Case study: stochastic simulation via Rademacher bootstrap

Maxim Raginsky

December 4, 2013

In this lecture, we will look at an application of statistical learning theory to the problem of efficient stochastic simulation, which arises frequently in engineering design. The basic question is as follows. Suppose we have a system with input space Z. The system has a tunable parameter θ that lies in some set Θ. We have a *performance index* $\ell : \mathsf{Z} \times \Theta \to [0,1]$, where we assume that the lower the value of ℓ, the better the performance. Thus, if we use the parameter setting $\theta \in \Theta$ and apply input $z \in \mathsf{Z}$, the performance of the corresponding system is given by the scalar $\ell(z,\theta) \in [0,1]$. Now let's suppose that the input to the system is actually a *random variable* $Z \in \mathsf{Z}$ with some distribution $P \in \mathscr{P}(\mathsf{Z})$. Then we can define the *operating characteristic*

$$L(\theta) \triangleq \mathbb{E}_P[\ell(Z,\theta)] \equiv \int_{\mathsf{Z}} \ell(z,\theta) P_Z(dz), \qquad \theta \in \Theta. \tag{1}$$

The goal is to find an *optimal operating point* $\theta^* \in \Theta$ that achieves (or comes arbitrarily close to) $\inf_{\theta\in\Theta} L(\theta)$.

In practice, the problem of minimizing $L(\theta)$ is quite difficult for large-scale systems. First of all, computing the integral in (1) may be a challenge. Secondly, we may not even know the distribution P_Z. Thirdly, there may be more than one distribution of the input, each corresponding to different operating regimes and/or environments. For this reason, engineers often resort to Monte Carlo simulation techniques: Assuming we can efficiently sample from P_Z, we draw a large number of independent samples $Z_1, Z_2, \ldots, Z_n$ and compute

$$\widehat{\theta_n} = \operatorname*{arg\,min}_{\theta\in\Theta} L_n(\theta) \equiv \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i,\theta),$$

where $L_n(\cdot)$ denotes the empirical version of the operating characteristic (1). Given an accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta \in (0,1)$, we simply need to draw enough samples, so that

$$L(\widehat{\theta}_n) \le \inf_{\theta\in\Theta} L(\theta) + \varepsilon$$

with probability at least $1-\delta$, regardless of what the true distribution P_Z happens to be.

This is, of course, just another instance of the ERM algorithm we have been studying extensively. However, there are two issues. One is how many samples we need to guarantee that

the empirically optimal operating point will be good. The other is the complexity of actually computing an empirical minimizer.

The first issue has already come up in the course under the name of *sample complexity* of learning. The second issue is often handled by relaxing the problem a bit: We choose a probability distribution Q over Θ (assuming it can be equipped with an appropriate σ-algebra) and, instead of minimizing $L(\theta)$ over $\theta \in \Theta$, set some *level parameter* $\alpha \in (0,1)$, and seek any $\widehat{\theta} \in \Theta$, for which there exists some *exceptional set* $\Lambda \subset \Theta$ with $Q(\Lambda) \leq \alpha$, such that

$$\inf_{\theta} L(\theta) - \varepsilon \leq L(\widehat{\theta}) \leq \inf_{\theta \in \Theta \setminus \Lambda} L(\theta) + \varepsilon \tag{2}$$

with probability at least $1-\delta$. Unless the actual optimal operating point θ^* happens to lie in the exceptional set Λ, we will come to within ε of the optimum with confidence at least $1-\delta$. Then we just need to draw a large enough number n of samples $Z_1, \ldots, Z_n$ from P_Z and a large enough number m of samples $\theta_1, \ldots, \theta_m$ from Q, and then compute

$$\widehat{\theta} = \operatorname*{arg\,min}_{\theta \in \{\theta_1, \ldots, \theta_m\}} L_n(\theta).$$

In the next several lectures, we will see how statistical learning theory can be used to develop such simulation procedures. Moreover, we will learn how to use Rademacher averages[1] to determine how many samples we need *in the process of learning*. The use of statistical learning theory for simulation has been pioneered in the context of control by M. Vidyasagar [Vid98, Vid01]; the refinement of his techniques using Rademacher averages is due to Koltchinskii et al. [KAA+00a, KAA+00b]. We will essentially follow their presentation, but with slightly better constants.

We will follow the following plan. First, we will revisit the abstract ERM problem and its sample complexity. Then we will introduce a couple of refined tools pertaining to Rademacher averages. Next, we will look at *sequential* algorithms for empirical approximation, in which the sample complexity is not set *a priori*, but is rather determined by a data-driven *stopping rule*. And, finally, we will see how these sequential algorithms can be used to develop robust and efficient stochastic simulation strategies.

1 Empirical Risk Minimization: a quick review

Recall the *abstract Empirical Risk Minimization problem*: We have a space Z, a class $\mathscr{P}$ of probability distributions over Z, and a class $\mathscr{F}$ of measurable functions $f : \mathsf{Z} \to [0,1]$. Given an i.i.d. sample Z^n drawn according to some unknown $P \in \mathscr{P}$, we compute

$$\widehat{f}_n \triangleq \operatorname*{arg\,min}_{f \in \mathscr{F}} P_n(f) \equiv \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i).$$

[1] More precisely, their stochastic counterpart, in which we do not take the expectation over the Rademacher sequence, but rather use it as a *resource* to aid the simulation.

