Topics in High-Dimensional Statistics

Lecture 4: Empirical Risk Minimization II
Concentration, Symmetrization and Contraction

Contents

1 Empirical processes

Let $\mathcal{Z}$ be a measurable space, P be a probability measure on $\mathcal{Z}$ and $\{Z_i\}_{i=1}^n$ be i.i.d. with distribution P. Given a collection $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$, the empirical process of P indexed by $\mathcal{F}$ is the stochastic process

$$(X_f)_{f\in\mathcal{F}},$$

defined by

$$X_f := \int f \,\mathrm{d}P - \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

In the sequel, we will use a convenient and commonly used notation to simplify computations. Namely, given two probability measures μ, ν over $\mathcal{Z}$, we will denote

$$(\mu - \nu)f := \int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu.$$

With this convention, the empirical process of P indexed by f writes

$$X_f = (P - P_n)f,$$

where P_n refers to the empirical measure

$$P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

As will be clarified in this lecture, and the following, the study of empirical processes is central to the analysis of empirical risk minimization. In particular, the supremum

$$|P - P_n|_{\mathcal{F}} := \sup_{f\in\mathcal{F}}(P - P_n)f,$$

plays a central role as described next.

2 First connection between ERM and empirical processes

Consider the problem of empirical risk minimization introduced in the previous lecture. Using the notation previously introduced, suppose for simplicity that there exists

$$\bar{\theta} \in \arg\min_{\theta\in\Theta} R(\theta).$$

Then, we have the following lemma.

Lemma 2.1. *The excess risk $\mathcal{E}(\theta_n)$ of the empirical risk minimizer θ_n over the parameter set Θ satisfies*

$$\mathcal{E}(\theta_n) \leq |P - P_n|_{\mathcal{F}},$$

where

$$\mathcal{F} := \{\gamma(\theta, .) - \gamma(\bar{\theta}, .) : \theta \in \Theta\},$$

where $\gamma(\theta, .) - \gamma(\bar{\theta}, .)$ refers to the function $z \in \mathcal{Z} \mapsto \gamma(\theta, z) - \gamma(\bar{\theta}, z)$.

Proof. First, observe that using the notation introduced in Section 1 and the remarks made in Section 2 of the previous lecture, we have that for any $\theta \in \Theta$, possibly depending on the data,

$$R(\theta) = P\gamma(\theta, .) \quad \text{and} \quad R_n(\theta) = P_n\gamma(\theta, .).$$

By definition of the excess risk and $\bar{\theta}$, we have

$$\begin{aligned} \mathcal{E}(\theta_n) &= R(\theta_n) - R(\bar{\theta}) \\ &= P(\gamma(\theta_n, .) - \gamma(\bar{\theta}, .)). \end{aligned}$$

Now, observe that by definition of the empirical risk minimizer θ_n, we have for all $\theta \in \Theta$ that $R_n(\theta_n) \leq R_n(\theta)$, which can be written equivalently as

$$P_n(\gamma(\theta_n, .) - \gamma(\theta, .)) \leq 0.$$

Using this observation for $\theta = \bar{\theta}$ implies that

$$\begin{aligned} \mathcal{E}(\theta_n) &= P(\gamma(\theta_n, .) - \gamma(\bar{\theta}, .)) \\ &\leq (P - P_n)(\gamma(\theta_n, .) - \gamma(\bar{\theta}, .)) \\ &\leq \sup_{\theta \in \Theta}(P - P_n)(\gamma(\theta, .) - \gamma(\bar{\theta}, .)) \\ &= |P - P_n|_{\mathcal{F}}, \end{aligned} \tag{2.1}$$

which concludes the proof. □

The rest of the lecture will present tools to further control the quantity

$$|P - P_n|_{\mathcal{F}},$$

in the context of general empirical processes.

3 Concentration

The first result we present follows from an application of McDiarmid's inequality. Below, $\mathcal{F}$ denotes an arbitrary collection of measurable functions $f : \mathcal{Z} \to \mathbb{R}$.

Theorem 3.1. *Suppose that all functions in $\mathcal{F}$ are $[a, b]$-valued, for some $a < b$. Then, for all $n \geq 1$ and all $\delta \in (0, 1)$, the inequality*

$$|P - P_n|_{\mathcal{F}} \leq \mathbb{E}|P - P_n|_{\mathcal{F}} + (b - a)\sqrt{\frac{1}{2n} \log\left(\frac{1}{\delta}\right)},$$

hold with probability larger than $1 - \delta$.

