for any $\delta > 0$, the Klein-Rio version of Talagrand's lower-tail inequality gives

$$e^{-x} \geq \mathbb{P}\left(\tilde{Z} \leq \mathbb{E}\tilde{Z} - \sqrt{2x(n\sigma^2 + 2\mathbb{E}\tilde{Z})} - x\right) \geq \mathbb{P}\left(\tilde{Z} \leq (1-\delta)\mathbb{E}\tilde{Z} - \sqrt{2xn\sigma^2} - \frac{1+\delta}{\delta}x\right).$$

Similarly, using (99),

$$\mathbb{P}\left(Z \geq (1+\delta)\mathbb{E}Z + \sqrt{2xn\sigma^2} + \frac{3+\delta}{3\delta}x\right) \leq e^{-x}.$$

Recall also that $\mathbb{E}[Z] \leq 2\mathbb{E}[\tilde{Z}]$. Then, we have on the intersection of the complement of the events in the last two inequalities, for $\delta = 1/5$ (say),

$$\begin{aligned} Z &< \frac{6}{5}\mathbb{E}[Z] + \sqrt{2xn\sigma^2} + \frac{16}{3}x \leq \frac{12}{5}\mathbb{E}[\tilde{Z}] + \sqrt{2xn\sigma^2} + \frac{16}{3}x \\ &< \frac{12}{5}\left[\frac{5}{4}\tilde{Z} + \frac{5}{4}\sqrt{2xn\sigma^2} + \frac{15}{2}x\right] + \sqrt{2xn\sigma^2} + \frac{16}{3}x \\ &= 3\tilde{Z} + 4\sqrt{2xn\sigma^2} + \frac{70}{3}x; \end{aligned}$$

i.e., this inequality holds with probability $1 - 2e^{-x}$. □

Note that different values of δ produce different coefficients in the above theorem.

8.3 Empirical risk minimization and concentration inequalities

Let $X, X_1, \ldots, X_n, \ldots$ be i.i.d. random variables defined on a probability space and taking values in a measurable space $\mathcal{X}$ with common distribution P. In this section we highlight the usefulness of concentration inequalities, especially Talagrand's inequality, in empirical risk minimization (ERM); see [Koltchinskii, 2011] for a thorough study of this topic.

Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. In what follows, the values of a function $f \in \mathcal{F}$ will be interpreted as "losses" associated with certain "actions" (e.g., $\mathcal{F} = \{f(x) \equiv f(z,y) = (y - \beta^\top z)^2 : \beta \in \mathbb{R}^d\}$ and $X = (Z, Y) \sim P$).

We will be interested in the problem of risk minimization:

response

$$\min_{f \in \mathcal{F}} Pf \tag{102}$$

in the cases when the distribution P is unknown and has to be estimated based on the data $X_1, \ldots, X_n$. Since the empirical measure $\mathbb{P}_n$ is a natural estimator of P, the true risk can be estimated by the corresponding empirical risk, and the risk minimization problem has to be replaced by the *empirical risk minimization* (ERM):

$$\min_{f \in \mathcal{F}} \mathbb{P}_n f. \tag{103}$$

$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \mathbb{P}_n f = \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n l(g(x_i), y_i)$

$\beta' x_i$

As is probably clear by now, many important methods of statistical estimation such as maximum likelihood and more general M-estimation are versions of ERM.

$f_{\mathcal{F}} = \arg\min_{h\in\mathcal{F}} Ph$

Definition 8.12. *The excess risk of $f \in \mathcal{F}$ is defined as*

$$\mathcal{E}(f) \equiv \mathcal{E}_P(f) := Pf - \inf_{h\in\mathcal{F}} Ph. = Pf - Pf_{\mathcal{F}}$$

approximation error

Recall that we have already seen an important application of ERM in the problem of classification in Example 7.10. Here is another important application.

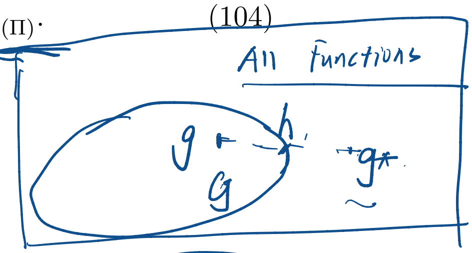
Example 8.13 (Regression)**.** *Suppose that we observe $X_1 \equiv (Z_1, Y_1), \ldots, X_n \equiv (Z_n, Y_n)$ i.i.d. $X \equiv (Z, Y) \sim P$ on $\mathcal{X} \equiv \mathcal{Z} \times T$, $T \subset \mathbb{R}$, and the goal is to study the relationship between Y and Z. We study regression with quadratic loss $\ell(y, u) := (y - u)^2$ given a class of of measurable functions $\mathcal{G}$ from $\mathcal{Z}$ to T; the distribution of Z will be denoted by Π. This problem can be thought of as a special case of ERM with*

$$\mathcal{F} := \{(\ell \bullet g)(z, y) \equiv (y - g(z))^2 : g \in \mathcal{G}\}.$$

Suppose that the true regression function is $g_*(z) := \mathbb{E}[Y|Z = z]$, for $z \in \mathcal{Z}$. In this case, the excess risk of $f(z, y) = (y - g(z))^2 \in \mathcal{F}$ (for some $g \in \mathcal{G}$) is given by[77]

$$\mathcal{E}_P(f) = \mathcal{E}_P(\ell \bullet g) = \underbrace{\|g - g_*\|^2_{L_2(\Pi)}}_{Pf} - \inf_{h\in\mathcal{G}} \|h - g_*\|^2_{L_2(\Pi)}. \tag{104}$$

If $\mathcal{G}$ is such that $g_* \in \mathcal{G}$ then $\mathcal{E}_P(\ell \bullet g) = \|g - g_*\|^2_{L_2(\Pi)}$, for all $g \in \mathcal{G}$.



Let

$$\hat{f} \equiv \hat{f}_n \in \arg\min_{f\in\mathcal{F}} \mathbb{P}_n f$$

be a solution of the ERM problem (103). The function $\hat{f}_n$ is used as an approximation of the solution of the true risk minimization problem (102) and its excess risk $\mathcal{E}_P(\hat{f}_n)$ is a natural measure of accuracy of this approximation.

1. model is conditional distribution
2. loss function is log loss function

It is worth pointing out that a crucial difference between ERM and classical M-estimation, as discussed in Sections 5 and 6, is that in the analysis of ERM we do not (usually) assume that the data generating distribution P belongs to the class of models considered (e.g., $\inf_{h\in\mathcal{F}} Ph$ need not be 0). Moreover, in M-estimation, typically the focus is on recovering a parameter of interest in the model (which is expressed as the population M-estimator) whereas in ERM the focus is mainly on deriving optimal (upper and lower) bounds for the excess risk $\mathcal{E}_P(\hat{f}_n)$.

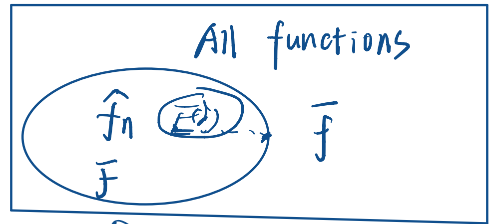
f

It is of interest to find tight upper bounds on the excess risk[78] of $\hat{f}$ that hold with a high probability. Such bounds usually depend on certain "geometric" properties of the function class $\mathcal{F}$ and on various measures of its "complexity" that determine the accuracy of approximation of the true risk Pf by the empirical risk $\mathbb{P}_n f$ in a neighborhood of a proper size of the minimal set of the true risk.