We would like for $P(\widehat{f}_n)$ to be close to $\inf_{f\in\mathscr{F}} P(f)$ with high probability. To that end, we have derived the bound

$$P(\widehat{f}_n) - \inf_{f\in\mathscr{F}} P(f) \le 2\|P_n - P\|_{\mathscr{F}},$$

where, as before, we have defined the *uniform deviation*

$$\|P_n - P\|_{\mathscr{F}} \triangleq \sup_{f\in\mathscr{F}} \left|P_n(f) - P(f)\right| = \sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}_P f(Z)\right|.$$

Hence, if n is sufficiently large so that, for every $P \in \mathscr{P}$, $\|P_n - P\|_{\mathscr{F}} \le \varepsilon/2$ with P-probability at least $1-\delta$, then $P(\widehat{f}_n)$ will be ε-close to $\inf_{f\in\mathscr{F}} P(f)$ with probability at least $1-\delta$. This motivates the following definition:

Definition 1. *Given the pair* $(\mathscr{F},\mathscr{P})$*, an accuracy parameter* $\varepsilon > 0$*, and a confidence parameter* $\delta \in (0,1)$*, the* sample complexity of empirical approximation *is*

$$N(\varepsilon;\delta) \triangleq \min\left\{n \in \mathbb{N} : \sup_{P\in\mathscr{P}} \mathbb{P}\left\{\|P_n - P\|_{\mathscr{F}} \ge \varepsilon\right\} \le \delta\right\}. \tag{3}$$

In other words, for any $\varepsilon > 0$ and any $\delta \in (0,1)$, $N(\varepsilon/2;\delta)$ is an *upper bound* on the number of samples needed to guarantee that $P(\widehat{f}_n) \le \inf_{f\in\mathscr{F}} P(f) + \varepsilon$ with probability (confidence) at least $1-\delta$.

2 Empirical Rademacher averages

As before, let Z^n be an i.i.d. sample of length n from some $P \in \mathscr{P}(\mathsf{Z})$. On multiple occasions we have seen that the performance of the ERM algorithm is controlled by the *Rademacher average*

$$R_n(\mathscr{F}(Z^n)) \triangleq \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right], \tag{4}$$

where $\sigma^n = (\sigma_1,\ldots,\sigma_n)$ is an n-tuple of i.i.d. Rademacher random variables independent of Z^n. More precisely, we have stablished the fundamental *symmetrization inequality*

$$\mathbb{E}\|P_n - P\|_{\mathscr{F}} \le 2\mathbb{E}R_n(\mathscr{F}(Z^n)), \tag{5}$$

as well as the concentration bounds

$$\mathbb{P}\left\{\|P_n - P\|_{\mathscr{F}} \ge \mathbb{E}\|P_n - P\|_{\mathscr{F}} + \varepsilon\right\} \le e^{-2n\varepsilon^2} \tag{6}$$

$$\mathbb{P}\left\{\|P_n - P\|_{\mathscr{F}} \le \mathbb{E}\|P_n - P\|_{\mathscr{F}} - \varepsilon\right\} \le e^{-2n\varepsilon^2} \tag{7}$$

These results show two things:

1. The uniform deviation $\|P_n - P\|_{\mathscr{F}}$ tightly concentrates around its expected value.

2. The expected value $\mathbb{E}\|P_n - P\|_{\mathscr{F}}$ is bounded from above by $\mathbb{E}R_n(\mathscr{F}(Z^n))$.

It turns out that the expected Rademacher average $\mathbb{E}R_n(\mathscr{F}(Z^n))$ also furnishes a *lower bound* on $\mathbb{E}\|P_n - P\|_{\mathscr{F}}$:

Lemma 1 (Desymmetrization inequality)**.** *For any class $\mathscr{F}$ of measurable functions $f : \mathsf{Z} \to [0,1]$, we have*

$$\frac{1}{2}\mathbb{E}R_n(\mathscr{F}(Z^n)) - \frac{1}{2\sqrt{n}} \le \frac{1}{2n}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[f(Z_i) - P(f)]\right|\right] \le \mathbb{E}\|P_n - P\|_{\mathscr{F}}. \tag{8}$$

Proof. We will first prove the second inequality in (8). To that end, for each $1 \le i \le n$ and each $f \in \mathscr{F}$, let us define $U_i(f) \triangleq f(Z_i) - P(f)$. Then $\mathbb{E}U_i(f) = 0$. Let $\overline{Z}_1, \ldots, \overline{Z}_n$ be an independent copy of $Z_1, \ldots, Z_n$. Then we can define $\overline{U}_i(f), 1 \le i \le n$, similarly. Moreover, since $\mathbb{E}U_i(f) = 0$, we can write

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[f(Z_i) - P(f)]\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i U_i(f)\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i\left[U_i(f) - \mathbb{E}\overline{U}_i(f)\right]\right|\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[U_i(f) - \overline{U}_i(f)]\right|\right].
\end{aligned}$$

Since, for each i, $U_i(f)$ and $\overline{U}_i(f)$ are i.i.d., the difference $U_i(f) - \overline{U}_i(f)$ is a symmetric random variable. Therefore,

$$\left\{\sigma_i[U_i(f) - \overline{U}_i(f)] : 1 \le i \le n\right\} \overset{(\mathrm{d})}{=} \left\{U_i(f) - \overline{U}_i(f) : 1 \le i \le n\right\}.$$

Using this fact and the triangle inequality, we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[U_i(f) - \overline{U}_i(f)]\right|\right] &= \mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n [U_i(f) - \overline{U}_i(f)]\right|\right] \\
&\le 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n U_i(f)\right|\right] \\
&= 2\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n |f(Z_i) - P(f)\right|\right] \\
&= 2n\cdot\mathbb{E}\|P_n - P\|_{\mathscr{F}}.
\end{aligned}$$

To prove the first inequality in (8), we write

$$\begin{aligned}
\mathbb{E}R_n(\mathscr{F}(Z^n)) &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i\left[f(Z_i)-P(f)+P(f)\right]\right|\right] \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[f(Z_i)-P(f)]\right|\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathscr{F}} P(f)\cdot\left|\sum_{i=1}^n \sigma_i\right|\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[f(Z_i)-P(f)]\right|\right] + \frac{1}{n}\mathbb{E}\left|\sum_{i=1}^n \sigma_i\right| \\
&\le \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i[f(Z_i)-P(f)]\right|\right] + \frac{1}{\sqrt{n}}.
\end{aligned}$$

Rearranging, we get the desired inequality. □

In this section, we will see that we can get a lot of mileage out of the *stochastic version* of the Rademacher average. To that end, let us define

$$r_n(\mathscr{F}(Z^n)) \triangleq \frac{1}{n}\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(Z_i)\right|. \tag{9}$$

