8 Talagrand's concentration inequality for the suprema of the empirical process

The main goal of this chapter is to motivate and formally state (without proof) Talagrand's inequality for the suprema of the empirical process. We will also see a few applications of this result. If we have time, towards the end of the course, I will develop the tools necessary and prove the main result. To fully appreciate the strength of the main result, we start with a few important tail bounds for the sum of independent random variables. The following discussion extends and improves Hoeffding's inequality (Lemma 3.9).

In most of results in this chapter we only assume that the $\mathcal{X}$-valued random variables $X_1, \ldots, X_n$ are independent; they need not be identically distributed.

8.1 Preliminaries

Recall Hoeffding's inequality: Let $X_1, \ldots, X_n$ be independent and centered random variables such that $X_i \in [a_i, b_i]$ w.p.1 and let $S_n := \sum_{i=1}^n X_i$. Then, for any $t \geq 0$,

$$\mathbb{P}\left(S_n \geq t\right) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2} \quad \text{and} \quad \mathbb{P}\left(S_n \leq -t\right) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}. \tag{89}$$

A crucial ingredient in the proof of the above result was Lemma 3.8 which stated that for a centered $X \in [a, b]$ w.p.1 we have $\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 (b-a)^2/8}$, for $\lambda \geq 0$.

Note that if $b_i - a_i$ is much larger than the standard deviation σ_i of X_i then, although the tail probabilities prescribed by Hoeffding's inequality for S_n are of the normal type[71], they correspond to normal variables with the 'wrong' variance. The following result incorporates the standard deviation of the random variable and is inspired by the moment generating function of Poisson random variables[72].

Theorem 8.1. *Let X be a centered random variable such that $|X| \leq c$ a.s, for some $c < \infty$, and $\mathbb{E}[X^2] = \tau^2$. Then*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\tau^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right), \qquad \textit{for all } \lambda > 0. \tag{90}$$

As a consequence, if X_i, $1 \leq i \leq n$, are centered, independent and a.s. bounded by $c < \infty$ in absolute value, then setting

$$\sigma^2 := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2], \tag{91}$$

[71] Recall that if the X_i's are i.i.d. and centered with variance σ^2, by the CLT for fixed $t > 0$, $\lim_{n\to\infty} \mathbb{P}\left(S_n \geq t\sqrt{n}\right) = 1 - \Phi\left(\frac{t}{\sigma}\right) \leq \frac{\sigma}{\sqrt{2\pi}t} \exp\left(-\frac{t^2}{2\sigma^2}\right)$, where the last inequality uses a standard bound on the normal CDF.

[72] Recall that if X has Poisson distribution with parameter a (i.e., $\mathbb{E}X = \mathrm{Var}(X) = a$) then $\mathbb{E}[e^{\lambda(X-a)}] = e^{-a(\lambda+1)} \sum_{k=0}^\infty e^{\lambda k} a^k / k! = e^{a(e^\lambda - 1 - \lambda)}$.

and $S_n = \sum_{i=1}^n X_i$, *we have*

$$\mathbb{E}[e^{\lambda S_n}] \le \exp\left(\frac{n\sigma^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right), \qquad \textit{for all } \lambda > 0, \tag{92}$$

and the same inequality holds for $-S_n$.

Proof. Since $\mathbb{E}(X) = 0$, expansion of the exponential gives

$$\mathbb{E}[e^{\lambda X}] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!} \le \exp\left(\sum_{i=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!}\right).$$

Since $|\mathbb{E}X^k| \le c^{k-2}\tau^2$, for all $k \ge 2$, this exponent can be bounded by

$$\left|\sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}X^k}{k!}\right| \le \lambda^2\tau^2 \sum_{k=2}^{\infty} \frac{(\lambda c)^{k-2}}{k!} = \frac{\tau^2}{c^2} \sum_{k=2}^{\infty} \frac{(\lambda c)^k}{k!} = \frac{\tau^2}{c^2}(e^{\lambda c} - 1 - \lambda c).$$

This gives inequality (90). Inequality (92) follows from (90) by using the independence of the X_i's. The above also applies to $Y_i = -X_i$ which yields the result for $-S_n$. □