Proof. Introduce the function $g : \mathcal{Z}^n \to \mathbb{R}$ defined by

$$g(z_1, \ldots, z_n) = \sup_{f \in \mathcal{F}} \left(Pf - \frac{1}{n} \sum_{i=1}^{n} f(z_i) \right).$$

Then, for all $1 \leq i \leq n$ and all $z_1, \ldots, z_{i-1}, z_i, z_i', z_{i+1}, \ldots, z_n \in \mathcal{Z}$, it may be easily verified that

$$\begin{aligned}
&|g(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \\
&\leq \frac{1}{n} \sup_{f \in \mathcal{F}} |f(z_i) - f(z_i')| \\
&\leq \frac{b - a}{n}.
\end{aligned}$$

The result therefore follows from an application of McDiarmid's inequality. $\square$

The last result exploits only the boundedness property of the functions in $\mathcal{F}$. Building upon the pioneering work of Talagrand, Bousquet provides in [2] a major improvement of the previous result which involves a notion of variance of the empirical process defined by

$$\sigma^2(\mathcal{F}) := \sup_{f \in \mathcal{F}} \mathrm{Var}_P f \quad \text{where} \quad \mathrm{Var}_P f := P(f - Pf)^2. \tag{3.1}$$

The term $\sigma^2(\mathcal{F})$ is sometimes called the wimpy variance of the empirical process and appears as one of the natural means to define the variance of the empirical process [see, for instance, page 305 in 1, for alternative definitions and dicussions].

Theorem 3.2 (2)**.** *Suppose that all functions in $\mathcal{F}$ are $[a,b]$-valued, for some $a<b$. Then, for all $n \geq 1$ and all $t>0$,*

$$|P-P_n|_{\mathcal{F}} \leq \mathbb{E}|P-P_n|_{\mathcal{F}} + \sqrt{\frac{2t}{n}\left(\sigma^2(\mathcal{F}) + 2(b-a)\,\mathbb{E}|P-P_n|_{\mathcal{F}}\right)} + \frac{(b-a)t}{3n},$$

with probability larger than $1-e^{-t}$.

4 Symmetrization

Concentration properties of the supremum $|P-P_n|_{\mathcal{F}}$ allow essentially to reduce its control to that of its expectation. The rest of the lecture will therefore focus on bounding the quantity

$$\mathbb{E}|P-P_n|_{\mathcal{F}}.$$

To this aim, an important result is know as the symmetrization principle that draws a link between the expected suprema of empirical processes and that of Rademacher processes which, via conditionning, happen to be sub-gaussian processes as discussed below.

To discuss this principle, introduce independent random signs $\sigma_1,\ldots,\sigma_n$ (also referred to as Rademacher random variables), i.e., random variables such that $\mathbb{P}(\sigma_i=-1)=\mathbb{P}(\sigma_i=1)=1/2$, supposed independent from the Z_i's. The stochastic process

$$(\mathfrak{R}_n(f))_{f\in\mathcal{F}},$$

defined by

$$\mathfrak{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i), \tag{4.1}$$

is called the Rademacher process indexed by $\mathcal{F}$.

By the symmetrization principle, one usually refers to a set of results relating the supremum of the empirical process to that of the Rademacher process. While early forms of the symmetrization principle where developped in the works of [6], modern versions of the principle find their roots in the influential work of [3]. A general variant of the principle may be stated as follows.

Theorem 4.1. *For any convex and non-decreasing function $G:\mathbb{R}\to\mathbb{R}$ and all $n\geq 1$,*

$$\mathbb{E}\,G\left[\sup_{f\in\mathcal{F}}(P-P_n)f\right] \leq \mathbb{E}\,G\left[2\sup_{f\in\mathcal{F}}\mathfrak{R}_n(f)\right]. \tag{4.2}$$

