中国科学技术大学
University of Science and Technology of China

Localization and uniform laws

2022 年 12 月 11 日

We begin our exploration with a detailed study of the relation between the population and the usual $L^2(\mathbb{P})$-norm is given by

$$\|f\|^2_{L^2(\mathbb{P})} := \int_X f^2(x)\mathbb{P}(dx) = \mathbb{E}\left[f^2(X)\right] = \|f\|_2^2$$

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

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

It induces the empirical L^2-norm

$$\|f\|^2_{L^2(\mathbb{P}_n)} := \frac{1}{n}\sum_{i=1}^n f^2(x_i) = \int_X f^2(x)\mathbb{P}_n(dx) = \|f\|_n^2$$

Since each $x_i \sim \mathbb{P}$, the linearity of expectation guarantees that

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

Consequently, the law of large numbers implies that $\|f\|_n^2$ converges to $\|f\|_2^2$. For instance, if the function f is uniformly bounded, that is, if

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

then Hoeffding's inequality (Proposition 2.5 Bernstein-type bound) implies that

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

We begin by stating a theorem that controls the deviations in the random variable $|\|f\|_n - \|f\|_2|$.
For a given radius $\delta > 0$ and function class $\mathscr{F}$, consider the localized population Rademacher complexity

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

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 $\{x_i\}_{i=1}^n$.

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

- (A1)$\overline{\mathcal{R}}_n(\delta;\mathscr{F}) \le \frac{\delta^2}{b}$.
- (A2)for any $f \in \mathscr{F}$ and scalar $\alpha \in [0,1]$, the function αf also belongs to $\mathscr{F}$.
- (A3)there is a constant $b < \infty$ such that $\|f\|_\infty \le b$ for all $f \in \mathscr{F}$.

(1) for any $t \ge \delta_n$, we have

$$\left|\|f\|_n^2 - \|f\|_2^2\right| \le \frac{1}{2}\|f\|_2^2 + \frac{t^2}{2} \quad \text{for all } f \in \mathscr{F}$$

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

$$\left|\|f\|_n - \|f\|_2\right| \le c_0\delta_n \quad \text{for all } f \in \mathscr{F}$$

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

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

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

And we shall use it to prove a version of the theorem with $c_0 = 1$.

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

indexed by $r \in (0, 1]$. We also define the auxiliary events

$$\mathcal{A}_0(r) := \left\{ Z_n(r) \geq r^2/2 \right\}$$

and

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

We let $\mathcal{E}_0$ and $\mathcal{E}_1$, respectively, denote the complement of events that inequality (1) or inequality (2) are violated.

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

$$\mathcal{E}_0 \subseteq \mathcal{A}_0(t) \quad \text{and} \quad \mathcal{E}_1 \subseteq \mathcal{A}_0(\delta_n) \cup \mathcal{A}_1(\delta_n)$$

Proof Beginning with the inclusion (i), we divide the analysis into two cases. First, suppose that there exists some function with norm $\|f\|_2 \leq t$. For this function, we must have $\left|\|f\|_n^2 - \|f\|_2^2\right| > \frac{t^2}{2}$, showing that $Z_n(t) > \frac{t^2}{2}$. Otherwise, suppose that the inequality (1) 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.

Turning to the inclusion (ii), it is equivalent to show that $\mathcal{A}_0^c(\delta_n) \cap \mathcal{A}_1^c(\delta_n) \subseteq \mathcal{E}_1^c$. 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 $|\|f\|_n - \|f\|_2| \leq \delta_n$, or we have $\|f\|_n \geq \delta_n$, in which case

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

- Case 2: Next consider a function $f \in \mathscr{F}$ with $\|f\|_2 > \delta_n$. In this case, on the complement of $\mathcal{A}_1$, we have

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

In order to control the events $\mathcal{A}_0(r)$ and $\mathcal{A}_1(r)$, 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

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

Proof

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

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 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

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

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 $g = f^2 - \mathbb{E}\left[f^2(X)\right]$,

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

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

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

- Setting $r = s = \delta_n$ in the bound yields

$$\mathbb{P}\left[\mathcal{A}_0\left(\delta_n\right)\right] \leq e^{-c_1 n \delta_n^2},$$

- This difficulty can be addressed by using a so-called "peeling" argument. For $m = 1, 2, \ldots$, define the events

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

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 $M \leq 4\log\left(1/\delta_n\right)$.
By the union bound, we have $\mathbb{P}\left(\mathcal{A}_1\right) \leq \sum_{m=1}^{M} \mathbb{P}\left(\mathcal{A}_1 \cap \mathcal{S}_m\right)$. Now if the event $\mathcal{A}_1 \cap \mathcal{S}_m$ occurs, such that