It is standard to derive tail probability bounds for a random variable based on a bound for its moment generating function. We proceed to implement this idea and obtain four such bounds, three of them giving rise, respectively, to the *Bennett, Prokhorov* and *Bernstein* classical inequalities for sums of independent random variables and one where the bound on the tail probability function is inverted. It is convenient to introduce the following notation:

$$\begin{aligned} \phi(x) &= e^{-x} - 1 + x, \qquad \text{for } x \in \mathbb{R} \\ h_1(t) &= (1+t)\log(1+t) - t, \qquad \text{for } t \ge 0. \end{aligned}$$

Proposition 8.2. *Let* Z *be a random variable whose moment-generating function satisfies the bound*

$$\mathbb{E}(e^{\lambda Z}) \le \exp\big(\nu(e^{\lambda} - 1 - \lambda)\big), \qquad \lambda > 0, \tag{93}$$

for some $\nu > 0$. *Then, for all* $t \ge 0$,

$$\mathbb{P}(Z \ge t) \le e^{-\nu h_1(t/\nu)} \le \exp\left(-\frac{3t}{4}\log\left(1 + \frac{2t}{3\nu}\right)\right) \le \exp\left(-\frac{t^2}{2\nu + 2t/3}\right) \tag{94}$$

and

$$\mathbb{P}\Big(Z \ge \sqrt{2\nu x} + x/3\Big) \le e^{-x}, \qquad x \ge 0. \tag{95}$$

Proof. Observe that by Markov's inequality and the given bound $\mathbb{E}[e^{\lambda Z}]$, we obtain

$$\mathbb{P}(Z \ge t) = \mathbb{P}(e^{\lambda Z} \ge e^{\lambda t}) \le \inf_{\lambda > 0} e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] \le e^{\nu \inf_{\lambda>0}\{\phi(-\lambda) - \lambda t/\nu\}}.$$

It can be checked that for $z > -1$ (think of $z = t/\nu$)

$$\inf_{\lambda \in \mathbb{R}}\{\phi(-\lambda) - \lambda z\} = z - (1+z)\log(1+z) = -h_1(z).$$

This proves the first inequality in (94). We can also show that (by checking the value of the corresponding functions at $t = 0$ and then comparing derivatives)

$$h_1(t) \geq \frac{3t}{4} \log\left(1 + \frac{2t}{3}\right) \geq \frac{t^2}{2 + 2t/3}, \qquad \text{for } t > 0,$$

thus completing the proof of the three inequalities in (94).

To prove (95), we begin by observing that (by Taylor's theorem) $(1-\lambda/3)(e^\lambda - \lambda - 1) \leq \lambda^2/2, \lambda \geq 0$. Thus, if

$$\varphi(\lambda) := \frac{\nu\lambda^2}{2(1-\lambda/3)}, \qquad \lambda \in [0, 3),$$

then inequality (93) yields

$$\mathbb{P}(Z \geq t) \leq \inf_{0 \leq \lambda < 3} e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] \leq \exp\left[\inf_{0 \leq \lambda < 3}(\varphi(\lambda) - \lambda t)\right] = \exp\left[-\sup_{0 \leq \lambda < 3}(\lambda t - \varphi(\lambda))\right] = e^{-\gamma(t)},$$

where we have used the fact that $\nu(e^\lambda - 1 - \lambda) \leq \varphi(\lambda)$ and $\gamma(s) := \sup_{\lambda \in [0,3)}(\lambda s - \varphi(\lambda))$, for $s > 0$. Then it can be shown[73] that $\gamma^{-1}(x) = \sqrt{2\nu x} + x/3$. Therefore, letting $t = \gamma^{-1}(x)$ (i.e., $x = \gamma(t)$) in the above display yields (95). □

Let $X_i, 1 \leq i \leq n$, be independent centered random variables a.s. bounded by $c < \infty$ in absolute value. Let $S_n := \sum_{i=1}^n X_i$ and define $Z := S_n/c$. Then,

$$\mathbb{E}[e^{\lambda Z}] = \prod_{i=1}^n \mathbb{E}[e^{(\lambda/c)X_i}] \leq \prod_{i=1}^n \exp\left(\frac{\mathbb{E}[X_i^2]}{c^2}(e^\lambda - 1 - \lambda)\right) = \exp\left(\frac{n\sigma^2}{c^2}(e^\lambda - 1 - \lambda)\right)$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2]$. Thus, Z satisfies the hypothesis of Proposition 8.2 with $\nu := n\sigma^2/c^2$. Therefore we have the following exponential inequalities, which go by the names of *Bennet's, Prokhorov's and Bernstein's*[74] (in that order).