Proof. Consider an artificial set (or ghost sample) $Z'_1, \ldots, Z'_n$ of independent random variables with same distribution P as (and independent from) the original sample $Z_1, \ldots, Z_n$. Suppose, without loss of generality, that the variables $Z'_1, \ldots, Z'_n$ are also independent from the random signs $\sigma_1, \ldots, \sigma_n$. For brevity, define the notation $\mathbb{E}_z[.] = \mathbb{E}[\,.\,|Z_1, \ldots, Z_n]$ and set, for all $f \in \mathcal{F}$,

$$P'_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z'_i} \quad \text{and} \quad \mathfrak{R}'_n(f) = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z'_i).$$

Then, since $Pf = \mathbb{E}_z[P'_n f]$, we obtain

$$\begin{aligned}
\mathbb{E}G\left[\sup_{f\in\mathcal{F}}(P - P_n)f\right] &= \mathbb{E}G\left[\sup_{f\in\mathcal{F}} \mathbb{E}_z(P'_n - P_n)f\right] \\
&\le \mathbb{E}G\left[\mathbb{E}_z \sup_{f\in\mathcal{F}}(P'_n - P_n)f\right] && (4.3)\\
&\le \mathbb{E}G\left[\sup_{f\in\mathcal{F}}(P'_n - P_n)f\right], && (4.4)
\end{aligned}$$

where (4.3) follows by monotonicity of G and where (4.4) follows from Jensen's inequality. Next, since for all $f \in \mathcal{F}$ the variables $(P'_n - P_n)f$ and $\mathfrak{R}'_n(f) - \mathfrak{R}_n(f)$, have same distribution, we deduce from (4.4) that

$$\begin{aligned}
&\mathbb{E}G\left[\sup_{f\in\mathcal{F}}(P - P_n)f\right] \\
&\le \mathbb{E}G\left[\sup_{f\in\mathcal{F}}(\mathfrak{R}'_n(f) - \mathfrak{R}_n(f))\right] \\
&\le \mathbb{E}G\left[\sup_{f\in\mathcal{F}}\mathfrak{R}'_n(f) + \sup_{f\in\mathcal{F}}(-\mathfrak{R}_n(f))\right] && (4.5)\\
&\le \frac{1}{2}\mathbb{E}G\left[2\sup_{f\in\mathcal{F}}\mathfrak{R}'_n(f)\right] + \frac{1}{2}\mathbb{E}G\left[2\sup_{f\in\mathcal{F}}(-\mathfrak{R}_n(f))\right] && (4.6)\\
&= \mathbb{E}G\left[2\sup_{f\in\mathcal{F}}\mathfrak{R}_n(f)\right], && (4.7)
\end{aligned}$$

where (4.5) follows from the monotonicity of G, where (4.6) follows from the convexity of G and where finally (4.7) derives from the fact that, for all $f \in \mathcal{F}$, both the variables $\mathfrak{R}'_n(f)$ and $-\mathfrak{R}_n(f)$ have the same distribution as $\mathfrak{R}_n(f)$. This concludes the proof of the first statement.

□

5 Contraction

Theorem 5.1. *Let $T \subset \mathbb{R}^n$ and $L > 0$ be fixed. For $i = 1, \dots, n$, let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz functions satisfying $\varphi_i(0) = 0$. Let $G : \mathbb{R} \to \mathbb{R}$ be convex and non-decreasing. Then, for any function $F : \mathbb{R}^n \to \mathbb{R}$,*

$$\mathbb{E}\, G\left[\sup_{t\in T}\{\, F(t) + \frac{1}{n}\sum_{i=1}^{n}\sigma_i\varphi_i(t_i)\,\}\right] \le \mathbb{E}\, G\left[\sup_{t\in T}\{\, F(t) + \frac{L}{n}\sum_{i=1}^{n}\sigma_i t_i\,\}\right],$$

where t_i denotes the i-th coordinate of $t \in \mathbb{R}^n$.

The proof may be found, for instance, in [5] (Theorem 4.12) or in [4] (Theorem 2.2). A typical application of the previous result is the following.

