Topics in High-Dimensional Statistics

Lecture 5: Empirical Risk Minimization III
Fast rates

Contents

1 Excess risk and increments of empirical processes

The general analysis provided so far proves that, under fairly general assumptions, the excess risk converges to 0 (with high probability or in expectation) at the rate of $1/\sqrt{n}$. A finer analysis, due to the works of people like Massart [3] or Koltchinskii [2] for instance, show that under a specific assumption on the cost function γ, known as the Bernstein assumption, the excess risk converges to 0 (with high probability or in expectation) at the rate of $1/n^\alpha$ for $1/2 \leq \alpha \leq 1$. The central idea that drives this analysis is that the excess risk can be connected to the fixed points of the increments of the empirical process as described next.

First we recall some notation. 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 parameter space Θ and a cost (or contrast function) $\gamma : \Theta \times \mathcal{Z} \to \mathbb{R}$, the risk and empirical risk are defined respectively, for $\theta \in \Theta$, by

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

where

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

denotes the empirical measure. We denote as usual

$$\bar{\theta} \in \arg\min_{\theta\in\Theta} R(\theta) \quad \text{and} \quad \theta_n \in \arg\min_{\theta\in\Theta} R_n(\theta),$$

and we are interested in upper bounding the excess risk of the empirical risk minimizer θ_n which, for the sake of simplicity, we will denote δ_n in this lecture, i.e.,

$$\delta_n := R(\theta_n) - R(\bar{\theta}).$$

The analysis provided so far was based on the basic observation that

$$\delta_n \leq \sup_{f\in\mathcal{F}} (P - P_n)f, \tag{1.1}$$

where

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

A close look at the proof of this inequality reveals that it is quite brutal and can be improved.

To that aim, we introduce the **increment of the empirical process**

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

where, for all $\delta > 0$,

$$\mathcal{F}(\delta) := \{\gamma(\theta, .) - \gamma(\bar{\theta}, .) : P(\gamma(\theta, .) - \gamma(\bar{\theta}, .)) \leq \delta\}.$$

We are now in position to make the central observation that allows a finer analysis of the excess risk δ_n.

Lemma 1.1. *For all $n \geq 1$, we have*

$$\delta_n \leq \nu_n(\delta_n).$$

Proof. We simply modify the proof of (1.1) by writing

$$\begin{aligned}
\delta_n &= P(\gamma(\theta_n) - \gamma(\bar{\theta}, .)) \\
&\leq (P - P_n)(\gamma(\theta_n) - \gamma(\bar{\theta}, .)) \\
&\leq \sup\{(P - P_n)(\gamma(\theta) - \gamma(\bar{\theta}, .)) : P(\gamma(\theta) - \gamma(\bar{\theta}, .)) \leq \delta_n\} \\
&= \nu_n(\delta_n),
\end{aligned}$$

where, in the second line, we have used that $P_n(\gamma(\theta_n) - \gamma(\bar{\theta}, .)) \leq 0$ by definition of the empirical risk minimizer. □

To interpret this observation, a few comments are in order. First, note that we have by construction that $\nu_n(0) = 0$. Second, the lemma above indicates that for all $n \geq 1$ there exists some $\delta \geq 0$ for which the graph of the map $\delta \mapsto \nu_n(\delta)$ is above (or meets) that of the identity $\delta \mapsto \delta$. Lastly, and provided

$$\delta_\infty := \sup_{\theta \in \Theta} P(\gamma(\theta, .) - \gamma(\bar{\theta}, .)) < +\infty,$$

the map $\delta \mapsto \nu_n(\delta)$ is constant for $\delta \geq \delta_\infty$. Hence, provided we indeed have $\delta_\infty < +\infty$, the function ν_n admits a largest fixed point δ_n^* defined by

$$\delta_n^* := \sup\{\delta \geq 0 : \delta = \nu_n(\delta)\}.$$

The previous lemma indicates that $\delta_n \leq \delta_n^*$. The next result allows to build upon this observation to bound δ_n with high probability.