The key difference between (4) and (9) is that, in the latter, we do *not* take the expectation over the Rademacher sequence σ^n. In other words, both $R_n(\mathscr{F}(Z^n))$ and $r_n(\mathscr{F}(Z^n))$ are random variables, but the former depends only on the training data Z^n, while the latter also depends on the n Rademacher random variables $\sigma_1,\ldots,\sigma_n$. We see immediately that $R_n(\mathscr{F}(Z^n)) = \mathbb{E}[r_n(\mathscr{F}(Z^n))|Z^n]$ and $\mathbb{E}R_n(\mathscr{F}(Z^n)) = \mathbb{E}r_n(\mathscr{F}(Z^n))$, where the expectation on the right-hand side is over both Z^n and σ^n. The following result will be useful:

Lemma 2 (Concentration inequalities for Rademacher averages)**.** *For any $\varepsilon > 0$,*

$$\mathbb{P}\left\{r_n(\mathscr{F}(Z^n)) \ge \mathbb{E}R_n(\mathscr{F}(Z^n)) + \varepsilon\right\} \le e^{-n\varepsilon^2/2} \tag{10}$$

and

$$\mathbb{P}\left\{r_n(\mathscr{F}(Z^n)) \le \mathbb{E}R_n(\mathscr{F}(Z^n)) - \varepsilon\right\} \le e^{-n\varepsilon^2/2}. \tag{11}$$

Proof. For each $1 \le i \le n$, let $U_i \triangleq (Z_i, \sigma_i)$. Then $r_n(\mathscr{F}(Z^n))$ can be represented as a real-valued function $g(U^n)$. Moreover, it is easy to see that this function has bounded differences with $c_1 = \ldots = c_n = 2/n$. Hence, McDiarmid's inequality tells us that for any $\varepsilon > 0$

$$\mathbb{P}\left\{g(U^n) \ge \mathbb{E}g(U^n) + \varepsilon\right\} \le e^{-n\varepsilon^2/2},$$

and the same holds for the probability that $g(U^n) \le \mathbb{E}g(U^n) - \varepsilon$. This completes the proof. □

3 Sequential learning algorithms

In a sequential learning algorithm, the sample complexity is a *random variable*. It is not known in advance, but rather is computed from data in the process of learning. In other words, instead of using a training sequence of fixed length, we keep drawing independent samples until we decide that we have acquired enough of them, and then compute an empirical risk minimizer.

To formalize this idea, we need the notion of a *stopping time*. Let $\{U_n\}_{n=1}^{\infty}$ be a random process. A random variable τ taking values in $\mathbb{N}$ is called a *stopping time* if and only if, for each $n \geq 1$, the occurrence of the event $\{\tau = n\}$ is determined by $U^n = (U_1, \ldots, U_n)$. More precisely:

Definition 2. *For each n, let Σ_n denote the σ-algebra generated by U^n (in other words, Σ_n consists of all events that occur by time n). Then a random variable τ taking values in $\mathbb{N}$ is a stopping time if and only if, for each $n \geq 1$, the event $\{\tau = n\} \in \Sigma_n$.*

In other words, denoting by U^{∞} the entire sample path $(U_1, U_2, \ldots)$ of our random process, we can view τ as a function that maps U^{∞} into $\mathbb{N}$. For each n, the indicator function of the event $\{\tau = n\}$ is a function of U^{∞}:

$$\mathbf{1}_{\{\tau=n\}} \equiv \mathbf{1}_{\{\tau(U^{\infty})=n\}}.$$

Then τ is a stopping time if and only if, for each n and for all U^{∞}, V^{∞} with $U^n = V^n$ we have

$$\mathbf{1}_{\{\tau(U^{\infty})=n\}} = \mathbf{1}_{\{\tau(V^{\infty})=n\}}.$$

Our sequential learning algorithms will work as follows. Given a desired accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta > 0$, let $\overline{n}(\varepsilon, \delta)$ be the initial sample size; we will assume that $\overline{n}(\varepsilon, \delta)$ is a nonincreasing function of both ε and δ. Let $\mathscr{T}(\varepsilon, \delta)$ denote the set of all stopping times τ such that

$$\sup_{P \in \mathscr{P}} \mathbb{P}\{\|P_\tau - P\|_{\mathscr{F}} \leq \varepsilon\} \geq \delta.$$

Now if $\tau \in \mathscr{T}(\varepsilon, \delta)$ and we let

$$\widehat{f}_\tau \triangleq \operatorname*{arg\,min}_{f \in \mathscr{F}} P_\tau(f) \equiv \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{\tau} \sum_{i=1}^{\tau} f(Z_i),$$

then we immediately see that

$$\sup_{P \in \mathscr{P}} \left\{ P(\widehat{f}_\tau) \geq \inf_{f \in \mathscr{F}} P(f) + 2\varepsilon \right\} \leq \delta.$$

Of course, the whole question is how to construct an appropriate stopping time *without knowing P*.

Definition 3. *A parametric family of stopping times $\{\nu(\varepsilon, \delta) : \varepsilon > 0, \delta \in (0,1)\}$ is called* strongly efficient (SE) *(w.r.t. $\mathscr{F}$ and $\mathscr{P}$) if there exist constants $K_1, K_2, K_3 \geq 1$, such that for all $\varepsilon > 0, \delta \in (0,1)$*

$$\nu(\varepsilon, \delta) \in \mathscr{T}(K_1\varepsilon, \delta) \tag{12}$$

and for all $\tau \in \mathscr{T}(\varepsilon, \delta)$

$$\sup_{P \in \mathscr{P}} \mathbb{P}\{\nu(K_2\varepsilon, \delta) > \tau\} \leq K_3\delta. \tag{13}$$

In other words, Eq. (12) says that any SE stopping time $\{\nu(\varepsilon,\delta)\}$ guarantees that we can approximate statistical expectations by empirical expectations with accuracy $K_1\varepsilon$ and confidence $1-\delta$; similarly, Eq. (13) says that, with probability at least $1-K_3\delta$, we will require at most as many samples as would be needed by *any* sequential algorithm for empirical approximation with accuracy ε/K_2 and confidence $1-\delta$.