[77] Exercise (HW3): Show this.

[78] Note that we have studied upper bounds on the excess risk in the problem of classification in Example 7.10.

In the following we describe a rather general approach to derivation of such bounds in an abstract framework of ERM. We start with some definitions.



Definition 8.14. *The δ-minimal set of the risk is defined as*

$$\mathcal{F}(\delta) := \{f \in \mathcal{F} : \mathcal{E}_P(f) \le \delta\}.$$

The L_2-diameter of the δ-minimal set is denoted by

$$D(\delta) \equiv D_P(\mathcal{F};\delta) := \sup_{f_1,f_2 \in \mathcal{F}(\delta)} \{P[(f_1 - f_2)^2]\}^{1/2}.$$

$\hat f = \hat f_n \in \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{P}_n f$

$\bar f \in \operatorname{argmin}_{f \in \mathcal{F}} Pf$

Suppose, for simplicity, that the infimum of the risk Pf is attained at $\bar f \in \mathcal{F}$ (the argument can be easily modified if the infimum is not attained in the class). Denote

$$\hat\delta := \mathcal{E}_P(\hat f).$$

$\mathcal{F}(\hat\delta) := \{f \in \mathcal{F} : \mathcal{E}_P(f) \le \mathcal{E}_P(\hat f)\}$

the infimum of the empirical risk

Then $\hat f, \bar f \in \mathcal{F}(\hat\delta)$ and $\mathbb{P}_n \hat f \le \mathbb{P}_n \bar f$. Therefore,

≤ 0

$$\begin{aligned}
\hat\delta &= \mathcal{E}_P(\hat f) = P(\hat f - \bar f) \le \mathbb{P}_n(\hat f - \bar f) + (P - \mathbb{P}_n)(\hat f - \bar f) \\
&\le \sup_{f_1,f_2 \in \mathcal{F}(\hat\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \qquad (105)\\
&\le \sup_{f_1,f_2 \in \mathcal{F}} |(\mathbb{P}_n - P)(f_1 - f_2)|.
\end{aligned}$$

$\le U_n(\hat\delta)$

upper bound of excess risk.

Previously, we had used the last inequality to upper bound the excess risk in classification; see Example 7.10. In this section we will use the implicit characterization of $\hat\delta$ in (105) to improve our upper bound. This naturally leads us to the study of the following (local) measure of empirical approximation:

$$\phi_n(\delta) \equiv \phi_n(\mathcal{F};\delta) := \mathbb{E}\left[\sup_{f_1,f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)|\right]. \qquad (106)$$

Idea: Imagine there exists a nonrandom upper bound

$$\boxed{U_n(\delta) \ge \sup_{f_1,f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)|} \qquad (107)$$

that holds *uniformly* in δ with a high probability. Then, with the same probability, the excess risk $\hat\delta = \mathcal{E}_P(\hat f)$ will be bounded[79] by the largest solution of the inequality

$$\hat\delta \le U_n(\hat\delta) \longrightarrow \delta \le U_n(\delta). \qquad (108)$$

$\delta_n(\mathcal{F})$

By solving the above inequality one can obtain $\delta_n(\mathcal{F})$ (which satisfies (108)) such that $\mathbb{P}(\mathcal{E}_P(\hat f_n) > \delta_n(\mathcal{F}))$ is small[80]. Thus, constructing an upper bound on the excess risk essentially reduces to solving a fixed point inequality of the type $\delta \le U_n(\delta)$.

[79] As $\hat\delta \le \sup_{f_1,f_2 \in \mathcal{F}(\hat\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \le U_n(\hat\delta)$, $\hat\delta$ satisfies inequality (108).

[80] We will formalize this later.

upper bound with a high probability

$V_n = 2uE[Z] + n\sigma^2$.

$P(Z \geq EZ + \sqrt{2V_n x} + ux/3) \leq e^{-x}, \; x \geq 0$.

Let us describe in more detail what we mean by the above intuition. There are many different ways to construct upper bounds on the sup-norm of empirical processes. A very general approach is based on Talagrand's concentration inequalities. For example, if the functions in $\mathcal{F}$ take values in the interval $[0,1]$, then[81] by (99) we have, for $t > 0$,[82] $x=t$. $u=\frac{1}{n}$

Z EZ.

$$\mathbb{P}\left(\sup_{f_1,f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \geq \phi_n(\delta) + \frac{1}{\sqrt{n}}\sqrt{2t\,(2\phi_n(\delta) + D^2(\delta))} + \frac{t}{3n} \right) \leq e^{-t}. \quad (109)$$

Then, using the facts: (i) $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and (ii) $2\sqrt{ab} \leq a/K + Kb$, for any $a, b, K > 0$, we have

$\phi_n(\delta) + D(\delta)\sqrt{\frac{2t}{n}} + \frac{t}{n} + \phi_n(\delta) + \frac{t}{3n} = 2\phi_n(\delta) + D(\delta)\sqrt{\frac{2t}{n}} + \frac{4t}{3n}$

$$\sqrt{2t\,(D^2(\delta) + 2\phi_n(\delta))} \leq \sqrt{2tD^2(\delta)} + 2\sqrt{t\phi_n(\delta)} \leq D(\delta)\sqrt{2t} + \frac{t}{\sqrt{n}} + \sqrt{n}\phi_n(\delta).$$

$\leq 2\phi_n(\delta) + 2D(\delta)\sqrt{\frac{t}{n}} + \frac{2t}{n}$.

Thus, from (109), for all $t > 0$, we have[83]

$$\mathbb{P}\left(\sup_{f_1,f_2 \in \mathcal{F}(\delta)} |(\mathbb{P}_n - P)(f_1 - f_2)| \geq \bar{U}_n(\delta;t) \right) \leq e^{-t} \quad (110)$$

where

$$\bar{U}_n(\delta;t) := 2\left(\phi_n(\delta) + D(\delta)\sqrt{\frac{t}{n}} + \frac{t}{n} \right). \quad (111)$$

This observation provides a way to construct a function $U_n(\delta)$ such that (107) holds with a high probability "uniformly" in δ — by first defining such a function at a discrete set of values of δ and then extending it to all values by monotonicity. We will elaborate on this shortly. Then, by solving the inequality (108) one can construct a bound on $\mathcal{E}_P(\hat{f}_n)$, which holds with "high probability" and which is often of correct order of magnitude.

8.3.1 A formal result on excess risk in ERM

Let us now try to state a formal result in this direction. To simplify notation, assume that the functions in $\mathcal{F}$ take values in $[0,1]$. Let $\{\delta_j\}_{j\geq 0}$ be a decreasing sequence of positive numbers with $\delta_0 = 1$ and let $\{t_j\}_{j\geq 0}$ be a sequence of positive numbers. Define $U_n : (0,\infty) \to \mathbb{R}$, via (111), as

$\delta_j \leq \delta_0 = 1$

$$U_n(\delta) := \bar{U}_n(\delta_j; t_j), \qquad \text{for} \quad \delta \in (\delta_{j+1}, \delta_j], \quad (112)$$

and $U_n(\delta) := U_n(1)$ for $\delta > 1$. Denote

$$\delta_n(\mathcal{F}) := \sup\{\delta \in (0,1] : \delta \leq U_n(\delta)\}. \quad (113)$$

[81] This assumption just simplifies a few mathematical expressions; there is nothing sacred about the interval $[0,1]$, we could have done it for any constant compact interval.