Lemma 1.2. *Let $\nu(\delta)$, $\delta \geq 0$, be non-negative random variables such that, almost surely, $\nu(\delta) \leq \nu(\delta')$ if $\delta \leq \delta'$. Let $\beta(\delta, t)$, $\delta \geq 0$, $t \geq 0$, be real numbers such that $\beta(\delta, t) \leq \beta(\delta, t')$, as soon as $t \leq t'$, and such that*

$$\mathbb{P}(\nu(\delta) \geq \beta(\delta, t)) \leq e^{-t}.$$

Finally, let $\hat{\delta}$ be a nonnegative random variable, a priori upper bounded by a constant $\bar{\delta} > 0$, and such that

$$\hat{\delta} \leq \nu(\hat{\delta}).$$

Then, defining, for all $t \geq 0$*,*

$$\beta(t) = \inf\left\{\alpha > 0 : \sup_{\delta \geq \alpha} \frac{\beta\left(\delta, \frac{t\delta}{\alpha}\right)}{\delta} \leq 1\right\},$$

we obtain, for all $t \geq 0$*,*

$$\mathbb{P}(\hat{\delta} \geq \beta(t)) \leq 2e^{-t}.$$

Proof. The bound is obvious if $t \leq \log 2$ since, in this case, the upper bound is larger than 1. Therefore, we will assume $t > \log 2$. The proof will be divided into two steps.

• *Step 1.* Let $\delta_j, j \geq 0$ be a strictly decreasing sequence of positive numbers with $\delta_0 = \bar{\delta}$ and let $t_j, j \geq 0$, be an arbitrary sequence of positive numbers. For all $\delta \geq 0$, denote

$$\bar{\beta}(\delta) = \sum_{j=0}^{+\infty} \beta(\delta_j, t_j)\mathbf{1}\{\delta_{j+1} < \delta \leq \delta_j\},$$

and set

$$\delta^\star = \sup\{\delta \geq 0 : \delta \leq \bar{\beta}(\delta)\}.$$

The goal of this first step is to prove that

$$\forall \delta \geq \delta^\star : \quad \mathbb{P}(\hat{\delta} \geq \delta) \leq \sum_{\delta_j \geq \delta} e^{-t_j}.$$

Fix $\delta > \delta^\star$. For all $j \geq 0$, define the events $E_j = \{\nu(\delta_j) < \bar{\beta}(\delta_j)\}$ and set

$$E = \bigcap_{\delta_j \geq \delta} E_j.$$

Then, it may be easily verified that

$$\mathbb{P}(E) \geq 1 - \sum_{\delta_j \geq \delta} e^{-t_j}.$$

On the event E, and for all $\delta' \geq \delta$, we have $\nu(\delta') \leq \bar{\beta}(\delta')$, almost surely, by monotonicity of ν and by definition of $\bar{\beta}$. Thus, on the event $\{\hat{\delta} \geq \delta\} \cap E$ we obtain

$$\hat{\delta} \leq \nu(\hat{\delta}) \leq \bar{\beta}(\hat{\delta}),$$

which implies that $\delta \leq \hat{\delta} \leq \delta^\star$. Since this contradicts $\delta > \delta^\star$, we deduce that $\{\hat{\delta} \geq \delta\} \subset E^c$ which implies that

$$\mathbb{P}(\hat{\delta} \geq \delta) \leq \sum_{\delta_j \geq \delta} e^{-t_j}.$$

By continuity, this also holds for $\delta = \delta^\star$.