Definition 4. *A family of stopping times* $\{\nu(\varepsilon,\delta) : \varepsilon > 0, \delta \in (0,1)\}$ *is* weakly efficient (WE) *for* $(\mathscr{F},\mathscr{P})$ *if there exist constants* $K_1, K_2, K_3 \geq 1$*, such that for all* $\varepsilon > 0, \delta \in (0,1)$

$$\nu(\varepsilon,\delta) \in \mathscr{T}(K_1\varepsilon,\delta) \tag{14}$$

and

$$\sup_{P\in\mathscr{P}} \mathbb{P}\{\nu(K_2\varepsilon,\delta) > N(\varepsilon;\delta)\} \leq K_3\delta. \tag{15}$$

If $\nu(\varepsilon,\delta)$ is a WE stopping time, then Eq. (14) says that we can solve the empirical approximation problem with accuracy $K_1\varepsilon$ and confidence $1-\delta$; Eq. (15) says that, with probability at most $1-\delta$, the sample complexity will be less than the sample complexity of empirical approximation with accuracy ε/K_2 and confidence $1-\delta$.

If $N(\varepsilon;\delta) \geq \overline{n}(\varepsilon,\delta)$, then $N(\varepsilon,\delta) \in \mathscr{T}(\varepsilon,\delta)$. Hence, any WE stopping time is also SE. The converse, however, is not true.

3.1 A strongly efficient sequential learning algorithm

Let $\{Z_n\}_{n=1}^{\infty}$ be an infinite sequence of i.i.d. draws from some $P \in \mathscr{P}$; let $\{\sigma_n\}_{n=1}^{\infty}$ be an i.i.d. Rademacher sequence independent of $\{Z_n\}$. Choose

$$\overline{n}(\varepsilon,\delta) \geq \left\lfloor \frac{2}{\varepsilon^2} \log \frac{2}{\delta(1-e^{-\varepsilon^2/2})} \right\rfloor + 1 \tag{16}$$

and let

$$\nu(\varepsilon,\delta) \triangleq \min\left\{n \geq \overline{n}(\varepsilon,\delta) : r_n(\mathscr{F}(Z^n)) \leq \varepsilon\right\}. \tag{17}$$

This is clearly a stopping time for each $\varepsilon > 0$ and each $\delta \in (0,1)$.

Theorem 1. *The family* $\{\nu(\varepsilon,\delta) : \varepsilon > 0, \delta \in (0,1)\}$ *defined in* (17) *with* $\overline{n}(\varepsilon,\delta)$ *set according to* (16) *is SE for any class* $\mathscr{F}$ *of measurable functions* $f : \mathsf{Z} \to [0,1]$ *and* $\mathscr{P} = \mathscr{P}(\mathsf{Z})$ *with* $K_1 = 5, K_2 = 6, K_3 = 1$.

Proof. Let $\overline{n} = \overline{n}(\varepsilon,\delta)$. We will first show that, for any $P \in \mathscr{P}(\mathsf{Z})$,

$$\|P_n - P\|_{\mathscr{F}} \leq 2r_n(\mathscr{F}(Z^n)) + 3\varepsilon, \qquad \forall n \geq \overline{n} \tag{18}$$

with probability at least $1-\delta$. Since for $n = \nu(\varepsilon,\delta) \geq \overline{n}$ we have $r_n(\mathscr{F}(Z^n)) \leq \varepsilon$, we will immediately be able to conclude that

$$\mathbb{P}\left\{\|P_{\nu(\varepsilon,\delta)} - P\|_{\mathscr{F}} \geq 5\varepsilon\right\} \leq \delta,$$

which will imply that $\nu(\varepsilon,\delta) \in \mathscr{T}(5\varepsilon,\delta)$. Now we prove (18). First of all, applying Lemma 2 and the union bound, we can write

$$\begin{aligned}\mathbb{P}\left\{\bigcup_{n\geq\overline{n}}\left\{r_n(\mathscr{F}(Z^n)) \geq \mathbb{E}R_n(\mathscr{F}(Z^n)) + \varepsilon\right\}\right\} &\leq \sum_{n\geq\overline{n}} e^{-n\varepsilon^2/2} \\ &= e^{-\overline{n}\varepsilon^2/2}\sum_{n\geq 0} e^{-n\varepsilon^2/2} \\ &= \frac{e^{-\overline{n}\varepsilon^2/2}}{1-e^{-\varepsilon^2/2}} \\ &\leq \delta/2.\end{aligned}$$

From the symmetrization inequality (5), we know that $\mathbb{E}\|P_n - P\|_{\mathscr{F}} \leq 2\mathbb{E}R_n(\mathscr{F}(Z^n))$. Moreover, using (6) and the union bound, we can write

$$\begin{aligned}\mathbb{P}\left\{\bigcup_{n\geq\overline{n}}\left\{\|P_n - P\|_{\mathscr{F}} \geq \mathbb{E}\|P_n - P\|_{\mathscr{F}} + \varepsilon\right\}\right\} &\leq \sum_{n\geq\overline{n}} e^{-2n\varepsilon^2} \\ &\leq \sum_{n\geq\overline{n}} e^{-n\varepsilon^2/2} \\ &\leq \delta/2.\end{aligned}$$

Therefore, with probability at least $1-\delta$,

$$\|P_n - P\|_{\mathscr{F}} \leq \mathbb{E}\|P_n - P\|_{\mathscr{F}} + \varepsilon \leq 2\mathbb{E}R_n(\mathscr{F}(Z^n)) + \varepsilon \leq 2r_n(\mathscr{F}(Z^n)) + 3\varepsilon, \qquad \forall n \geq \overline{n}$$

which is (18). This shows that (12) holds for $\nu(\varepsilon,\delta)$ with $K_1 = 5$.

Next, we will prove that, for any $P \in \mathscr{P}(\mathrm{Z})$,

$$\mathbb{P}\left\{\min_{\overline{n}\leq n<\nu(6\varepsilon,\delta)} \|P_n - P\|_{\mathscr{F}} < \varepsilon\right\} \leq \delta. \tag{19}$$

In other words, (19) says that, with probability at least $1-\delta$, $\|P_n - P\|_{\mathscr{F}} \geq \varepsilon$ for all $\overline{n} \leq n < \nu(6\varepsilon,\delta)$. This means that, for any $\tau \in \mathscr{T}(\varepsilon,\delta)$, $\nu(6\varepsilon,\delta) \leq \tau$ with probability at least $1-\delta$, which will give us (13) with $K_2 = 6$ and $K_3 = 1$.