Theorem 8.4. *Let $X_i, 1 \leq i \leq n$, be independent centered random variables a.s. bounded by $c < \infty$ in absolute value. Set $\sigma^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]/n$ and $S_n := \sum_{i=1}^n X_i$. Then, for all $x \geq 0$,*

$$\mathbb{P}(S_n \geq t) \leq e^{-\left(\frac{n\sigma^2}{c^2}\right)h_1\left(\frac{tc}{n\sigma^2}\right)} \leq \exp\left(-\frac{3t}{4c}\log\left(1 + \frac{2tc}{3n\sigma^2}\right)\right) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + 2ct/3}\right) \quad (96)$$

[73] Exercise (HW3): Complete this.

[74] It is natural to ask whether Theorem 8.4 extends to unbounded random variables. In fact, Bernstein's inequality does hold for random variables X_i with finite exponential moments, i.e., such that $\mathbb{E}[e^{\lambda|X_i|}] < \infty$, for some $\lambda > 0$, as shown below.

Lemma 8.3 (Bernstein's inequality)**.** *Let $X_i, 1 \leq i \leq n$, be centered independent random variables such that, for all $k \geq 2$ and all $1 \leq i \leq n$,*

$$\mathbb{E}|X_i|^k \leq \frac{k!}{2}\sigma_i^2 c^{k-2},$$

and set $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$, $S_n := \sum_{i=1}^n X_i$. Then,

$$\mathbb{P}(S_n \geq t) \leq \exp\left(-\frac{t^2}{2n\sigma^2 + 2ct}\right), \qquad \text{for } t \geq 0.$$

and

$$\mathbb{P}\Big(S_n \geq \sqrt{2n\sigma^2 x} + cx/3\Big) \leq e^{-x}, \qquad x \geq 0.$$

Bennett's inequality is the sharpest, but Prokhorov's and Bernstein's inequalities are easier to interpret. Prokhorov's inequality exhibits two regimes for the tail probabilities of S_n: if $tc/(n\sigma^2)$ is small, then the logarithm is approximately $2tc/(3n\sigma^2)$, and the tail probability is only slightly larger than $e^{-t^2/(2n\sigma^2)}$ (which is Gaussian-like), whereas, if $tc/(n\sigma^2)$ is not small or moderate, then the exponent for the tail probability is of the order of $-[3t/(4c)]\log[2tc/(3n\sigma^2)]$ (which is 'Poisson'-like[75]). Bernstein's inequality keeps the Gaussian-like regime for small values of $tc/(n\sigma^2)$ but replaces the Poisson regime by the larger, hence less precise, exponential regime.

Example 8.5 (Deviation bound with fixed probability)**.** *Let us try to shed some light on the differences between Bernstein's inequality (i.e., the rightmost side of (96)) and Hoeffding's inequality (see (89)). We can first attempt to find the value of t which makes the bound on the rightmost side of (96) exactly equal to α, i.e., we want to solve the equation*

$$\exp\left(-\frac{t^2}{2(n\sigma^2 + ct/3)}\right) = \alpha.$$

This leads to the quadratic equation

$$t^2 - \frac{2tc}{3}\log\frac{1}{\alpha} - 2n\sigma^2\log\frac{1}{\alpha} = 0,$$

whose nonnegative solution is given by

$$t = \frac{c}{3}\log\frac{1}{\alpha} + \sqrt{\frac{c^2}{9}\left(\log\frac{1}{\alpha}\right)^2 + 2n\sigma^2\log\frac{1}{\alpha}} \;\leq\; \sigma\sqrt{2n\log\frac{1}{\alpha}} + \frac{2c}{3}\log\frac{1}{\alpha}.$$

where in the last inequality we used the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. Thus, Bernstein's inequality implies that $S_n \leq \sigma\sqrt{2n\log\frac{1}{\alpha}} + \frac{2c}{3}\log\frac{1}{\alpha}$ with probability at least $1-\alpha$. Now if $X_1, \ldots, X_n$ are i.i.d. with mean zero, variance σ^2 and bounded in absolute value by c, then this yields

$$\bar{X}_n \leq \frac{\sigma}{\sqrt{n}}\sqrt{2\log\frac{1}{\alpha}} + \frac{2c}{3n}\log\frac{1}{\alpha} \tag{97}$$

with probability (w.p.) at least $1-\alpha$; compare this the Hoeffding's bound which yields $\bar{X}_n \leq c\sqrt{\frac{2}{n}\log\frac{1}{\alpha}}$ w.p. at least $1-\alpha$; see (11). Note that if $\bar{X}_n$ is normal, then $\bar{X}_n$ will be bounded by the first term in the right hand side of (97) w.p. at least $1-\alpha$. Therefore the above deviation bound agrees with the normal approximation bound except for the smaller order term (which if of order $1/n$; the leading term being of order $1/\sqrt{n}$).