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

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

$$\mathbb{P}\left[\mathcal{A}_1\right] \leq \sum_{m=1}^{M} e^{-c_2 n \delta_n^2} \leq e^{-c_2 n \delta_n^2 + \log M} \leq e^{-\frac{c2 n \delta_n^2}{2}},$$

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

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

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

$$\frac{\mathcal{G}_n\left(\delta;\mathscr{F}^*\right)}{\delta} \leq \frac{\delta}{2\sigma}$$

Corollary 14.3 Let $N_n\left(t;\mathbb{B}_n(\delta;\mathscr{F})\right)$ 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 $L^2\left(\mathbb{P}_n\right)$-norm. Then the empirical version of critical inequality is satisfied for any $\delta > 0$ such that

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

This result is essentially identical to Corollary 13.7.

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 is upper bounded as

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

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 is upper bounded as

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

Lemma 13.22 yields the claim.

A potentially limiting aspect of Theorem 14.1 is that it requires the underlying function class to be b-uniformly bounded.
Concretely, in the current context, for any fixed function $f \in \mathscr{F}$, applying the lower tail bound in the chapter 2 to the i.i.d. sequence $\{f(x_i)\}_{i=1}^n$ yields the guarantee

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

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

$$\mathbb{E}\left[f^4(x)\right] \leq C^2 \mathbb{E}\left[f^2(x)\right] \quad \text{for all } f \in \mathscr{F} \text{ with } \|f\|_2 \leq 1.$$

It is certainly implied by the global condition

$$\mathbb{E}\left[f^4(x)\right] \leq C^2 \left(\mathbb{E}\left[f^2(x)\right]\right)^2 \quad \text{for all } f \in \mathscr{F}.$$

Theorem 14.12 Consider a star-shaped class F of functions, each zero-mean under $\mathbb{P}$, and such that the fourth-moment condition holds uniformly over $\mathscr{F}$.

- (A1) $\frac{\overline{\mathcal{R}}_n(\delta;\mathscr{F})}{\delta} \leq \frac{\delta}{128C}$, where the constant C appears in (A2).
- (A2) $\mathbb{E}\left[f^4(x)\right] \leq C^2\mathbb{E}\left[f^2(x)\right]$ for all $f \in \mathscr{F}$ with $\|f\|_2 \leq 1$.

Suppose that the sample size n is large enough to ensure that there is a solution $\delta_n \leq 1$ to the inequality. Then for any $\delta \in [\delta_n, 1]$,we have

$$\|f\|_n^2 \geq \frac{1}{2}\|f\|_2^2 \quad \text{for all } f \in \mathscr{F}\backslash\mathbb{B}_2(\delta)$$

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

- 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 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

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

For a level $\tau > 0$ to be chosen, consider the truncated quadratic

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

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 $\|f\|_n^2 \geq \|f_\tau\|_n^2$.
The remainder of the proof consists of showing that a suitable choice of truncation level τ ensures that

$$\|f_\tau\|_2^2 \geq \frac{3}{4}\|f\|_2^2 \quad \text{for all } f \in \partial\mathbb{B}_2(\delta)$$

and

$$\mathbb{P}\left[Z_n \geq \frac{1}{4}\delta^2\right] \leq c_1 e^{-c_2 n\delta^2} \quad \text{where } Z_n := \sup_{f \in \partial\mathbb{B}_2(\delta)} \left| \|f_\tau\|_n^2 - \|f_\tau\|_2^2 \right|$$

Let $I[|f(x)| \geq \tau]$ be a zero-one indicator for the event $|f(x)| \geq \tau$, we have

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

where the last step uses the Cauchy-Schwarz inequality. Combining the moment bound with Markov's inequality yields

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

Setting $\tau^2 = 4C^2$ yields the bound $\|f\|_2^2 - \|f_\tau\|_2^2 \leq \frac{1}{4}\|f\|_2^2$.

Beginning with the expectation, a standard symmetrization argument (see Proposition 4.11) guarantees that

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

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 guarantees that

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

Our previous choice $\tau = 2C$ ensures that $\mathbb{E}_x\left[Z_n\right] \leq \frac{1}{8}\delta^2$.

Next we prove an upper tail bound on the random variable Z_n, in particular using Talagrand's theorem for empirical processes. By construction, we have $\left\|f_\tau^2\right\|_\infty \leq \tau^2 = 4C^2$, and uniformly all f with $\|f\|_2 = \delta$.