To prove (19), we have by (7) and the union bound that

$$\mathbb{P}\left\{\bigcup_{n\geq\overline{n}}\left\{\|P_n - P\|_{\mathscr{F}} \leq \mathbb{E}\|P_n - P\|_{\mathscr{F}} - \varepsilon\right\}\right\} \leq \delta/2.$$

By the desymmetrization inequality (8), we have

$$\mathbb{E}\|P_n - P\|_{\mathscr{F}} \geq \frac{1}{2}\mathbb{E}R_n(\mathscr{F}(Z^n)) - \frac{1}{2\sqrt{n}}, \qquad \forall n.$$

Finally, by the concentration inequality (10) and the union bound,

$$\mathbb{P}\left\{\bigcup_{n\geq\overline{n}}\left\{r_n(\mathscr{F}(Z^n)) \geq \mathbb{E}R_n(\mathscr{F}(Z^n)) + \varepsilon\right\}\right\} \leq \delta/2.$$

Therefore, with probability at least $1-\delta$,

$$\|P_n - P\|_{\mathscr{F}} \geq \frac{1}{2} r_n(\mathscr{F}(Z^n)) - \frac{1}{2\sqrt{n}} - \frac{3\varepsilon}{2}, \qquad \forall n \geq \overline{n}.$$

If $\overline{n} \leq n < \nu(6\varepsilon,\delta)$, then $r_n(\mathscr{F}(Z^n)) > 6\varepsilon$. Therefore, using the fact that $n \geq \overline{n}$ and $\overline{n}(\varepsilon,\delta)^{-1/2} \leq \varepsilon$, we see that, with probability at least $1-\delta$,

$$\|P_n - P\|_{\mathscr{F}} > \frac{3\varepsilon}{2} - \frac{1}{2\sqrt{n}} \geq \frac{3\varepsilon}{2} - \frac{1}{2\sqrt{\overline{n}}} \geq \varepsilon, \qquad \overline{n} \leq n < \nu(6\varepsilon,\delta).$$

This proves (19), and we are done. □

3.2 A weakly efficient sequential learning algorithm

Now choose

$$\overline{n}(\varepsilon,\delta) \geq \left\lfloor \frac{2}{\varepsilon^2} \log \frac{4}{\delta} \right\rfloor + 1, \tag{20}$$

for each $k = 0,1,2,\ldots$ let $n_k \triangleq 2^k \overline{n}(\varepsilon,\delta)$, and let

$$\nu(\varepsilon,\delta) \triangleq \min\left\{ n_k : r_{n_k}(\mathscr{F}(Z^{n_k})) \leq \varepsilon \right\}. \tag{21}$$

Theorem 2. *The family* $\{\nu(\varepsilon,\delta) : \varepsilon > 0, \delta \in (0,1/2)\}$ *defined in* (21) *with* $\overline{n}(\varepsilon,\delta)$ *set according to* (20) *is WE for any class* $\mathscr{F}$ *of measurable functions* $f : \mathsf{Z} \to [0,1]$ *and* $\mathscr{P} = \mathscr{P}(\mathsf{Z})$ *with* $K_1 = 5$, $K_2 = 18$, $K_3 = 3$.

Proof. As before, let $\overline{n} = \overline{n}(\varepsilon,\delta)$. The proof of (14) is similar to what we have done in the proof of Theorem 1, except we use the bounds

$$\begin{aligned}
\mathbb{P}\left\{ \bigcup_{k=0}^{\infty} \left\{ r_{n_k}(\mathscr{F}(Z^{n_k})) \geq \mathbb{E} R_{n_k}(\mathscr{F}(Z^{n_k})) + \varepsilon \right\} \right\} &\leq \sum_{k=0}^{\infty} e^{-2^k \overline{n} \varepsilon^2/2} \\
&= e^{-\overline{n}\varepsilon^2/2} + e^{-\overline{n}\varepsilon^2/2} \sum_{k=1}^{\infty} e^{-\frac{\overline{n}\varepsilon^2}{2}(2^k-1)} \\
&\leq e^{-\overline{n}\varepsilon^2/2} + e^{-\overline{n}\varepsilon^2/2} \sum_{k=1}^{\infty} e^{-(2^k-1)} \\
&\leq e^{-\overline{n}\varepsilon^2/2} + e^{-\overline{n}\varepsilon^2/2} \sum_{k=1}^{\infty} e^{-k} \\
&\leq 2e^{-\overline{n}\varepsilon^2/2} \\
&\leq \delta/2,
\end{aligned}$$

where in the third step we have used the fact that $\overline{n}\varepsilon^2/2 \geq 1$. Similarly,

$$\mathbb{P}\left\{ \bigcup_{k=0}^{\infty} \left\{ \|P_{n_k} - P\|_{\mathscr{F}} \leq \mathbb{E}\|P_{n_k} - P\|_{\mathscr{F}} + \varepsilon \right\} \right\} \leq \delta^2.$$

Therefore,

$$\|P_{n_k} - P\|_{\mathscr{F}} \le 2r_{n_k}(\mathscr{F}(Z^{n_k})) + 3\varepsilon, \qquad \forall k = 0,1,2,\ldots$$

and consequently

$$\mathbb{P}\left\{\|P_{\nu(\varepsilon,\delta)} - P\|_{\mathscr{F}} \ge 5\varepsilon\right\} \le \delta,$$

which proves (14).