Corollary 5.2. *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz and such that $\varphi(0) = 0$. Then, for any collection $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$, we have*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathfrak{R}_n(\varphi \circ f)\right] \le L\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathfrak{R}_n(f)\right].$$

Proof. Conditioning on the data $\{Z_i\}_{i=1}^n$ and applying the previous Theorem with $F(t) = 0$ and $G(t) = t$ implies

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathfrak{R}_n(\varphi \circ f)|\{Z_i\}_{i=1}^n\right] \le L\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathfrak{R}_n(f)|\{Z_i\}_{i=1}^n\right].$$

Taking the expectation on both sides gives the desired result. □

6 Chaining: General result

A classical result states that if $X_1, \dots, X_n$ are all sub-gaussian with variance proxy at most σ^2, we have

$$\mathbb{E}\max_{1\le i\le n} X_i \le \sqrt{2\sigma^2 \log n}. \tag{6.1}$$

The chaining method can be understood as an important tool to generalize the above classical result to the case of a more general sub-gaussian process $(X_t)_{t\in T}$. The main insight of the chaining method is that, if $(X_t)_{t\in T}$ is a sub-gaussian process as defined below, the quantity

$$\mathbb{E}\sup_{t\in T} X_t,$$

can be controlled in terms of a complexity measure of the index set T.

We first recall the definition of covering numbers. Let (E, d) be a pseudo-metric[1] space and $T \subset E$. For any $\varepsilon > 0$, define an ε-net for T as any collection $\mathcal{N}$ of points in E such that

$$T \subset \bigcup_{t \in \mathcal{N}} B(t, \varepsilon), \tag{6.2}$$

where $B(t, \varepsilon) := \{s \in E : d(s,t) \leq \varepsilon\}$. Finally, the ε-covering number of T in E is defined

$$N(T, d, \varepsilon) := \inf\{|\mathcal{N}| : \mathcal{N} \text{ is an } \varepsilon\text{-net for } T \text{ in } E\}.$$

Definition 6.1. *Let (E, d) be a pseudo-metric space. A centered stochastic process $(X_t)_{t \in E}$ be is said to be sub-gaussian if*

$$\forall s, t \in E, \forall \lambda \in \mathbb{R} : \log \mathbb{E}\left[e^{\lambda(X_s - X_t)}\right] \leq \frac{\lambda^2 d^2(s,t)}{2},$$

i.e., if $X_s - X_t$ is sub-gaussian with variance proxy at most $d^2(s,t)$.

Theorem 6.2. *Let (E, d) be a pseudo-metric space and $(X_t)_{t \in E}$ be a centered sub-gaussian stochastic process. Let $T \subset E$ and denote $D(T)$ its diameter. Then, for all $T \subset E$ and all $0 \leq \varepsilon \leq D(T)/2$,*

$$\mathbb{E} \sup_{t \in T} X_t \leq \mathbb{E}\left[\sup_{d(s,t) \leq 4\epsilon} |X_s - X_t|\right] + 12 \int_{\varepsilon}^{\frac{D(T)}{2}} \sqrt{\log N(T, d, u)}\, \mathrm{d}u.$$

Proof. Fix $t_0 \in T$. For all integers $j \geq 0$, denote $\delta_j = 2^{-j} D(T)$. For $j \geq 0$, let $\mathcal{N}_j$ be a δ_j-net of T in E of minimum cardinality, with the convention that $\mathcal{N}_0 = \{t_0\}$. For all $j \geq 0$, let $\pi_j : T \to \mathcal{N}_j$ be any function such that, for all $t \in \mathcal{T}$,

$$d(t, \pi_j(t)) \leq \delta_j. \tag{6.3}$$

Finally, denote

$$\Delta(u) := \sup_{d(s,t) \leq u} |X_s - X_t|.$$

Then, for any integer $J \geq 1$ and any $t \in T$, write

$$\begin{aligned} X_t - X_{t_0} &= X_t - X_{\pi_J(t)} + \sum_{j=1}^{J} (X_{\pi_j(t)} - X_{\pi_{j-1}(t)}) \\ &\leq \Delta(\delta_J) + \sum_{j=1}^{J} \sup_{t \in T} (X_{\pi_j(t)} - X_{\pi_{j-1}(t)}). \end{aligned}$$

[1] All the properties of a metric space are satisfied except that different points may be at zero distance from each other.