[75] Note that if X has Poisson distribution with parameter a (i.e., $\mathbb{E}X = \mathrm{Var}(X) = a$) then

$$\mathbb{P}(X - a \geq t) \leq \exp\left[-\frac{3t}{4}\log\left(1 + \frac{2t}{3a}\right)\right], \qquad t \geq 0.$$

Example 8.6 (When X_i's are i.i.d. Bernoulli)**.** *Suppose that X_i's are i.i.d. Bernoulli with probability of success $p \in (0,1)$. Then, using (97), we see that using the Bernstein's inequality yields that $\bar{X}_n \le \sqrt{\frac{p(1-p)}{n}}\sqrt{2\log\frac{1}{\alpha}} + \frac{2}{3n}\log\frac{1}{\alpha}$ holds w.p. at least $1-\alpha$; compare this with Hoeffding's inequality which yields $\bar{X}_n \le \sqrt{\frac{2}{n}\log\frac{1}{\alpha}}$ w.p. at least $1-\alpha$. Note that Bernstein's inequality is superior here if $p(1-p)$ is a fairly small. In particular, if $\mathrm{Var}(X_1) = \frac{1}{n}$ (i.e., $p \approx \frac{1}{n}$), then the two upper bounds reduce to $\frac{1}{n}\sqrt{2\log\frac{1}{\alpha}} + \frac{2}{3n}\log\frac{1}{\alpha}$ and $\sqrt{\frac{2}{n}\log\frac{1}{\alpha}}$ respectively, showing that Bernstein's inequality is so much better in this case.*

8.2 Talagrand's concentration inequality

Talagrand's concentration inequality for the supremum of the empirical process [Talagrand, 1996a] is one of the most useful results in modern empirical process theory, and also one of the deepest results in the theory. This inequality may be thought of as a Bennett, Prokhorov or Bernstein inequality uniform over an infinite collection of sums of independent random variables, i.e., for the supremum of the empirical process. As such, it constitutes an exponential inequality of the best possible kind. Below we state Bousquet's version of the upper half of Talagrand's inequality.

Theorem 8.7 (Talagrand's inequality, [Talagrand, 1996a, Bousquet, 2003])**.** *Let $X_i, i = 1,\dots,n$, be independent $\mathcal{X}$-valued random variables. Let $\mathcal{F}$ be a countable family of measurable real-valued functions on $\mathcal{X}$ such that $\|f\|_\infty \le U < \infty$ and $\mathbb{E}[f(X_1)] = \ldots = \mathbb{E}[f(X_n)] = 0$, for all $f \in \mathcal{F}$. Let*

$$Z := \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \qquad or \qquad Z = \sup_{f\in\mathcal{F}} \Big|\sum_{i=1}^n f(X_i)\Big|$$

and let the parameters σ^2 and ν_n be defined as

$$U^2 \ge \sigma^2 \ge \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_i)] \qquad and \qquad \nu_n := 2U\mathbb{E}[Z] + n\sigma^2.$$

Then[76]*, for all $t \ge 0$,*

$$\mathbb{P}(Z \ge \mathbb{E}Z + t) \le e^{-\left(\frac{\nu_n}{U^2}\right)h_1\left(\frac{tU}{\nu_n}\right)} \le \exp\left(-\frac{3t}{4U}\log\left(1 + \frac{2tU}{3\nu_n}\right)\right) \le \exp\left(\frac{-t^2}{2\nu_n + 2tU/3}\right) \quad (98)$$

and

$$\mathbb{P}\Big(Z \ge \mathbb{E}Z + \sqrt{2\nu_n x} + Ux/3\Big) \le e^{-x}, \qquad x \ge 0. \quad (99)$$