Now we prove (15). Let $N = N(\varepsilon,\delta)$, the sample complexity of empirical approximation that we have defined in (3). Let us choose k so that $n_k \le N < n_{k+1}$, which is equivalent to $2^k\overline{n} \le N < 2^{k+1}\overline{n}$. Then

$$\mathbb{P}\{\nu(18\varepsilon,\delta) > N\} \le \mathbb{P}\{\nu(18\varepsilon,\delta) > n_k\}.$$

We will show that the probability on the right-hand side is less than 3δ. First of all, since $N \ge \overline{n}$ (by hypothesis), we have $n_k \ge \overline{n}/2 \ge 1/\varepsilon^2$. Therefore, with probability at least $1-\delta$

$$\|P_{n_k} - P\|_{\mathscr{F}} \ge \frac{1}{2}r_{n_k}(\mathscr{F}(Z^{n_k})) - \frac{1}{2\sqrt{n_k}} - \frac{9\varepsilon}{2} \ge \frac{1}{2}r_{n_k}(\mathscr{F}(Z^{n_k})) - 5\varepsilon. \tag{22}$$

If $\nu(18\varepsilon,\delta) > n_k$, then by definition $r_{n_k}(\mathscr{F}(Z^{n_k})) > 18\varepsilon$. Writing $r_{n_k} = r_{n_k}(\mathscr{F}(Z^{n_k}))$ for brevity, we see get

$$\begin{aligned}
\mathbb{P}\{\nu(18\varepsilon,\delta) > n_k\} &\le \mathbb{P}\left\{r_{n_k} > 18\varepsilon\right\} \\
&= \mathbb{P}\left\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathscr{F}} \ge 18\varepsilon\right\} + \mathbb{P}\left\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathscr{F}} < 4\varepsilon\right\} \\
&\le \mathbb{P}\left\{\|P_{n_k} - P\|_{\mathscr{F}} \ge 4\varepsilon\right\} + \mathbb{P}\left\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathscr{F}} < 4\varepsilon\right\}.
\end{aligned}$$

If $r_{n_k} > 18\varepsilon$ but $\|P_{n_k} - P\|_{\mathscr{F}} < 4\varepsilon$, the event in (22) cannot occur. Indeed, suppose it does. Then it must be the case that $4\varepsilon > 9\varepsilon - 5\varepsilon = 4\varepsilon$, which is a contradiction. Therefore,

$$\mathbb{P}\left\{r_{n_k} > 18\varepsilon, \|P_{n_k} - P\|_{\mathscr{F}} < 4\varepsilon\right\} \le \delta,$$

and hence

$$\mathbb{P}\{\nu(18\varepsilon,\delta) > n_k\} \le \mathbb{P}\left\{\|P_{n_k} - P\|_{\mathscr{F}} \ge 4\varepsilon\right\} + \delta.$$

For each $f \in \mathscr{F}$ and each $n \in \mathbb{N}$ define

$$S_n(f) \triangleq \sum_{i=1}^{n} [f(Z_i) - P(f)]$$

and let $\|S_n\|_{\mathscr{F}} \triangleq \sup_{f\in\mathscr{F}} |S_n(f)|$. Then

$$\mathbb{P}\left\{\|P_{n_k} - P\|_{\mathscr{F}} \ge 4\varepsilon\right\} = \mathbb{P}\left\{\|S_{n_k}\|_{\mathscr{F}} \ge 4\varepsilon n_k\right\} \le \mathbb{P}\left\{\|S_{n_k}\|_{\mathscr{F}} \ge 2\varepsilon N\right\}.$$

Since $n_k \le N$, the $\mathscr{F}$-indexed stochastic processes $S_{n_k}(f)$ and $S_N(f) - S_{n_k}(f)$ are independent. Therefore, we use a technical result stated as Lemma 4 in the appendix with $\xi_1 = S_{n_k}$ and $\xi_2 = S_N(f) - S_{n_k}(f)$ to write

$$\mathbb{P}\left\{\|S_{n_k}\|_{\mathscr{F}} \ge 2\varepsilon N\right\} \le \frac{\mathbb{P}\{\|S_N\|_{\mathscr{F}} \ge \varepsilon N\}}{\inf_{f\in\mathscr{F}} \mathbb{P}\left\{|S_N(f) - S_{n_k}(f)| \le \varepsilon N\right\}}.$$

By definition of $N = N(\varepsilon, \delta)$, the probability in the numerator is at most δ. To analyze the probability in the denominator, we use Hoeffding's inequality to get

$$\begin{aligned}\inf_{f \in \mathscr{F}} \mathbb{P}\left\{|S_N(f) - S_{n_k}(f)| \le \varepsilon N\right\} &= 1 - \sup_{f \in \mathscr{F}} \mathbb{P}\left\{|S_N(f) - S_{n_k}(f)| > \varepsilon N\right\} \\ &\ge 1 - 2e^{-N\varepsilon^2/2} \\ &\ge 1 - \delta.\end{aligned}$$

Therefore,

$$\mathbb{P}\{\nu(18\varepsilon, \delta) > n_k\} \le \frac{\delta}{1-\delta} + \delta \le 3\delta$$

for $\delta < 1/2$. Therefore, $\{\nu(\varepsilon, \delta) : \varepsilon \in (0,1), \delta \in (0, 1/2)\}$ is WE with $K_1 = 5, K_2 = 18, K_3 = 3$. □

4 A sequential algorithm for stochastic simulation

Armed with these results on sequential learning algorithms, we can take up the question of constructing efficient simulation strategies. We fix an accuracy parameter $\varepsilon > 0$, a confidence parameter $\delta \in (0,1)$, and a level parameter $\alpha \in (0,1)$. Given two probability distributions, P on the input space Z and Q on the parameter space Θ, we draw a large i.i.d. sample $Z_1, \ldots, Z_n$ from P and a large i.i.d. sample $\theta_1, \ldots, \theta_m$ from Q. We then compute

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \{\theta_1, \ldots, \theta_m}} L_n(\theta),$$

where

$$L_n(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, \theta).$$

The goal is to pick n and m large enough so that, with probability at least $1 - \delta$, $\widehat{\theta}$ is an ε-minimizer of L to level α, i.e., with probability at least $1 - \delta$ there exists some set $\Lambda \subset \Theta$ with $Q(\Lambda) \le \alpha$, such that Eq. (2) holds with probability at least $1 - \delta$.