[82] According to the notation of (99), we can take $\sigma^2 = D^2(\delta)$, and then $\nu_n = 2n\phi_n(\mathcal{F};\delta) + nD^2(\delta)$.

[83] This form of the concentration inequality is usually called Bousquet's version of Talagrand's inequality.

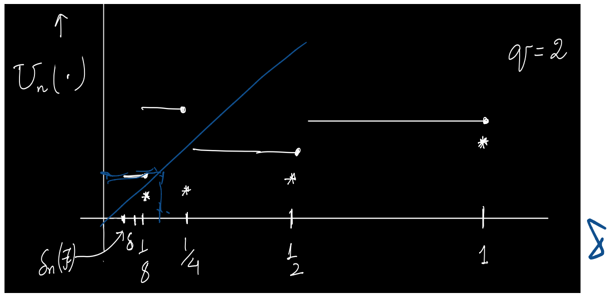


Figure 2: Plot of the piecewise constant function $U_n(\delta)$, for $\delta \geq \delta_n(\mathcal{F})$, along with the value of $\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta_j)}$, for $j = 0, 1, \ldots$, denoted by the $\star$'s.

$\delta \leq U_n(\delta)$

It is easy to check that $\delta_n(\mathcal{F}) \leq U_n(\delta_n(\mathcal{F}))$. Obviously, the definitions of U_n and $\delta_n(\mathcal{F})$ depend on the choice of $\{\delta_j\}_{j\geq 0}$ and $\{t_j\}_{j\geq 0}$ (we will choose specific values of these quantities later on). We start with the following simple inequality that provides a distribution dependent upper bound on the excess risk $\mathcal{E}_P(\hat{f}_n)$.

Theorem 8.15. *For all $\delta \geq \delta_n(\mathcal{F})$,*

$$\mathbb{P}\left(\mathcal{E}_P(\hat{f}_n) > \delta\right) \leq \sum_{j:\delta_j \geq \delta} e^{-t_j}. \tag{114}$$

Proof. It is enough to prove the result for any $\delta > \delta_n(\mathcal{F})$; then the right continuity of the distribution function of $\mathcal{E}_P(\hat{f}_n)$ would lead to the bound (114) for $\delta = \delta_n(\mathcal{F})$.

So, fix $\delta > \delta_n(\mathcal{F})$. Letting $\mathcal{F}'(\delta) := \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}(\delta)\}$, we know that

$$\mathcal{E}_P(\hat{f}) = \hat{\delta} \leq \sup_{f \in \mathcal{F}'(\hat{\delta})} |(\mathbb{P}_n - P)(f)| \equiv \|\mathbb{P}_n - P\|_{\mathcal{F}'(\hat{\delta})}. \tag{115}$$

Denote

$\hat{\delta}$

$$E_{n,j} := \left\{\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta_j)} \leq U_n(\delta_j)\right\}.$$

It follows from Bousquet's version of Talagrand's inequality (see (110)) that $\mathbb{P}(E_{n,j}) \geq 1 - e^{-t_j}$. Let

$$E_n := \cap_{j:\delta_j \geq \delta} E_{n,j}.$$

U_n

Then

$$\mathbb{P}(E_n) = 1 - \mathbb{P}(E_n^c) \geq 1 - \sum_{j:\delta_j \geq \delta} e^{-t_j}. \tag{116}$$

0 high probability

On the event E_n, for all $\sigma \geq \delta$, we have

$U_n(\cdot) = \bar{U}_n(\cdot)$

$$\|\mathbb{P}_n - P\|_{\mathcal{F}'(\sigma)} \leq U_n(\sigma). \tag{117}$$

E_n.

The above holds as: (i) $U_n(\cdot)$ is a piecewise constant function (with possible jumps only at δ_j's), (ii) the function $\sigma \mapsto \|\mathbb{P}_n - P\|_{\mathcal{F}'(\sigma)}$ is monotonically nondecreasing, and (iii) $\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta_j)} \leq U_n(\delta_j)$ on E_n, for j such that $\delta \geq \delta_j$; see Figure 8.3.1.

Claim: $\{\hat{\delta} \geq \delta\} \subset E_n^c$. We prove the claim using the method of contradiction. Thus, suppose that the above claim does not hold. Then, the event $\{\hat{\delta} \geq \delta\} \cap E_n$ is non-empty. On the event $\{\hat{\delta} \geq \delta\} \cap E_n$ we have

$$\hat{\delta} \leq \|\mathbb{P}_n - P\|_{\mathcal{F}'(\hat{\delta})} \leq U_n(\hat{\delta}), \tag{118}$$

where the first inequality follows from (115) and the second inequality holds via (117). This, in particular, implies that

$$\delta \leq \hat{\delta} \leq \delta_n(\mathcal{F}),$$

$\delta_n(\mathcal{F}) := \sup\{\delta \in (0,1] : \delta \leq U_n(\delta)\}$

where the last inequality follows from (118) and the maximality of $\delta_n(\mathcal{F})$ via (113). However the above display contradicts the assumption that $\delta > \delta_n(\mathcal{F})$. Therefore, we must have $\{\hat{\delta} \geq \delta\} \subset E_n^c$.

The claim now implies that $\mathbb{P}(\mathcal{E}_P(\hat{f}_n) \geq \delta) = \mathbb{P}(\hat{\delta} \geq \delta) \leq \mathbb{P}(E_n^c) \leq \sum_{j:\delta_j \geq \delta} e^{-t_j}$, via (116), thereby completing the proof. □

Although Theorem 8.15 yields a high probability bound on the excess risk of $\hat{f}_n$ (i.e., $\mathcal{E}_P(\hat{f}_n)$), we still need to upper bound $\delta_n(\mathcal{F})$ for the result to be useful. We address this next. We start with some notation. Given any $\psi : (0, \infty) \to \mathbb{R}$, denote by

$\delta \leq U_n(\delta)$

$$\psi^\dagger(\sigma) := \sup_{s \geq \sigma} \frac{\psi(s)}{s}. \tag{119}$$

$\delta(\mathcal{F})$.

$U_n(\delta(\mathcal{F}))$

Note that $\psi^\dagger$ is a nonincreasing function[84]. $\sigma_1 < \sigma_2$, $\psi^\dagger(\sigma_1) \geq \psi^\dagger(\sigma_2)$.