[76] This is a consequence of the following: consider the class of functions $\tilde{\mathcal{F}} = \{f/U : f \in \mathcal{F}\}$ (thus any $\tilde{f} \in \tilde{\mathcal{F}}$ satisfies $\|\tilde{f}\|_\infty \le 1$). Let $\tilde{Z} := Z/U$, $\tilde{\sigma}^2 := \sigma^2/U^2$, and $\tilde{\nu}_n := \nu_n/U^2$. Then,

$$\log \mathbb{E}[e^{\lambda(\tilde{Z} - \mathbb{E}\tilde{Z})}] \le \tilde{\nu}_n(e^\lambda - 1 - \lambda), \qquad \lambda \ge 0.$$

Notice the similarity between (98) and the Bennet, Prokhorov and Bernstein inequalities in (96) in Theorem 8.4: in the case when $\mathcal{F} = \{f\}$, with $\|f\|_\infty \le c$, and $\mathbb{E}[f(X_i)] = 0$, U becomes c, and ν_n becomes $n\sigma^2$, and the right-hand side of Talagrand's inequality becomes exactly the Bennet, Prokhorov and Bernstein inequalities. Clearly, Talagrand's inequality is essentially the best possible exponential bound for the empirical process.

Whereas the Bousquet-Talagrand upper bound for the moment generating function of the supremum Z of an empirical process for $\lambda \ge 0$ is best possible, there exist quite good results for $\lambda < 0$, but these do not exactly reproduce the classical exponential bounds for sums of independent random variables when specified to a single function. Here is the strongest result available in this direction.

Theorem 8.8 ([Klein and Rio, 2005])**.** *Under the same hypothesis and notation as in Theorem 8.7, we have*

$$\log \mathbb{E}[e^{-\lambda(\tilde{Z}-\mathbb{E}\tilde{Z})}] \le \frac{\tilde{V}_n}{9}(e^{3\lambda} - 1 - 3\lambda), \qquad 0 \le \lambda < 1,$$

where $\tilde{V}_n = V_n/U^2$ *and*

$$V_n := 2U\mathbb{E}[Z] + \sup_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E}[f^2(X_i)].$$

Then, for all $t \ge 0$,

$$\mathbb{P}(Z \le \mathbb{E}Z - t) \le e^{-\left(\frac{V_n}{9U^2}\right) h_1\left(\frac{3tU}{V_n}\right)} \le \exp\left(-\frac{t}{4U}\log\left(1 + \frac{2tU}{V_n}\right)\right) \le \exp\left(\frac{-t^2}{2V_n + 2tU}\right) \quad (100)$$

and

$$\mathbb{P}\left(Z \le \mathbb{E}Z - \sqrt{2V_n x} - Ux\right) \le e^{-x}, \qquad x \ge 0. \quad (101)$$

Remark 8.1. *In order to get concrete exponential inequalities from Theorems 8.7 and 8.8, we need to have good estimates of* $\mathbb{E}Z$ *and* $\sup_{f\in\mathcal{F}} \mathbb{E}[f^2(X_i)]$*. We have already seen many techniques to control* $\mathbb{E}Z$*. In particular, (85) gives such a bound.*

Example 8.9 (Dvoretzky-Kiefer-Wolfowitz)**.** *A first question we may ask is whether Talagrand's inequality recovers, up to constants, the DKW inequality. Let* F *be a distribution function in* $\mathbb{R}^d$ *and let* $\mathbb{F}_n$ *be the distribution function corresponding to* n *i.i.d. variables with distribution* F*. Let* $Z := n\|\mathcal{F}_n - F\|_\infty$ *We can take the envelope of the class* $\mathcal{F} := \{\mathbf{1}_{(-\infty,x]}(\cdot) : x \in \mathbb{R}^d\}$ *to be 1 (i.e.,* $U = 1$*), and* $\sigma^2 = 1/4$*.* $\mathcal{F}$ *is VC (with* $V(\mathcal{F}) = d$*) and inequality (85) gives*

$$\mathbb{E}[Z] = n\mathbb{E}\|\mathbb{F}_n - F\|_\infty \le c_1\sqrt{n},$$

where c_1 *depends only on* d*. Here,* $\nu_n \le 2c_1\sqrt{n} + n/4$*. We have to upper-bound the probability*

$$\mathbb{P}(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \ge x) = \mathbb{P}(Z \ge \sqrt{n}x) = \mathbb{P}(Z - \mathbb{E}Z \ge \sqrt{n}x - \mathbb{E}Z).$$