To that end, consider the following algorithm based on Theorem 2, proposed by Koltchinskii et al. [KAA+00a, KAA+00b]:

Algorithm 1
choose positive integers m and n such that $m \geq \frac{\log(2/\delta)}{\log[1/(1-\alpha)]}$ and $n \geq \lfloor \frac{50}{\varepsilon^2} \log \frac{8}{\delta} \rfloor + 1$ draw m independent samples $\theta_1, \ldots, \theta_m$ from Q draw n independent samples $Z_1, \ldots, Z_n$ from P_Z evaluate the stopping variable $\gamma = \max_{1 \leq j \leq m} \left\lvert \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(Z_i, \theta_j) \right\rvert$ where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher r.v.'s independent of θ^m and Z^n if $\gamma > \varepsilon/5$, then add n more i.i.d. samples from P_Z and repeat else stop and output $\widehat{\theta} = \arg\min_{\theta \in \{\theta_1, \ldots, \theta_n\}} L_n(\theta)$

Then we claim that, with probability at least $1-\delta$, $\widehat{\theta}$ is an ε-minimizer of L to level α. To see this, we need the following result [Vid03, Lemma 11.1]:

Lemma 3. *Let Q be a probability distribution on the parameter set Θ, and let $h : \Theta \to \mathbb{R}$ be a (measurable) real-valued function on Θ, bounded from above, i.e., $h(\theta) < +\infty$ for all $\theta \in \Theta$. Let $\theta_1, \ldots, \theta_m$ be m i.i.d. samples from Q, and let*

$$\bar{h}(\theta^m) \triangleq \max_{1 \leq j \leq m} h(\theta_m).$$

Then for any $\alpha \in (0,1)$

$$Q\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right) \leq \alpha \tag{23}$$

with probability at least $1-(1-\alpha)^m$.

Proof. For each $c \in \mathbb{R}$, let

$$F(c) \triangleq \mathbb{P}\left(\{\theta \in \Theta : h(\theta) \leq c\}\right).$$

Note that F is the CDF of the random variable $\xi = h(\theta)$ with $\theta \sim Q$. Therefore, it is right-continuous, i.e., $\lim_{c' \searrow c} F(c') = F(c)$. Now define

$$c_\alpha \triangleq \inf\{c : F(c) \geq 1-\alpha\}.$$

Since F is right-continuous, $F(c_\alpha) \geq 1-\alpha$. Moreover, if $c < c_\alpha$, then $F(c) < 1-\alpha$. Now let us suppose that $\bar{h}(\theta^m) \geq c_\alpha$. Then, since F is monotone nondecreasing,

$$\mathbb{P}\left(\left\{\theta \in \Theta : h(\theta) \leq \bar{h}(\theta^m)\right\}\right) = F\left(\bar{h}(\theta^m)\right) \geq F(c_\alpha) \geq 1-\alpha,$$

or, equivalently, if $\bar{h}(\theta^m) \geq c_\alpha$, then

$$\mathbb{P}\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right) \leq \alpha.$$

Therefore, if θ^m is such that

$$\mathbb{P}\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right) > \alpha,$$

then it must be the case that $\bar{h}(\theta^m) < c_\alpha$, which in turn implies that $F(\bar{h}(\theta^m)) < 1-\alpha$, the complement of the event in (23). But $\bar{h}(\theta^m) < c_\alpha$ means that $h(\theta_j) < c_\alpha$ for every $1 \le j \le m$. Since the θ_j's are independent, the events $\{h(\theta_j) < c_\alpha\}$ are independent, and each occurs with probability at most $1-\alpha$. Therefore,

$$\mathbb{P}\left(\left\{\theta^m \in \Theta^m : Q\left(\left\{\theta \in \Theta : h(\theta) > \bar{h}(\theta^m)\right\}\right)\right\}\right) \le (1-\alpha)^m,$$

which is what we intended to prove. □

We apply this lemma to the function $h(\theta) = -L(\theta)$. Then, provided m is chosen as described in Algorithm 1, we will have

$$Q\left(\left\{\theta \in \Theta : L(\theta) < \min_{1\le j\le j} L(\theta_m)\right\}\right) \le \delta/2.$$

Now consider the *finite* class of functions $\mathscr{F} = \{f_j(z) = \ell(z,\theta_j) : 1 \le j \le m\}$. By Theorem 2, the final output $\widehat{\theta} \in \{\theta_1,\ldots,\theta_m\}$ will satisfy

$$\left|L(\widehat{\theta}) - \min_{1\le j\le m} L(\theta_j)\right| \le \varepsilon$$

with probability at least $1-\delta/2$. Hence, with probability at least $1-\delta$ there exists a set $\Lambda \subset \Theta$ with $Q(\Lambda) \le \alpha$, such that (2) holds. Moreover, the total number of samples used up by Algorithm 1 will be, with probability at least $1-3\delta/2$, no more than

$$N_{\mathscr{F},P_Z}(\varepsilon/18,\delta/2) \equiv \min\{n \in \mathbb{N} : \mathbb{P}\left(\|P_n - P_Z\|_{\mathscr{F}} > \varepsilon/18\right) < \delta/2\}.$$

We can estimate $N_{\mathscr{F},P_Z}(\varepsilon/18,\delta/2)$ as follows. First of all, the function

$$\Delta(Z^n) \triangleq \|P_n - P_Z\|_{\mathscr{F}} \equiv \max_{1\le j\le m} |P_n(f_j) - P_Z(f_j)|$$

has bounded differences with $c_1 = \ldots = c_n = 1/n$. Therefore, by McDiarmid's inequality

$$\mathbb{P}\left(\Delta(Z^n) \ge \mathbb{E}\Delta(Z^n) + t\right) \le e^{-2nt^2}, \qquad \forall\, t > 0.$$

Secondly, since the class $\mathscr{F}$ is finite with $|\mathscr{F}| = m$, the symmetrization inequality (5) and the Finite Class Lemma give the bound

$$\mathbb{E}\|P_n - P_Z\|_{\mathscr{F}} \le 4\sqrt{\frac{\log m}{n}}.$$

Therefore, if we choose $t = \varepsilon/18 - 4\sqrt{n^{-1}\log m}$ and n is large enough so that $t > \varepsilon/20$ (say), then

$$\mathbb{P}\left(\|P_n - P\|_{\mathscr{F}} > \varepsilon/18\right) \le e^{-n\varepsilon^2/200}.$$