The study of $\psi^\dagger$ is naturally motivated by the study of the function $\frac{U_n(\delta)}{\delta}$ and when it crosses the value 1; cf. (113). As $\frac{U_n(\delta)}{\delta}$ may have multiple crossings of 1, we "regularize" $\frac{U_n(\delta)}{\delta}$ by studying $V_n^t(\delta)$ defined below (which can be thought of as a well-behaved monotone version of $U_n^\dagger$). For $q > 1$ an $t > 0$, denote

$\delta(\mathcal{F})$

$U_n = 2\left[\phi_n(\sigma) + \sqrt{D^2(\sigma)\frac{t}{n\sigma}} + \frac{t}{n\sigma}\right]$

$$V_n^t(\sigma) = 2q\left[\phi_n^\dagger(\sigma) + \sqrt{(D^2)^\dagger(\sigma)}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\right], \qquad \text{for } \sigma > 0. \tag{120}$$

Note that V_n^t is a strictly decreasing of σ in $(0, \infty)$. Let

$$\boxed{\sigma_n^t} \equiv \sigma_n^t(\mathcal{F}) := \inf\{\sigma > 0 : V_n^t(\sigma) \leq 1\}. \tag{121}$$

We will show next that $\sigma_n^t \geq \delta_n(\mathcal{F})$ (for a special choice of $\{\delta_j\}_{j\geq 0}$ and $\{t_j\}_{j\geq 0}$) and thus, by (8.15) and some algebraic simplification, we will obtain the following result. Given a concrete application, our goal would be to find upper bounds on σ_n^t; see Section 8.3.2 where we illustrate this technique for finding a high probability bound on the excess risk in bounded regression.

[84] Take $\sigma_1 < \sigma_2$. Then

$$\psi^\dagger(\sigma_1) = \sup_{s \geq \sigma_1} \frac{\psi(s)}{s} \geq \sup_{s \geq \sigma_2} \frac{\psi(s)}{s} = \psi^\dagger(\sigma_2).$$

$V_n^t(\delta_n^t) \leq 1$

$V_n^t(\sigma) \leq 1$

Theorem 8.16 (High probability bound on the excess risk of the ERM)**.** *For all $t > 0$,*

$$\mathbb{P}\left(\mathcal{E}_P(\hat{f}_n) > \sigma_n^t\right) \leq C_q e^{-t}. \tag{122}$$

where $C_q := \frac{q}{q-1} \vee e$. $\max\left\{\frac{q}{q-1}, e\right\}$

Proof. Fix $t > 0$ and let $\sigma > \sigma_n^t$. We will show that $\mathbb{P}\big(\mathcal{E}_P(\hat{f}_n) > \sigma\big) \leq C_q e^{-t}$. Then, by taking a limit as $\sigma \downarrow \sigma_n^t$, we obtain (122).

Define, for $j \geq 0$,

$$\delta_j := q^{-j} \qquad \text{and} \qquad t_j := t\frac{\delta_j}{\sigma}.$$

Recall the definitions of $U_n(\delta)$ and $\delta_n(\mathcal{F})$ (in (112) and (113)) using the above choice of the sequences $\{\delta_j\}_{j\geq 0}$ and $\{t_j\}_{j\geq 0}$. Then, for all $\delta \geq \sigma$, using (112),[85]

$U_n(\delta, t) := 2\left(\phi_n(\delta) + D(\delta)\sqrt{\frac{t}{n}} + \frac{t}{n}\right)$ (111)

$$\begin{aligned}
\frac{U_n(\delta)}{\delta} &= 2\left(\frac{\phi_n(\delta_j)}{\delta} + \frac{D(\delta_j)}{\sqrt{\delta}}\sqrt{\frac{t\delta_j}{\delta\sigma n}} + \frac{t\delta_j}{\delta\sigma n}\right) && \text{if } \delta \in (\delta_{j+1}, \delta_j] \\
&\leq 2q\left(\frac{\phi_n(\delta_j)}{\delta_j} + \frac{D(\delta_j)}{\sqrt{\delta_j}}\sqrt{\frac{t\delta_j}{\delta_j\sigma n}} + \frac{t\delta_j}{\delta_j\sigma n}\right) && \text{as } \delta > \delta_{j+1} = \frac{\delta_j}{q} \Rightarrow \frac{1}{\delta} < \frac{q}{\delta_j} \\
&\leq 2q\left(\sup_{s\geq\sigma}\frac{\phi_n(s)}{s} + \sqrt{\frac{t}{\sigma n}}\sup_{s\geq\sigma}\frac{D(\delta)}{\sqrt{\delta}} + \frac{t}{\sigma n}\right) && \text{as } \delta_j \geq \delta \geq \sigma \\
&= 2q\left(\phi_n^\dagger(\sigma) + \sqrt{(D^2)^\dagger(\sigma)}\sqrt{\frac{t}{\sigma n}} + \frac{t}{\sigma n}\right) = V_n^t(\sigma).
\end{aligned}$$

$q^{-j-1} = \frac{q^{-j}}{q} = \frac{\delta_j}{q}$

non-increasing function

inf

Since $\sigma > \sigma_n^t$ and the function V_n^t is strictly decreasing, we have $V_n^t(\sigma) < V_n^t(\sigma_n^t) \leq 1$, and hence, for all $\delta > \sigma$,

δ_n^t

$$\frac{U_n(\delta)}{\delta} \leq \boxed{V_n^t(\sigma) < 1.} \Rightarrow \delta > U_n(\delta) \geq \delta_n(\mathcal{F})$$

Therefore, $\delta > \delta_n(\mathcal{F}) := \sup\{s > 0 : 1 \leq \frac{U_n(s)}{s}\}$, and thus, $\sigma \geq \delta_n(\mathcal{F})$. Now, from Theorem 8.15 it follows that

$\sigma > V_n^t \geq \delta_n(\mathcal{F})$

$\sigma > \sigma_t$

$$\mathbb{P}\left(\mathcal{E}_P(\hat{f}_n) > \sigma\right) \leq \sum_{j:\delta_j\geq\sigma} e^{-t_j} \leq C_q e^{-t}$$

σ_n^t

$= \sum_{j:\delta_j\geq\sigma} e^{-t\delta_j/\sigma} \leq \sum_j \left(e^{q^j}\right)^{-t}$

where the last step follows from some algebra[86]. □

[85] For $\delta > \delta_0 \equiv 1$, the following sequence of displays also holds with $j = 0$.

[86] Exercise (HW3): Show this. Hint: we can write

$$\sum_{j:\delta_j\geq\sigma} e^{-t_j} = \sum_{j:\delta_j\geq\sigma} e^{-t\delta_j/\sigma} \leq \sum_{j\geq 0} e^{-tq^j} = \cdots \leq \frac{q}{q-1}e^{-t}, \qquad \text{for } t \geq 1.$$

$(e^{-t})^{q^j} \leq e^{-t} \cdot q^{-j}$

for all $\delta > \sigma$, $\frac{U_n(\delta)}{\delta} \leq V_n^t(\sigma)$

now we get $V_n^t(\sigma) < 1 \leq \frac{U_n(\delta_n(\mathcal{F}))}{\delta_n(\mathcal{F})}$

method of contradiction

thus $\sigma \geq \delta_n(\mathcal{F})$

8.3.2 Excess risk in bounded regression

Recall the regression setting in Example 8.13. Given a function $g : \mathcal{Z} \to T$, the quantity $(\ell \bullet g)(z, y) := \ell(y, g(z))$ is interpreted as the loss suffered when $g(z)$ is used to predict y. The problem of optimal prediction can be viewed as a *risk minimization*:

loss

$$\mathbb{E}[\ell(Y, g(Z))] =: P(\ell \bullet g)$$

over $g : \mathcal{Z} \to T$. We start with the regression problem with *bounded response* and with quadratic loss. To be specific, assume that Y takes values in $T = [0, 1]$ and $\ell(y, u) := (y-u)^2$. Suppose that we are given a class of measurable real-valued functions $\mathcal{G}$ on $\mathcal{Z}$. We denote by $\mathcal{F} := \{\ell \bullet g : g \in \mathcal{G}\}$. Suppose that the true regression function is $g_*(z) := \mathbb{E}[Y|Z = z]$, for $z \in \mathcal{Z}$, which is not assumed to be in $\mathcal{G}$. Recall that the *excess risk* $\mathcal{E}_P(\ell \bullet g)$ in this problem is given by (104). $\mathcal{E}_P(f) = \mathcal{E}_P(\ell\cdot g) = \|g-g_*\|^2_{L_2(\Pi)} - \inf_{h\in\mathcal{G}} \|h-g_*\|^2_{L_2(\Pi)}$