Note that for $x > 2\sqrt{n}$, this probability is zero (as $Z \le 2n$). For $x > 2c_1$, $t := \sqrt{n}x - \mathbb{E}Z > \sqrt{n}(x - c_1) > 0$, and thus we can apply the last inequality in (98). Hence, for $2\sqrt{n} \ge x > 2c_1$,

$$\begin{aligned}\mathbb{P}(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \ge x) &\le \exp\left(-\frac{(\sqrt{n}x - \mathbb{E}Z)^2}{2(2c_1\sqrt{n} + n/4) + 2(\sqrt{n}x - \mathbb{E}Z)/3}\right) \\ &\le \exp\left(-\frac{n(x - c_1)^2}{c_3 n}\right) \le \exp\left(-\frac{x^2}{4c_3}\right),\end{aligned}$$

where we have used (i) for $2\sqrt{n} \ge x$ the denominator in the exponential term is upper bounded by $2(2c_1\sqrt{n} + n/4) + 4n/3$ which is in turn upper bounded by $c_3 n$ (for some $c_3 > 0$); (ii) for $x > 2c_1$, $(x - c_1)^2 > x^2/4$ (as $x - c_1 \ge x - x/2 = x/2$). Thus, for some constants $c_2, c_3 > 0$ that depend only on d, we can show that for all $x > 0$,

$$\mathbb{P}(\sqrt{n}\|\mathbb{F}_n - F\|_\infty \ge x) \le c_2 e^{-x^2/(4c_3)}.$$

Example 8.10 (Data-driven inequalities)**.** *In many statistical applications, it is of importance to have data-dependent "confidence sets" for the random quantity $\|\mathbb{P}_n - P\|_{\mathcal{F}}$. This quantity is a natural measure of the accuracy of the approximation of an unknown distribution by the empirical distribution $\mathbb{P}_n$. However, $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ itself depends on the unknown distribution P and is not directly available.*

To obtain such data dependent bounds on $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ we have to replace the unknown quantities $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$, σ^2 and U by suitable estimates or bounds. Suppose for the sake of simplicity, σ^2 and U are known, and the only problem is to estimate or bound the expectation $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$. We have discussed so far how to bound the expectation $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$. However, such bounds typically depend on other unknown constants and may not be sharp. Talagrand's inequalities (99) and (101), and symmetrization allow us to replace $\mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$ by a completely data-based surrogate. In the following we give such a (finite-sample) high-probability upper bound on $\|\mathbb{P}_n - P\|_{\mathcal{F}}$; see [Giné and Nickl, 2016, Section 3.4.2] for more on this topic.

Theorem 8.11. *Let $\mathcal{F}$ be a countable collection of real-valued measurable functions on $\mathcal{X}$ with absolute values bounded by 1/2. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{X}$ with a common probability law P. Let $\varepsilon_1, \ldots, e_n$ be i.i.d. Rademacher random variables independent from the sequence $\{X_i\}$ and let $\sigma^2 \ge \sup_{f\in\mathcal{F}} Pf^2$. Then, for all n and $x \ge 0$,*

$$\mathbb{P}\left(\|\mathbb{P}_n - P\|_{\mathcal{F}} \ge 3\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\|_{\mathcal{F}} + 4\sqrt{\frac{2\sigma^2 x}{n}} + \frac{70}{3n}x\right) \le 2e^{-x}.$$

Proof. Set $Z := \|\sum_{i=1}^n (f(X_i) - Pf)\|_{\mathcal{F}}$ and set $\tilde{Z} := \|\sum_{i=1}^n \epsilon_i f(X_i)\|_{\mathcal{F}}$. Note that $\tilde{Z}$ is also the supremum of an empirical process: the variables are $\tilde{X}_i = (\varepsilon_i, X_i)$, defined on $\{-1, 1\} \times \mathcal{X}$, and the functions are $\tilde{f}(\epsilon, x) := \epsilon f(x)$, for $f \in \mathcal{F}$. Thus, Talagrand's inequalities apply to both Z and $\tilde{Z}$. Then, using the fact

$$\sqrt{2x(n\sigma^2 + 2\mathbb{E}\tilde{Z})} \le \sqrt{2xn\sigma^2} + 2\sqrt{x\mathbb{E}\tilde{Z}} \le \sqrt{2xn\sigma^2} + \frac{1}{\delta}x + \delta\mathbb{E}\tilde{Z},$$

Theorem 3.1] For any class of measurable function $\mathcal{F}$

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \le 2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}$$

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

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

Similarly, using (99),

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

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

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

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:

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

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):

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

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.