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

Consequently, Talagrand's inequality implies that

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

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

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 $\mathscr{F}_{\text{lin}} = \left\{ f_\theta \mid \theta \in \mathbb{R}^d \right\}$.
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 for the moments of a Gaussian random vector, we have $\mathbb{E}\left[f_\theta^4(x)\right] = 3\left(\mathbb{E}\left[f_\theta^2(x)\right]\right)^2$, showing that condition holds uniformly with $C^2 = 3$. Note that C does not depend on the variance of $f_\theta(x)$, which can be arbitrarily large.

Example 14.13 (Linear functions and random matrices, continued) Recall the linear function class $\mathscr{F}_{\text{lin}}$ introduced previously in Example 14.10. In particular, supposing that the design vector x has a zero-mean distribution with covariance matrix $\boldsymbol{\Sigma}$.

Writing each $x = \sqrt{\Sigma} w$, where $w \sim \mathcal{N}(0, \Sigma)$. Consequently, by definition of the local Rademacher complexity, we have

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

Note that the random variables $\{\varepsilon_i w_i\}_{i=1}^n$ are i.i.d. and standard Gaussian. Consequently, previous results from Chapter 2 guarantee that $\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i\right\|_2 \leq \sqrt{\frac{d}{n}}$. Putting together the pieces, we conclude that $\delta_n^2 \asymp \frac{d}{n}$.

Therefore, for this particular ensemble, Theorem 14.12 implies that, as long as $n \gtrsim d$, then

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

with high probability. The function $f_\theta(x) = \langle x, \theta \rangle$ has $L^2(\mathbb{P})$-norm

$$\|f_\theta\|_2^2 = \theta^{\mathrm{T}} \mathbb{E}\left[xx^{\mathrm{T}}\right] \theta = \|\sqrt{\Sigma}\theta\|_2^2 \quad \text{for each } f_\theta \in \mathscr{F}.$$

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

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

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

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

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

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

Proof We split the argument into two cases:
Case 1: Suppose that $\delta_n \geq \varepsilon_n$. Consequently, we may apply Theorem 13.5 with $t=\delta_n$ to find that

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

On the other hand, Theorem 14.12 implies that

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

Putting together the pieces yields that

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

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

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

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

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

On one hand, we have

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

where the final inequality follows from Theorem 14.12.
On the other hand, let $I[\mathcal{A}]$ be a zero-one indicator for the event $\mathcal{A} := \{\delta_n < \varepsilon_n\}$. Then applying Theorem 13.5 with $t = \varepsilon_n$ yields

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

Putting together the pieces yields the claim.

A predictor is a function f that maps a covariate $x \in \mathcal{X}$ to a prediction $\hat{y} = f(x) \in \tilde{\mathcal{Y}}$.
The goodness of a predictor f is measured in terms of a cost function $\mathcal{L} : \tilde{\mathcal{Y}} \times \tilde{Y} \to \mathbb{R}$, whose value $\mathcal{L} : \tilde{\mathcal{Y}} \times \tilde{Y}$ corresponds to the cost of predicting $\hat{y} \in \tilde{\mathcal{Y}}$ when the underlying true response is some $y \in \mathcal{Y}$.

Given a collection of n samples $\{(x_i, y_i)\}$, a natural way in which to determine a predictor is by minimizing the empirical cost:

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

The population cost function is

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

Define the function $\mathcal{L}_f : X \times \mathcal{Y} \to \mathbb{R}_+$ via $\mathcal{L}_f(x, y) = \mathcal{L}(f(x), y)$, and let us write

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

. Our goal is thus to understand when a minimizer of the empirical cost is a near minimizer of the population cost.

Uniform law of large numbers

Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. We hope that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right|$$

uniformly converges to zero in probability as $n \to \infty$ over the class $\mathscr{F}$.

Theorem (Glivenko-Cantelli)

For any distribution, the empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF in the uniform norm, meaning that

$$\left\| \widehat{F}_n - F \right\|_\infty \xrightarrow{a.s.} 0.$$

Our question can be understand as **deriving a Glivenko-Cantelli law for the so-called cost class $\{\mathcal{L}_f \mid f \in \mathscr{F}\}$.**

Property of cost function

- For a given constant $L > 0$, we say that the cost function $\mathcal{L}$ is L-Lipschitz in its first argument if

$$|\mathcal{L}(z, y) - \mathcal{L}(z, y)| \leq L|z - \bar{z}|$$

  for all pairs $z, \tilde{z} \in \widetilde{y}$ and $y \in \mathcal{Y}$.