In order to apply Theorem 8.16 to find a high probability bound on the excess risk of the ERM $\hat{f} \equiv \ell \bullet \hat{g}$ (see (103)) in this problem, which is determined by σ_n^t via (121), we have to find upper bounds for $V_n^t(\cdot)$ (which in turn depends on the functions $\phi_n^\dagger$ and $\sqrt{(D^2)^\dagger}$).

min $\mathbb{P}_n f$
$f \in \mathcal{F}$

As a first step we relate the excess risk of any $f \equiv \ell \bullet g \in \mathcal{F}$ to $g \in \mathcal{G}$. The following lemma provides an easy way to bound the excess risk of f from below in the case of a *convex class* $\mathcal{G}$, an assumption we make in the sequel.

$\sigma_n^t = \sigma_n^t(\tau) = \inf\{\tau > 0 : V_n^t(\tau) \le 1\}$

Lemma 8.17. *If $\mathcal{G}$ is a convex class of functions, then*

$$2\mathcal{E}_P(\ell \bullet g) \geq \|g - \bar{g}\|^2_{L_2(\Pi)}$$

where $\bar{g} := \operatorname{argmin}_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)}$ is assumed to exist.

Below we make some observations that will be crucial to find σ_n^t. nonrandom upper bound.

1. It follows from Lemma 8.17 that

   $\|g-\bar{g}\|^2_{L_2(\Pi)} \le 2\mathcal{E}_P(\ell\cdot g) \le 2\delta$.

$$\mathcal{F}(\delta) = \{f \in \mathcal{F} : \mathcal{E}_P(f) \leq \delta\} \subset \{\ell \bullet g : g \in \mathcal{G}, \|g - \bar{g}\|^2_{L_2(\Pi)} \leq 2\delta\}. \tag{123}$$

2. For any two functions $g_1, g_2 \in \mathcal{G}$ and all $z \in \mathcal{Z}, y \in [0, 1]$, we have

   $= |y-g_1(z)-y+g_2(z)|\,|y-g_1(z)+y-g_2(z)|$

$$\begin{aligned} |(\ell \bullet g_1)(z, y) - (\ell \bullet g_2)(z, y)| &= |(y - g_1(z))^2 - (y - g_2(z))^2| \\ &= |g_1(z) - g_2(z)|\,|2y - g_1(z) - g_2(z)| \leq 2\,|g_1(z) - g_2(z)|, \end{aligned}$$

   which implies ≤ 2. since $T = [0,1]$

$$P\left[(\ell \bullet g_1 - \ell \bullet g_2)^2\right] \leq 4\|g_1 - g_2\|^2_{L_2(\Pi)}.$$

   Recalling that $D(\delta) := \sup_{f_1, f_2 \in \mathcal{F}(\delta)} \{P[(f_1 - f_2)^2]\}^{1/2}$, we have

$$\begin{aligned} D(\delta) &\leq 2 \sup\left\{\|g_1 - g_2\|_{L_2(\Pi)} : g_k \in \mathcal{G}, \|g_k - \bar{g}\|^2_{L_2(\Pi)} \leq 2\delta \text{ for } k = 1, 2\right\} \\ &\leq \cancel{2(2\sqrt{2\delta})} \end{aligned} \tag{124}$$

$4\sqrt{2\delta}$

where the last step follows from the triangle inequality: $\|g_1 - g_2\|_{L_2(\Pi)} \le \|g_1 - \bar{g}\|_{L_2(\Pi)} + \|g_2 - \bar{g}\|_{L_2(\Pi)}$. Hence, by (124),

$$\sqrt{(D^2)^\dagger(\sigma)} = \sqrt{\sup_{\delta \ge \sigma} \frac{D^2(\delta)}{\delta}} \le 4\sqrt{2}.$$

3. By symmetrization inequality (recall that we use $\epsilon_1, \ldots, \epsilon_n$ to be i.i.d. Rademacher variables independent of the observed data), and letting $\mathcal{F}'(\delta) := \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}(\delta)\}$, and using (123),

$f_1(X_i) - f_2(X_i)$ Theorem 3.1

$$\begin{aligned}
\phi_n(\delta) = \mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}'(\delta)} &\le 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}'(\delta)} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i f(X_i)\Big|\right] \\
&\le 2\,\mathbb{E}\left[\sup_{g_k \in \mathcal{G}: \|g_k - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i\,(\ell \bullet g_1 - \ell \bullet g_2)(X_i)\Big|\right] \\
&\le 4\,\mathbb{E}\left[\sup_{g \in \mathcal{G}: \|g - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i\,(\ell \bullet g - \ell \bullet \bar{g})(X_i)\Big|\right].
\end{aligned}$$

$k = 1, 2.$

Since $\ell(y, \cdot)$ is Lipschitz with constant 2 on the interval $[0,1]$ one can use the *contraction inequality*[87] to get

$L = 2.$

$$\phi_n(\delta) \le 8\,\mathbb{E}\left[\sup_{g \in \mathcal{G}: \|g - \bar{g}\|^2_{L_2(\Pi)} \le 2\delta} \frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i (g - \bar{g})(Z_i)\Big|\right] := \psi_n(\delta).$$

As a result, we get (recall (119)) $\psi^\dagger(\sigma) := \sup_{\delta \ge \sigma} \frac{\psi(\delta)}{\delta}.$

$$\phi_n^\dagger(\sigma) \le \psi_n^\dagger(\sigma).$$

The following result is now a corollary of Theorem 8.16.

Theorem 8.18. *Let $\mathcal{G}$ be a convex class of functions from $\mathcal{Z}$ into $[0,1]$ and let $\hat{g}_n$ denotes the LSE of the regression function, i.e.,*

$$\hat{g}_n := \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \{Y_i - g(X_i)\}^2.$$

Then, there exist constants $K > 0$ such that for all $t > 0$,

$$\mathbb{P}\left\{\|\hat{g}_n - g_*\|^2_{L_2(\Pi)} \ge \inf_{g \in \mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + \left(\psi_n^\sharp\Big(\frac{1}{4q}\Big) + K\frac{t}{n}\right)\right\} \le C_q e^{-t}, \tag{125}$$

[87] Ledoux-Talagrand contraction inequality (Theorem 4.12 of [Ledoux and Talagrand, 1991]): If $\varphi_i : \mathbb{R} \to \mathbb{R}$ satisfies $|\varphi_i(a) - \varphi_i(b)| \le L|a - b|$ for all $a, b \in \mathbb{R}$, then

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \varphi_i(h(x_i))\right] \le L\,\mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(x_i)\right].$$

In the above application we take $\varphi_i(u) = (Y_i - u)^2$ for $u \in [0,1]$.

