the minimum is not achieved. (*Hint:* Consider a sequence of differentiable approximations to the density function placing mass $1/n$ at each of the data points.)

Exercise 14.10 (Hellinger distance and Kullback–Leibler divergence) Prove the lower bound (14.57b) on the Kullback–Leibler divergence in terms of the squared Hellinger distance.

Exercise 14.11 (Bounds on histogram density estimation) Recall the histogram estimator defined by the basis (14.62), and suppose that we apply it to estimate a density f^* on the unit interval $[0, 1]$ that is differentiable with $\|f'\|_\infty \leq 1$. Use the oracle inequality from Corollary 14.24 to show that there is a universal constant c such that

$$\|\widehat{f} - f^*\|_2^2 \leq c n^{-2/3} \tag{14.69}$$

with high probability.

https://doi.org/10.1017/9781108627771.014 Published online by Cambridge University Press