Taking the expectation on both sides of the last inequality, and recalling that X_{t_0} is centered, we obtain

$$\mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\Delta(\delta_J) + \sum_{j=1}^{J} \mathbb{E}\sup_t \left(X_{\pi_j(t)} - X_{\pi_{j-1}(t)}\right). \tag{6.4}$$

By definition of the $\mathcal{N}_j$'s, we have, for all $j \geq 1$,

$$\begin{aligned} |\{(\pi_j(t), \pi_{j-1}(t)) : t \in T\}| &\leq N(T, d, \delta_j) N(T, d, \delta_{j-1}) \\ &\leq N(T, d, \delta_j)^2, \end{aligned}$$

since the $\delta \mapsto N(T, d, \delta)$ is non-increasing. Now observe that, for all $t \in T$, we have $d(\pi_j(t), \pi_{j-1}(t)) \leq 3\delta_j$ by the triangle inequality. Therefore, the sub-gaussian assumption, together with the basic inequality (6.1), imply that, for all $j \geq 1$,

$$\begin{aligned} \mathbb{E}\sup_{t\in T} \left(X_{\pi_j(t)} - X_{\pi_{j-1}(t)}\right) &\leq 6\delta_j \sqrt{\log N(T, d, \delta_j)} \\ &= 12(\delta_j - \delta_{j+1}) \sqrt{\log N(T, d, \delta_j)}. \end{aligned} \tag{6.5}$$

Hence, combining (6.4) and (6.5) leads to

$$\begin{aligned} \mathbb{E}\sup_{t\in T} X_t &\leq \mathbb{E}\Delta(\delta_J) + 12\sum_{j=1}^{J} (\delta_j - \delta_{j+1}) \sqrt{\log N(T, d, \delta_j)} \\ &\leq \mathbb{E}\Delta(\delta_J) + 12\sum_{j=1}^{J} \int_{\delta_{j+1}}^{\delta_j} \sqrt{\log N(T, d, u)}\, \mathrm{d}u \\ &= \mathbb{E}\Delta(\delta_J) + 12 \int_{\delta_{J+1}}^{\frac{D(T)}{2}} \sqrt{\log N(T, d, u)}\, \mathrm{d}u. \end{aligned} \tag{6.6}$$

Now, take $0 < \varepsilon < D(T)/4$ and let J be the largest integer such that $\varepsilon \leq \delta_{J+1}$. Then, since $\delta_{J+2} < \varepsilon$, we have $\delta_J \leq 4\varepsilon$. As a result, applying inequality (6.6) for this value of J implies that

$$\mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\Delta(4\epsilon) + 12 \int_{\epsilon}^{\frac{D(T)}{2}} \sqrt{\log N(T, d, u)}\, \mathrm{d}u,$$

which is the desired result. Note finally that the bound of the Theorem is trivially true if $D(T)/4 \leq \varepsilon \leq D(T)/2$. $\square$

7 Chaining for Rademacher processes

Let (E, d) be the pseudo metric space $L^2(P_n)$, i.e., the set of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ such that $P_n f^2 < +\infty$, endowed with the pseudo-metric

$$d^2(f, g) := P_n(f - g)^2 = \frac{1}{n} \sum_{i=1}^{n} (f(Z_i) - g(Z_i))^2.$$

Lemma 7.1. *Conditionally on the data $\{Z_i\}_{i=1}^n$, the rescaled Rademacher process*

$$(\sqrt{n}\, \mathfrak{R}_n(f))_{f \in L^2(P_n)},$$

is (centered and) sub-gaussian. That is, for all $f, g \in L^2(P_n)$, the variable $\sqrt{n}\, (\mathfrak{R}_n(f) - \mathfrak{R}_n(g))$ is sub-gaussian with variance proxy at most $P_n(f - g)^2$.

Proof. Exercise. □

Corollary 7.2. *For any collection $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$, we have for any $0 < \varepsilon \le D_n(\mathcal{F})/2$,*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \mathfrak{R}_n(f) | \{Z_i\}_{i=1}^n\right] \le 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{\frac{D_n(\mathcal{F})}{2}} \sqrt{\log N(\mathcal{F}, \|. - .\|_{L^2(P_n)}, u)}\, \mathrm{d}u,$$

where $D_n(\mathcal{F})$ denotes the diameter of $\mathcal{F}$ in $L^2(P_n)$.

Proof. Exercise. □

References

[1] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[2] O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus de l'Académie des Sciences de Paris*, 334:495–500, 2002.

[3] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12:929–989, 1984.

[4] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. In *Lectures from the 38th Summer School on Probability Theory held in Saint-Flour in July, 2008*. Springer, 2011.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[6] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.