$P(\mathcal{E}_P(\hat{f}_n) > \sigma_n^t) \le Cq e^{-t}$ (122)

where for any $\psi : (0, \infty) \to \mathbb{R}$, $\psi^\sharp$ *is defined as*[88]

$$\psi^\sharp(\varepsilon) := \inf\left\{\sigma > 0 : \psi^\dagger(\sigma) \le \varepsilon\right\}. \tag{126}$$

Proof. Note that in this case, by (104), $\mathcal{E}_P(\hat{g}_n) = \|\hat{g}_n - g_*\|^2_{L_2(\Pi)} - \inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)}$. To use Theorem 8.16 we need to upper bound the quantity σ_n^t defined in (121). Recall the definition of $V_n^t(\sigma)$ from (120). By the above observations 1-3, we have

$$V_n^t(\sigma) \le 2q\left[\psi_n^\dagger(\sigma) + 4\sqrt{2}\sqrt{\frac{t}{n\sigma}} + \frac{t}{n\sigma}\right] \tag{127}$$

We are only left to show that $\sigma_n^t := \inf\{\sigma : V_n^t(\sigma) \le 1\} \le \psi_n^\sharp(\frac{1}{4q}) + K\frac{t}{n}$, for a sufficiently large K, which will be implied if we can show that $V_n^t\left(\psi_n^\sharp(\frac{1}{2q}) + K\frac{t}{n}\right) \le 1$ (since then $\psi_n^\sharp(\frac{1}{2q}) + K\frac{t}{n} \in \{\sigma : V_n^t(\sigma) \le 1\}$ and the result follows from the minimality of σ_n^t). Note that, by the nonincreasing nature of each of the terms on the right hand side of (127),

$$\begin{aligned} V_n^t\left(\psi_n^\sharp(\frac{1}{4q}) + K\frac{t}{n}\right) &\le 2q\left[\psi_n^\dagger(\psi_n^\sharp(\frac{1}{4q})) + 4\sqrt{2}\sqrt{\frac{t}{n(Kt/n)}} + \frac{t}{n(Kt/n)}\right] \\ &\le 2q\left[\frac{1}{4q} + \frac{4\sqrt{2}}{\sqrt{K}} + \frac{1}{K}\right] < 1, \end{aligned}$$

where $K > 0$ is chosen so that $\frac{4\sqrt{2}}{\sqrt{K}} + \frac{1}{K} < \frac{1}{2}$ (note that $\psi_n^\dagger(\psi_n^\sharp(\frac{1}{4q})) \le \frac{1}{4q}$). $\square$

Example 8.19 (Finite dimensional classes)**.** *Suppose that* $\mathcal{L} \subset L_2(\Pi)$ *is a finite dimensional linear space with* $\dim(\mathcal{L}) = d < \infty$*. and let* $\mathcal{G} \subset \mathcal{L}$ *be a convex class of functions taking values in a bounded interval (for simplicity,* $[0, 1]$*). We would like to show that*

$$\mathbb{P}\left\{\|\hat{g}_n - g_*\|^2_{L_2(\Pi)} \ge \inf_{g\in\mathcal{G}} \|g - g_*\|^2_{L_2(\Pi)} + \left(\frac{d}{n} + K\frac{t}{n}\right)\right\} \le Ce^{-t} \tag{128}$$

with some constant $C, K > 0$.

It can be shown that[89] *that*

$$\psi_n(\delta) \le c\sqrt{\frac{d\delta}{n}}$$

with some constant $c > 0$*. Hence,*

$$\psi_n^\dagger(\sigma) = \sup_{\delta\ge\sigma} \frac{\psi_n(\delta)}{\delta} \le \sup_{\delta\ge\sigma} c\sqrt{\frac{d}{\delta n}} = c\sqrt{\frac{d}{\sigma n}}.$$

[88] Note that $\psi^\sharp$ can be thought of as the *generalized inverse* of $\psi^\dagger$. Thus, under the assumption that $\psi^\dagger$ is right-continuous, $\psi^\dagger(\sigma) \le \varepsilon$ if and only if $\sigma \ge \psi^\sharp(\varepsilon)$ (Exercise (HW3): Show this). Further note that with this notation $\sigma_n^t = V_n^{t,\sharp}(1)$.

[89] Exercise (HW3): Suppose that $\mathcal{L}$ is a finite dimensional subspace of $L_2(P)$ with $\dim(\mathcal{L}) = d$. Then $\mathbb{E}\left[\sup_{f\in\mathcal{L}:\|f\|_{L_2(P)}\le r} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \le r\sqrt{\frac{d}{n}}$.

As, $\psi_n^\dagger(\sigma) \le \varepsilon$ implies $\sigma \ge \psi_n^\sharp(\varepsilon)$, taking $\sigma := \frac{d}{n}$ and $q \ge \max\{1, 1/(4c)\}$, we see that

$$\psi_n^\dagger\left(\frac{d}{n}\right) \le c\sqrt{\frac{d}{\frac{d}{n}n}} \le \frac{1}{4q} \quad \Rightarrow \quad \psi_n^\sharp(\frac{1}{4q}) \le \frac{d}{n},$$

and Theorem 8.18 then implies (128); here $C \equiv C_q$ is taken as in Theorem 8.16 and K as in Theorem 8.18.

Exercise (HW3): Consider the setting of Example 8.19. Instead of using the refined analysis using (105) (and Talagrand's concentration inequality) as illustrated in this section, use the bounded differences inequality to get a crude upper bound on the excess risk of the ERM in this problem. Compare the obtained high probability bound to (128).

Exercise (HW3)[VC-subgraph classes]: Suppose that $\mathcal{G}$ is a convex VC-subgraph class of functions $g : \mathcal{Z} \to [0, 1]$ of VC-dimension V. Then, show that, the function $\psi_n(\delta)$ can be upper bounded by:

$$\psi_n(\delta) \le c\left[\sqrt{\frac{V\delta}{n}\log\frac{1}{\delta}} \vee \frac{V}{n}\log\frac{1}{\delta}\right].$$

Show that $\psi_n^\sharp(\varepsilon) \le \frac{cV}{n\varepsilon^2}\log\frac{n\varepsilon^2}{V}$. Finally, use Theorem 8.18 to obtain a high probability bound analogous to (125).

Exercise (HW3)[Nonparametric classes]: In the case when the metric entropy of the class $\mathcal{G}$ (random, uniform, bracketing, etc.; e.g., if $\log N(\varepsilon, \mathcal{G}, L_2(\mathbb{P}_n)) \le \left(\frac{A}{\varepsilon}\right)^{2\rho})$ is bounded by $O(\varepsilon^{-2\rho})$ for some $\rho \in (0, 1)$ (assuming that the envelope of $\mathcal{G}$ is 1), we typically have $\psi_n^\sharp(\varepsilon) \le O(n^{-1/(1+\rho)})$. Finally, use Theorem 8.18 to obtain a high probability bound analogous to (125).