• Step 2. Fix $\alpha > \beta(t)$. Then, for all $\delta \geq 0$, let

$$\bar{\beta}_\alpha(\delta, t) = \sum_{j=0}^{+\infty} \beta\left(\frac{\bar{\delta}}{2^j}, \frac{t\bar{\delta}}{\alpha 2^j}\right) \mathbf{1}\left\{\frac{\bar{\delta}}{2^{j+1}} < \delta \leq \frac{\bar{\delta}}{2^j}\right\}.$$

Observe that

$$\begin{aligned}
\frac{\bar{\beta}_\alpha(\alpha, t)}{\alpha} &= \sum_{j=0}^{+\infty} \frac{\bar{\delta}}{\alpha 2^j} \left(\frac{\bar{\delta}}{2^j}\right)^{-1} \beta\left(\frac{\bar{\delta}}{2^j}, \frac{t\bar{\delta}}{\alpha 2^j}\right) \mathbf{1}\left\{\frac{\bar{\delta}}{2^{j+1}} < \alpha \leq \frac{\bar{\delta}}{2^j}\right\} \\
&\leq \sum_{j=0}^{+\infty} \left(\frac{\bar{\delta}}{2^j}\right)^{-1} \beta\left(\frac{\bar{\delta}}{2^j}, \frac{t\bar{\delta}}{\alpha 2^j}\right) \mathbf{1}\left\{\frac{\bar{\delta}}{2^{j+1}} < \alpha \leq \frac{\bar{\delta}}{2^j}\right\} \\
&\leq \sup_{\delta \geq \alpha} \frac{\beta\left(\delta, \frac{t\delta}{\alpha}\right)}{\delta} \\
&\leq 1, \qquad (1.2)
\end{aligned}$$

where (1.2) follows from the monotonicity of $\beta(\delta, .)$ and the fact that $\alpha > \beta(t)$. In addition, (1.2) implies that

$$\alpha \geq \delta_t^\star = \sup\left\{\delta \geq 0 : \delta \leq \bar{\beta}_\alpha(\delta, t)\right\}.$$

Then, according to the first step above,

$$\mathbb{P}(\hat{\delta} \geq \alpha) \leq \sum_{\frac{\bar{\delta}}{2^j} \geq \alpha} e^{-\frac{t\bar{\delta}}{\alpha 2^j}}. \qquad (1.3)$$

The sum on the right hand side of (1.3) may be bounded as follows. Let

$$j^\star = \max\left\{j \geq 0 : \frac{\bar{\delta}}{2^j} \geq \alpha\right\}.$$

Then

$$\sum_{\frac{\bar{\delta}}{2^j} \geq \alpha} e^{-\frac{t\bar{\delta}}{\alpha 2^j}} = \sum_{j=0}^{j^\star} e^{-\frac{t\bar{\delta}}{\alpha 2^j}} \leq \sum_{j=0}^{j^\star} e^{-t2^{(j^\star - j)}} \leq \sum_{j=0}^{+\infty} e^{-t2^j},$$

and

$$\sum_{j=0}^{+\infty} e^{-t2^j} \leq e^{-t} + \sum_{j=1}^{+\infty} (2^j - 2^{j-1}) e^{-t2^j} \leq e^{-t} + \int_1^{+\infty} e^{-tu} \, \mathrm{d}u = 2e^{-t}.$$

Finally, we have proved that, for all $t \geq 0$ and for all $\alpha > \beta(t)$ we have

$$\mathbb{P}(\hat{\delta} \geq \alpha) \leq 2e^{-t}.$$

The result follows by continuity.

□

2 Bounding the increments of the empirical process

The previous section allows one to focus on finding explicit numbers $\beta_n(\delta, t)$ such that, for all $\delta, t > 0$, we have

$$\mathbb{P}(\nu_n(\delta) \geq \beta_n(\delta, t)) \leq e^{-t},$$

where

$$\nu_n(\delta) = \sup_{f \in \mathcal{F}(\delta)} (P - P_n) f.$$

The following result provides such explicit numbers based on Bousquet's concentration inequality [1].