- 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

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

  for all $f \in \mathscr{F}$.

The cost function $\mathcal{L}(z, y) = \frac{1}{2}(y - z)^2$ in nonparametric regression is not globally Lipschitz in general. Consider $y = f^*(x) + \epsilon$ in the special case of bounded noise $|\epsilon \leq c$ for some constant c. If we perform nonparametric regression over a b-uniformly bounded function class $\mathcal{F}$, then for all $f, g \in \mathcal{F}$, we have

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

so that the least squares satisfies the Lipschitz condition with $L = 2b + c$. For any $y \in \mathbb{R}$, the function $z \to \frac{1}{2}(y - z)^2$ is strongly convex with parameter $\gamma = 1$, so that $f \to \mathcal{L}_f$ satisfies the strong convexity condition with $\gamma = 1$.

Let $f^* \in \mathcal{F}$ minimize the population cost function $f \to \mathbb{P}(\mathcal{L}_f)$, and consider the shifted function class.

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

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

$$\bar{R}_n(\delta; \mathcal{F}^*) := \mathbb{E}_{x,\epsilon}\Big[\sup_{g \in \mathcal{F}^*, \|g\|_2 \le \delta} |\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(x_i)|\Big]$$

Uniform law for Lipschitz cost functions

Theorem (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

$$\overline{\mathcal{R}}_n\left(\delta; \mathscr{F}^*\right) \leq \delta^2.$$

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

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

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

Uniform law for Lipschitz cost functions

Theorem (Uniform law for Lipschitz cost functions)

(b) Suppose that the cost function is L-Lipschitz and γ-strongly convex. Then for any function $\widehat{f} \in \mathscr{F}$ such that $\mathbb{P}_n\left(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}\right) \leq 0$, we have

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

and

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

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

Part (a) can be used to guarantee consistency of a procedure that chooses fb F to minimize the empirical cost $f \to \mathbb{P}_n(\mathcal{L}_f)$ over $\mathcal{F}$. For any function class $\mathcal{F}$ with $\|\cdot\|_2$diameter at most D, the inequality implies that

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

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

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

$$Z_n(r) = \sup_{\|f-f^*\|_2 \le r} |\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*} - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*}|, \ r > 0.$$

Lemma For each $r > \delta_n$, $Z_n(r)$ satisfies the tail bound

$$\mathbb{P}[Z_n(r) \ge 8Lr\delta_n + \mu] \le c_1 \exp(-\frac{c_2 n\mu^*}{L^2 r_L^2 \mu})$$

Part (a) We difine the events $\mathcal{E}_0 := \{Z_n(\delta_n) \ge 9L\delta_n^2\}, \mathcal{E}_1 := \{f \in \mathcal{F} \| \mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*} - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*} \ge 10L\delta_n \| f - f^* \|_2$ and $\| f - f^* \|_2 \ge \delta_n\}$.

If there is some function $f \in \mathcal{F}$ that violates the bound, then at least one of the events $\mathcal{E}_0$ or $\mathcal{E}_1$ must occur. Applying Lemma above with u $= L\delta_n^2$ guarantees that

$$\mathbb{P}[\mathcal{E}_0] \leq c_1 \exp^{-c_2 n\delta_n^2},$$
$$\mathbb{P}[\mathcal{E}_1] \leq c_1 \exp^{-c_2' n\delta_n^2}, \text{ valid for all } \delta_n^2 \geq \frac{c}{n}.$$

Part(b) By examing the proof of part (a), we can see that it actually implies that $\|\hat{f} - f^*\|_2 \leq \delta_n$, or

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

If the former one holds, then so does the inequality (1)/ If the latter holds, then, combine with the strong convexity condition, we obtain $\|\hat{f} - f^*\|_2 \leq \frac{10L}{\gamma}$, Which also implies inequality (1). As for the inequality (2), with the fact that $\mathbb{P}_n(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) \leq 0$./

We assume the upper bound $b = 1$. By the lipschitz condition, we have $|\mathcal{L}_f - \mathcal{L}_{f^*}|_\infty \leq L\|f - f^*\|_\infty \leq 2L$. Moreover, we have

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

where the last inequality follows since $\|f - f^*\|_2 \leq r$. Consequently, by theorem 3.27, we have

$$\mathbb{P}[Z_n(r) \geq 2\mathbb{E}[Z_n(r)] \leq 2\mathbb{E}[\sup_{\|f-f^*\|\leq r} |] + \mu] \leq c_1 \exp\{-\frac{c_2 n\mu^2}{L^2 r^2 + L\mu}\}. \quad (3.1)$$