8.4 Kernel density estimation

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. P on $\mathbb{R}^d$, $d \ge 1$. Suppose P has density p with respect to the Lebesgue measure on $\mathbb{R}^d$, and $\|p\|_\infty < \infty$. Let $K : \mathbb{R}^d \to \mathbb{R}$ be any measurable function that integrates to one, i.e.,

$$\int_{\mathbb{R}^d} K(y)dy = 1$$

and $\|K\|_\infty < \infty$. Then the kernel density estimator (KDE) of p if given by

$$\widehat{p}_{n,h}(y) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{y - X_i}{h}\right) = h^{-d}\mathbb{P}_n\left[K\left(\frac{y-X}{h}\right)\right], \qquad \text{for } y \in \mathbb{R}^d.$$

Here h is called the smoothing bandwidth. Choosing a suitable bandwidth sequence $h_n \to 0$ and assuming that the density p is continuous, one can obtain a strongly consistent estimator $\widehat{p}_{n,h}(y) \equiv \widehat{p}_{n,h_n}(y)$ of $p(y)$, for any $y \in \mathbb{R}^d$.

It is natural to write the difference $\widehat{p}_n(y,h) - p(y)$ as the sum of a random term and a deterministic term:

$$\widehat{p}_{n,h}(y) - p(y) = \widehat{p}_{n,h}(y) - p_h(y) + p_h(y) - p(y)$$

$\hat{p}_{n,h}(y)$ empirical ---

true ---

where

$$p_h(y) := h^{-d} P\Big[K\Big(\frac{y - X}{h}\Big)\Big] = h^{-d} \int_{\mathbb{R}^d} K\Big(\frac{y - x}{h}\Big) p(x)dx = \int_{\mathbb{R}^d} K(u)p(y - hu)du$$

is a smoothed version of p. Convergence to zero of the second term can be argued based only on smoothness assumptions on p: if p is uniformly continuous, then it is easily seen that

$x \to y \Rightarrow f(x) \to f(y)$

$p_n(y) - p(y) \to 0$

$$\sup_{h \le b_n} \sup_{y \in \mathbb{R}^d} |p_h(y) - p(y)| \to 0$$

for any sequence $b_n \to 0$. On the other hand, the first term is just

$$\hat{p}_{n,h}(y) - p_n(y) = h^{-d}(\mathbb{P}_n - P)\Big[K\Big(\frac{y - X}{h}\Big)\Big]. \tag{129}$$

For a fixed $y \in \mathbb{R}^d$, it is easy to study the properties of the above display using the CLT as we are dealing with a sum of independent random variables $h^{-d}K\big(\frac{y-X_i}{h}\big)$, $i = 1, \ldots, n$. However, it is natural to ask whether the KDE $\widehat{p}_{n,h_n}$ converges to p uniformly (a.s.) for a sequence of bandwidths $h_n \to 0$ and, if so, what is the rate of convergence in that case? We investigate this question using tools from empirical processes.

The KDE $\widehat{p}_{n,h}(\cdot)$ is indexed by the bandwidth h, and it is natural to consider $\widehat{p}_{n,h}$ as a process indexed by both $y \in \mathbb{R}^d$ and $h > 0$. This leads to studying the class of functions

$$\mathcal{F} := \left\{ x \mapsto K\left(\frac{y - x}{h}\right) : y \in \mathbb{R}^d, h > 0 \right\}.$$

It is fairly easy to give conditions on the kernel K so that the class $\mathcal{F}$ defined above satisfies

covering number

$$N(\epsilon \|K\|_\infty, \mathcal{F}, L_2(Q)) \le (A/\epsilon)^V \tag{130}$$

for some constants $V \ge 2$ and $A \ge e^2$; see e.g., Lemma 7.22[90]. While it follows immediately from the GC theorem that

$$\sup_{h>0, y \in \mathbb{R}^d} \left| (\mathbb{P}_n - P)\Big[K\Big(\frac{y - X}{h}\Big)\Big] \right| \stackrel{a.s.}{\to} 0,$$

this does not suffice in view of the factor of h^{-d} in (129). In fact, we need a rate of convergence for $\sup_{h>0, y\in\mathbb{R}^d}(\mathbb{P}_n - P)\Big[K\Big(\frac{y-X}{h}\Big)\Big] \stackrel{a.s.}{\to} 0$. The following theorem gives such a result[91].

[90] For instance, it is satisfied for general $d \ge 1$ whenever $K(x) = \phi(q(x))$, with $q(x)$ being a polynomial in d variables and ϕ being a real-valued right continuous function of bounded variation.

[91] To study variable bandwidth kernel estimators [Einmahl and Mason, 2005] derived the following result, which can be proved with some extra effort using ideas from the proof of Theorem 8.21.

Theorem 8.20. *For any $c > 0$, with probability 1,*

$$\limsup_{n\to\infty} \sup_{c \log n/n \le h \le 1} \frac{\sqrt{nh}\|\widehat{p}_{n,h}(y) - p_h(y)\|_\infty}{\sqrt{\log(1/h) \vee \log\log n}} =: K(c) < \infty.$$

Theorem (8.20) implies for any sequences $0 < a_n < b_n \le 1$, satisfying $b_n \to 0$ and $na_n/\log n \to \infty$, with probability 1 ,

$$\sup_{a_n \le h \le b_n} \|\widehat{p}_{n,h} - p_h\|_\infty = O\left(\sqrt{\frac{\log(1/a_n) \vee \log\log n}{na_n}}\right),$$

which in turn implies that $\lim_{n\to\infty} \sup_{a_n \le h \le b_n} \|\widehat{p}_{n,h} - p_h\|_\infty \stackrel{a.s.}{\to} 0.$

Theorem 8.21. *Suppose that $h_n \downarrow 0$, $nh_n^d/|\log h_n| \to \infty$, $\log\log n/|\log h_n| \to \infty$ and $h_n^d \le \check{c} h_{2n}^d$ for some $\check{c} > 0$. Then*

$$\limsup_{n\to\infty} \frac{\sqrt{nh_n^d}\|\widehat{p}_{n,h_n}(\cdot) - p_{h_n}(\cdot)\|_\infty}{\sqrt{\log h_n^{-1}}} = C \quad a.s.$$

where $C < \infty$ is a constant that depends only on the VC characteristics of $\mathcal{F}$.