Hence, a fairly conservative estimate is

$$N_{\mathscr{F},P_Z}(\varepsilon/18,\delta/2) \le \max\left\{\left\lfloor\frac{200}{\varepsilon^2}\log\frac{2}{\delta}\right\rfloor + 1, \left\lfloor\left(\frac{720}{\varepsilon}\right)^2 \log m\right\rfloor + 1\right\}$$

It is instructive to compare Algorithm 1 with a simple Monte Carlo strategy:

Algorithm 0
choose positive integers m and n such that $m \geq \frac{\log(2/\delta)}{\log[1/(1-\alpha)]}$ and $n \geq \frac{1}{2\varepsilon^2}\log\frac{4m}{\delta}$ draw m independent samples $\theta_1,\ldots,\theta_m$ from Q draw n independent samples $Z_1,\ldots,Z_n$ from P_Z for $j=1$ to m compute $L_n(\theta_j) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i,\theta_j)$ end for output $\widehat{\theta} = \arg\min_{\theta\in\{\theta_1,\ldots,\theta_m} L_n(\theta_j)$

The selection of m is guided by the same considerations as in Algorithm 1. Moreover, for each $1 \leq j \leq m$, $L_n(\theta_j)$ is an average of n independent random variables $\ell(Z_i,\theta_j) \in [0,1]$, and $L(\theta_j) = \mathbb{E}L_n(\theta_j)$. Hence, Hoeffding's inequality says that

$$\mathbb{P}\left(\left\{Z^n \in \mathsf{Z}^n : \left|L_n(\theta_j) - L(\theta_j)\right| > \varepsilon\right\}\right) \leq 2e^{-2n\varepsilon^2}.$$

If we choose n as described in Algorithm 0, then

$$\begin{aligned}\mathbb{P}\left(\left|L_n(\widehat{\theta}) - \min_{1\leq j\leq m} L(\theta_j)\right| > \varepsilon\right) &\leq \mathbb{P}\left(\bigcup_{j=1}^m \left|L_n(\theta_j) - L(\theta_j)\right| > \varepsilon\right)\\ &\leq \sum_{j=1}^m \mathbb{P}\left(\left|L_n(\theta_j) - L(\theta_j)\right| > \varepsilon\right)\\ &\leq \delta/2.\end{aligned}$$

Hence, with probability at least $1-\delta$ there exists a set $\Lambda \subset \Theta$ with $Q(\Lambda) \leq \alpha$, so that (2) holds. It may seem at first glance that Algorithm 0 is more efficient than Algorithm 1. However, this is not the case in high-dimensional situations. There, one can actually show that, with probability practically equal to one, the empirical minimum of L can be *much larger* than the true minimum (cf. [KAA+00b] for a very vivid numerical illustration). This is an instance of the so-called *Curse of Dimensionality*, which adaptive schemes like Algorithm 1 can often avoid.

A Technical lemma

Lemma 4. *Let $\{\xi_1(f) : f \in \mathscr{F}\}$ and $\{\xi_2(f) : f \in \mathscr{F}\}$ be two independent $\mathscr{F}$-indexed stochastic processes with*

$$\|\xi_j\|_{\mathscr{F}} \triangleq \sup_{f\in\mathscr{F}} |\xi_j(f)| < \infty, \qquad j = 1,2.$$

Then for all $t > 0, c > 0$

$$\mathbb{P}\{\|\xi_1\|_{\mathscr{F}} \geq t + c\} \leq \frac{\mathbb{P}\{\|\xi_1 - \xi_2\|_{\mathscr{F}} \geq t\}}{\inf_{f\in\mathscr{F}} \mathbb{P}\left\{|\xi_2(f)| \leq c\right\}}. \tag{24}$$

Proof. If $\|\xi_1\|_{\mathscr{F}} \geq t + c$, then there exists some $f \in \mathscr{F}$, such that $|\xi_1(f)| \geq t + c$. Then for this particular f by the triangle inequality we see that

$$|\xi_2(f)| \leq c \qquad \Rightarrow \qquad |\xi_1(f) - \xi_2(f)| \geq t$$

Therefore,

$$\inf_{f\in\mathscr{F}} \mathbb{P}_{\xi_2}\Big\{|\xi_2(f)| \leq c\Big\} \leq \mathbb{P}_{\xi_2}\Big\{|\xi_2(f)| \leq c\Big\} \leq \mathbb{P}_{\xi_2}\Big\{\big|\xi_1(f) - \xi_2(f)\big| \geq t\Big\} \leq \mathbb{P}_{\xi_2}\Big\{\|\xi_1 - \xi_2\|_{\mathscr{F}} \geq t\Big\}.$$

The leftmost and the rightmost terms in the above inequality do not depend on the particular f, and the inequality between them is valid on the event $\{\|\xi_1\|_{\mathscr{F}} \geq t + c\}$. Therefore, integrating the two sides w.r.t. ξ_1 on this event, we get

$$\inf_{f\in\mathscr{F}} \mathbb{P}_{\xi_2}\Big\{|\xi_2(f)| \leq c\Big\} \cdot \mathbb{P}_{\xi_1}\Big\{\|\xi_1\|_{\mathscr{F}} \geq t + c\Big\} \leq \mathbb{P}_{\xi_1,\xi_2}\Big\{\|\xi_1 - \xi_2\|_{\mathscr{F}} \geq t\Big\}.$$

Rearranging, we get (24). □

References

[KAA+00a] V. Koltchinskii, C. T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Improved sample complexity estimates for statistical learning control of uncertain systems. *IEEE Transactions on Automatic Control*, 45(12):2383–2388, 2000.

[KAA+00b] V. Koltchinskii, C. T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Statistical learning control of uncertain systems: it is better than it seems. Technical Report EECE-TR-00-001, University of New Mexico, April 2000.

[Vid98] M. Vidyasagar. Statistical learning theory and randomized algorithms for control. *IEEE Control Magazine*, 18(6):162–190, 1998.

[Vid01] M. Vidyasagar. Randomized algorithms for robust controller synthesis using statistical learning theory. *Automatica*, 37:1515–1528, 2001.

[Vid03] M. Vidyasagar. *Learning and Generalization*. Springer, 2 edition, 2003.