And

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

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?
Some method: Histogram, k-Nearest Neighbor, Kernel density estimation.
In particular, suppose that we fix some base class of densities $\mathcal{F}$, and then maximize the likelihood of the observed samples over this class, (constrained form of MLE)

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \mathbb{P}_n(-\log(f(x))) = \arg\min_{f\in\mathcal{F}}\{-\frac{1}{n}\sum_{i=1}^{n}\log(f(x_i))\} \tag{4.1}$$

To be clear, the class of densities F must be suitably restricted for this estimator to be well defined. As an alternative to constraining the estimate, it also possible to define a regularized form of the nonparametric MLE

Assume the true density f^* is belong to $\mathcal{F}$.We 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

$$H^2(f\|g) := \frac{1}{2}\int_{\mathcal{X}} (\sqrt{f} - \sqrt{g})^2 d\mu. \tag{4.2}$$

Kullback-Leibler (KL) divergence is lower bounded by (a multiple of) the squared Hellinger distance-viz.

$$D(f\|g) \geq 2H^2(f\|g) \tag{4.3}$$

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 $\mathcal{G} = \{g = \sqrt{f} \text{ for some } f \in \mathcal{F}\}$ plays an important role in our analysis. As does the shifted square-root function class $\mathcal{G}^* = \mathcal{G} - \sqrt{f^*}$

Assume that there are positive constants (b, v) such that the square-root density class $\mathcal{G}$ is $\sqrt{b}$-uniformly bounded, and star-shaped around $\sqrt{f^*}$, and moreover that the unknown density $f^* \in \mathcal{F}$ is uniformly lower bounded as

$$^*(x) \geq v > 0 \text{ for all } x \in \mathcal{X}$$

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

$$\bar{R}_n(\delta; \mathcal{G}^* \leq \frac{\delta^2}{\sqrt{b + v}}$$

Corollary

Given a class of densities satisfying the previous onditions, let δ_n be any solution to the critical inequality such that $\delta_n^2 \geq (1 + \frac{b}{v})\frac{1}{n}$. Then the nonparametric density estimate $\hat{f}$ satisfies the ellinger bound

$$H^2(\hat{f} \| f^*) \leq c_0 \delta_n^2$$

with probability greater than $1 - c_1 \exp(-c_2 \frac{v}{b+\gamma} n\delta_n^2)$.

Proof Based on applying Theorem 14.20(b) to the transformed function class

$$H = \{\sqrt{\frac{f + f^*}{2f^*}} | f \in \mathcal{F}$$

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

$$\|h\|_\infty = \|sqrt\frac{f+f^*}{2f^*}\|_\infty \leq \sqrt{\frac{1}{2}(\frac{b}{v}+1} = \frac{1}{\sqrt{2v}}\sqrt{b+v}.$$

Moreover, for anh h, we have $h(x) \geq 1/\sqrt{2}$ for all x and whence the mean value theorem applied to the algorithm, combined with the trianlge inequality, implies

$$|\mathcal{L}_{h(x)} - \mathcal{L}_{\tilde{h}(x)}| \leq \sqrt{2}|h(x) - \tilde{h}(x)|, \text{ for all } x, h$$

showing the cost function is $\sqrt{2}$-Lipschitz. For any h with $h^* := \frac{f^*+f^*}{2f^*} = 1$, we have

$$\|h - h^*\|_2 = \mathbb{E}_{f^*}[\{(\frac{f+f^*}{2f^*})^{1/2} - 1\}^2] = 2H^2(\frac{f+f^*}{2}\|f^*) \leq D(f\|g)$$

which is equivalent to asserting $\mathbb{P}(\mathcal{L}_h - \mathcal{L}_h^*) \geq \|h - h^*\|_2^2$ meaning that the cost function is 2-strongly convex around h^*.

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

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \{ \frac{1}{2} \|f\|_2^2 - \mathbb{P}_n(f) \} = \arg\min_{f \in \mathcal{F}} \{ \frac{1}{2} \|f\|_2^2 - \frac{1}{n} \sum_{i=1}^{2} f(x_i) \}$$

We consider use series expansion for density estimation, 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

$$\mathcal{F}_{ortho}(T) := \{f = \sum_{m=1}^{T} \beta_m \phi_m | \beta \in \mathbb{R}^T, \beta_1 = 1\}. \tag{4.4}$$

As one concrete example, we might define the indicator functions

$$\phi_m(x = \begin{cases} 1 \ x \in (m-1, m]/T \\ 0 \ \text{otherwise} \end{cases} \tag{4.5}$$

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