Proof. We will use the following result:

Lemma 8.22 ([de la Peña and Giné, 1999, Theorem 1.1.5])**.** *If $X_i, i \in \mathbb{N}$, are i.i.d $\mathcal{X}$-valued random variables and $\mathcal{F}$ a class of measurable functions, then*

$$\mathbb{P}\left(\max_{1\le j\le n}\Big\|\sum_{i=1}^{j}(f(X_i) - Pf)\Big\|_{\mathcal{F}} > t\right) \le 9\,\mathbb{P}\Big(\Big\|\sum_{i=1}^{n}(f(X_i) - Pf)\Big\|_{\mathcal{F}} > \frac{t}{30}\Big).$$

For $k \ge 0$, let $n_k := 2^k$. Let $\lambda > 0$; to be chosen later. The monotonicity of $\{h_n\}$ (hence of $h_n \log h_n^{-1}$ once $h_n < e^{-1}$) and Lemma 8.22 imply (for $k \ge 1$)

$$\begin{aligned}
&\mathbb{P}\left(\max_{n_{k-1}<n\le n_k}\sqrt{\frac{nh_n^d}{\log h_n^{-1}}}\|\widehat{p}_{n,h_n}(y) - p_{h_n}(y)\|_\infty > \lambda\right)\\
=\;& \mathbb{P}\left(\max_{n_{k-1}<n\le n_k}\sqrt{\frac{1}{nh_n^d\log h_n^{-1}}}\sup_{y\in\mathbb{R}^d}\left|\sum_{i=1}^{n}\Big[K\Big(\frac{y-X_i}{h_n}\Big) - \mathbb{E}K\Big(\frac{y-X_i}{h_n}\Big)\Big]\right| > \lambda\right)\\
\le\;& \mathbb{P}\left(\frac{1}{\sqrt{n_{k-1}h_{n_k}^d\log h_{n_k}^{-1}}} \times \max_{1\le n\le n_k}\sup_{y\in\mathbb{R}^d, h_{n_k}\le h\le h_{n_{k-1}}}\left|\sum_{i=1}^{n}\Big[K\Big(\frac{y-X_i}{h}\Big) - \mathbb{E}K\Big(\frac{y-X_i}{h}\Big)\Big]\right| > \lambda\right)\\
\le\;& 9\mathbb{P}\left(\frac{1}{\sqrt{n_{k-1}h_{n_k}^d\log h_{n_k}^{-1}}} \times \sup_{y\in\mathbb{R}^d, h_{n_k}\le h\le h_{n_{k-1}}}\left|\sum_{i=1}^{n_k}\Big[K\Big(\frac{y-X_i}{h}\Big) - \mathbb{E}K\Big(\frac{y-X_i}{h}\Big)\Big]\right| > \frac{\lambda}{30}\right). \qquad (131)
\end{aligned}$$

We will study the subclasses

$$\mathcal{F}_k := \left\{K\Big(\frac{y-\cdot}{h}\Big) : h_{n_k} \le h \le h_{n_{k-1}}, y \in \mathbb{R}^d\right\}.$$

As

$$\mathbb{E}\Big[K^2\Big(\frac{y-X}{h}\Big)\Big] = \int_{\mathbb{R}^d} K^2\Big(\frac{y-x}{h}\Big)p(x)dx = h^d\int_{\mathbb{R}^d}K^2(u)p(y-uh)du \le h^d\|p\|_\infty\|K\|_2^2,$$

for the class $\mathcal{F}_k$, we can take

$$U_k := 2\|K\|_\infty, \qquad \text{and} \qquad \sigma_k^2 := h_{n_{k-1}}^d\|p\|_\infty\|K\|_2^2.$$

Since $h_{n_k} \downarrow 0$, and $nh_n^d/\log h_n^{-1} \to \infty$, there exists $k_0 < \infty$ such that for all $k \ge k_0$,

$$\sigma_k < U_k/2 \qquad \text{and} \qquad \sqrt{n_k}\sigma_k \ge \sqrt{V}U_k\sqrt{\log\frac{AU_k}{\sigma_k}}. \qquad \text{(check!)} \qquad (132)$$

Letting $Z_k := \mathbb{E}\Big\| \sum_{i=1}^{n_k}(f(X_i) - Pf)\Big\|_{\mathcal{F}_k}$, we can bound $\mathbb{E}[Z_k]$ by using Theorem 7.13 (see (84)), for $k \geq k_0$, to obtain

$$\mathbb{E}[Z_k] = \mathbb{E}\Big\| \sum_{i=1}^{n_k}(f(X_i) - Pf)\Big\|_{\mathcal{F}_k} \leq L\sigma_k\sqrt{n_k \log(AU_k/\sigma_k)}$$

for a suitable constant $L > 0$. Thus, using (132),

$$\nu_k := n_k\sigma_k^2 + 2U_k\mathbb{E}[Z_k] \leq \tilde{c}n_k\sigma_k^2$$

for a constant $\tilde{c} > 1$ and $k \geq k_0$. Choosing $x = c\log(AU_k/\sigma_k)$ in (99), for some $c > 0$, we see that

$$\begin{aligned}\mathbb{E}[Z_k] + \sqrt{2\nu_k x} + U_k x/3 &\leq \sigma_k\sqrt{n_k \log(AU_k/\sigma_k)}(L + \sqrt{2c\tilde{c}}) + cU_k\log(AU_k/\sigma_k)/3 \\ &\leq C\sigma_k\sqrt{n_k \log(AU_k/\sigma_k)},\end{aligned}$$

for some constant $C > 0$, where we have again used (132). Therefore, by Theorem 8.7,

$$\mathbb{P}\Big(Z_k \geq C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)}\Big) \leq \mathbb{P}(Z_k \geq \mathbb{E}[Z_k] + \sqrt{2\nu_k x} + U_k x/3) \leq e^{-c\log(AU_k/\sigma_k)}.$$

Notice that

$$\frac{30C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)}}{\sqrt{n_{k-1}h_{n_k}^d \log h_{n_k}^{-1}}} > \lambda \qquad \text{(check!)}$$

for some $\lambda > 0$, not depending on k. Therefore, choosing this λ the probability on the right hand-side of (131) can be expressed as

$$\mathbb{P}\left(\frac{Z_k}{\sqrt{n_{k-1}h_{n_k}^d \log h_{n_k}^{-1}}} > \frac{\lambda}{30}\right) \leq \mathbb{P}\Big(Z_k \geq C\sigma_k\sqrt{n_k\log(AU_k/\sigma_k)}\Big) \leq e^{-c\log(AU_k/\sigma_k)}.$$

Since

$$\sum_{k=k_0}^{\infty} e^{-c\log(AU_k/\sigma_k)} = c_1\sum_{k=k_0}^{\infty} h_{n_{k-1}}^{cd/2} \leq \tilde{c}_1\sum_{k=k_0}^{\infty}(\check{c})^{-cd/2} < \infty,$$

for constants $c_1, \tilde{c}_1 > 0$, we get, summarizing,

$$\sum_{k=1}^{\infty}\mathbb{P}\left(\max_{n_{k-1}<n\leq n_k}\sqrt{\frac{nh_n^d}{\log h_n^{-1}}}\|\widehat{p}_{n,h}(y) - p_h(y)\|_\infty > \lambda\right) < \infty.$$

Let $Y_n = \sqrt{\frac{nh_n^d}{\log h_n^{-1}}}\|\widehat{p}_{n,h} - p_h\|_\infty$. Letting $Y := \limsup_{n\to\infty} Y_n$, and using the Borel-Cantelli lemma we can see that $\mathbb{P}(Y > \lambda) = 0$. This yields the desired result using the zero-one law[92]. □

[92]For a fixed $\lambda \geq 0$, define the event $A := \{\limsup_{n\to\infty} Y_n > \lambda\}$. As this is a tail event, by the zero-one law it has probability 0 or 1. We thus have that for each λ, $\mathbb{P}(Y > \lambda) \in \{0, 1\}$. Defining $c := \sup\{\lambda : \mathbb{P}(Y > \lambda) = 1\}$, we get that $Y = c$ a.s. Note that $c < \infty$ as there exists $\lambda > 0$ such that $\mathbb{P}(Y > \lambda) = 0$, by the proof of Theorem 8.21.