Lemma 2.1. *Suppose that $\gamma(\theta, .)$ takes values in $[0, b]$ for all $\theta \in \Theta$. Then, for all $\delta > 0$ and $t > 0$, letting*

$$\sigma^2(\delta) := \sup_{f \in \mathcal{F}(\delta)} Pf^2 \quad \text{and} \quad \beta_n(\delta, t) := 2\mathbb{E}\nu_n(\delta) + \sigma(\delta)\sqrt{\frac{2t}{n}} + \frac{8bt}{3n},$$

we have

$$\mathbb{P}(\nu_n(\delta) \geq \beta_n(\delta, t)) \leq e^{-t}.$$

Proof. According to Bousquet's inequality, we know that

$$\nu_n(\delta) \geq \mathbb{E}\nu_n(\delta) + \sqrt{\frac{2t}{n}\left(\sigma^2(\delta) + 4b\,\mathbb{E}\nu_n(\delta)\right)} + \frac{2bt}{3n}$$

with probability at most e^{-t}. Now, using the fact that $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ and $2\sqrt{uv} \leq u + v$, it follows that

$$\mathbb{E}\nu_n(\delta) + \sqrt{\frac{2t}{n}\left(\sigma^2(\delta) + 4b\,\mathbb{E}\nu_n(\delta)\right)} + \frac{2bt}{3n} \leq 2\mathbb{E}\nu_n(\delta) + \sigma(\delta)\sqrt{\frac{2t}{n}} + \frac{8bt}{3n},$$

which completes the proof.

□

We are now brought to consider the two quantities

$$\mathbb{E}\nu_n(\delta) \quad \text{and} \quad \sigma^2(\delta),$$

independently, with the goal of tracking precisely the dependence on δ. We separate the analysis in the two following sections.

3 Bounding $\sigma^2(\delta)$: Bernstein assumption

This section introduces a general condition, satisfied in many interesting cases, under which the control of $\sigma(\delta)$ and $\mathbb{E}\nu_n(\delta)$, for all $\delta > 0$, can be done in a fruitful way.

Definition 3.1 (Bernstein condition)**.** *For $B > 0$ and $\beta \in (0, 1]$, the excess loss class*

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

satisfies the (P, B, β)-Bernstein condition if

$$\forall f \in \mathcal{F}, \quad Pf^2 \leq B(Pf)^\beta.$$

Lemma 3.2. *Suppose that the excess loss class $\mathcal{F}$ satisfies the (P, B, β)-Bernstein condition. Then,*

$$\sigma^2(\delta) \leq B\,\delta^\beta,$$

for all $\delta > 0$.

Proof. The proof follows from the observation that $\mathcal{F}(\delta) = \{f \in \mathcal{F} \,:\, Pf \leq \delta\}$ so that, using the Bernstein condition,

$$\sigma^2(\delta) = \sup\{\, Pf^2 \,:\, Pf \leq \delta \,\} \leq \sup\{\, B\,(Pf)^\beta \,:\, Pf \leq \delta \,\} \leq B\,\delta^\beta,$$

which is the desired result. □

Let us discuss one practical scenario of importance for which the Bernstein assumption is satisfied. Consider the supervised learning problem introduced in previous lectures. In this setting we have typically $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$, and the parameter set Θ is a collection of functions $\theta : \mathcal{X} \to \mathcal{Y}$. The generic random variable Z takes here the form of a random pair (X, Y) of distribution P where X (resp. Y) is understood as an input (resp. output) variable. Here, the cost function γ takes the form

$$\gamma(\theta, (x, y)) = \ell(y, \theta(x)),$$

for a prescribed loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Here, the Bernstein assumption is satisfied provided a few classical assumptions on the loss function ℓ are satisfied. Namely, suppose that the following conditions hold:

(1) There exists $L > 0$ such that, for all

$$y, u, u' \in \mathcal{Y} : \quad |\ell(y, u) - \ell(y, u')| \leq L|u - u'|.$$

(2) The set $\mathcal{Y} \subset \mathbb{R}$ is convex, the set Θ is a convex set of functions $\theta : \mathcal{X} \to \mathcal{Y}$ and there exists $\alpha > 0$ such that, for all $y \in \mathcal{Y}$, the function $\ell(y, .)$ is α-strongly convex[1].

Then, the excess loss class $\mathcal{F}$ satisfies the $(P, 4L^2/\alpha, 1)$-Bernstein condition. We leave the proof to the reader.

4 Bounding $\mathbb{E}\nu_n(\delta)$

We will give two results providing explicit bounds for $\mathbb{E}\nu_n(\delta)$ depending on the complexity of the class $\mathcal{F}$.

Theorem 4.1. *Suppose that all $f \in \mathcal{F}$ are $[-b, b]$-valued and that $\mathcal{F}$ satisfies the (P, B, β)-Bernstein condition.*

(1) *Suppose that there exists $c, d > 0$ such that, for all $0 < u \leq 1$,*

$$a.s., \quad N(\mathcal{F}, \|. - .\|_{L^2(P_n)}, ub) \leq \left(\frac{c}{u}\right)^d. \tag{4.1}$$

Then, for all $n \geq 1$ and all $\delta > 0$,

$$\mathbb{E}\nu_n(\delta) \leq c_1 \max\left\{\delta^{\frac{\beta}{2}}\sqrt{\frac{d}{n}\log\left(\frac{c_2}{\delta}\right)}, \frac{d}{n}\log\left(\frac{c_2}{\delta}\right)\right\},$$

for constants $c_1, c_2 > 0$ depending only on b, β, B and c.

(2) *Suppose that there exists $c > 0$, $\alpha \in (0, 2)$ such that, for all $0 < u \leq 1$,*

$$a.s., \quad \log N(\mathcal{F}, \|. - .\|_{L^2(P_n)}, ub) \leq \left(\frac{c}{u}\right)^\alpha. \tag{4.2}$$

Then, for all $n \geq 1$ and all $\delta > 0$,

$$\mathbb{E}\nu_n(\delta) \leq \max\left\{c_3\frac{\delta^{\frac{\beta(2-\alpha)}{4}}}{\sqrt{n}}, c_4\frac{\delta^{-\frac{\beta\alpha}{2}}}{n}\right\},$$

for constants $c_3, c_4 > 0$ depending only on b, β, B, c and α.

Proof. See [2]. □

[1] Recall that this means that the map $u \mapsto \ell(y, u) - \frac{\alpha}{2}u^2$ is convex

5 Fast rates

Combining all the above, one proves the following.

Theorem 5.1. *Suppose that all $f \in \mathcal{F}$ are $[-b,b]$-valued and that $\mathcal{F}$ satisfies the (P,B,β)-Bernstein condition.*

(1) *Suppose that there exists $c,d>0$ such that, for all $0<u\leq 1$,*

$$a.s., \quad N(\mathcal{F}, \|.-.\|_{L^2(P_n)}, ub) \leq \left(\frac{c}{u}\right)^d. \tag{5.1}$$

Then, there exists $C>0$ depending only on b,β,B and c such that for all $n\geq 1$ and all $t>0$,

$$R(\theta_n) - \inf_{\theta\in\Theta} R(\theta) \leq \max\left\{\frac{d\log n}{n}, \frac{t}{n}\right\}^{\frac{1}{2-\beta}},$$

with probability at least $1-2e^{-t}$.

(2) *Suppose that there exists $c>0$, $\alpha\in(0,2)$ such that, for all $0<u\leq 1$,*

$$a.s., \quad \log N(\mathcal{F}, \|.-.\|_{L^2(P_n)}, ub) \leq \left(\frac{c}{u}\right)^\alpha. \tag{5.2}$$

Then, there exists $C'>0$ depending only on b,β,B,c and α such that for all $n\geq 1$ and all $t>0$,

$$R(\theta_n) - \inf_{\theta\in\Theta} R(\theta) \leq C' \max\left\{\left(\frac{1}{n}\right)^{\frac{2}{2(2-\beta)+\beta\alpha}}, \left(\frac{t}{n}\right)^{\frac{1}{2-\beta}}\right\},$$

with probability at least $1-2e^{-t}$.

Proof. See [2] □

References

[1] 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.

[2] 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.

[3] P. Massart. Concentration inequalities and model selection. In *Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour in July, 2003*. Springer, 2007.