Topics in learning theory*

Lecture 5: Empirical risk minimization (I)

Contents

1 Empirical risk minimization **2**

2 Estimation-Approximation tradeoff **2**

3 Risk bounds for finite classes **3**

3.1 A general result 3

3.2 Faster rates for strongly convex losses 4

In this chapter, we come back to the problem of statistical learning introduced in Lecture 1 and explore a basic principle known as empirical risk minimization (ERM). Recall that, in the statistical learning setup, one is given:

- a decision set Θ,
- an outcome set $\mathcal{Z}$,
- a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$,
- and finally a learning sample
$$\{Z_i\}_{i=1}^n,$$
composed of i.i.d. $\mathcal{Z}$-valued random variables with same distribution as (and independent from) a generic random variable Z.

In this setting, the goal is to construct a data-driven decision $\hat{\theta}_n$ that minimizes the excess risk
$$\mathcal{E}(\hat{\theta}_n) := R(\hat{\theta}_n) - R^*,$$
with high probability or in expectation, where
$$R(\hat{\theta}_n) = \mathbb{E}[\ell(\hat{\theta}_n, Z) | \{Z_i\}_{i=1}^n],$$
and
$$R^* = \inf_{\theta \in \Theta} \mathbb{E}[\ell(\theta, Z)].$$

*Teaching material can be found at https://www.qparis-math.com/teaching.

1 Empirical risk minimization

Empirical risk minimization is the natural statistical procedure consisting in minimizing an approximation of the risk constructed from data. Precisely, empirical risk minimization consists in chosing

$$\hat{\theta}_n \in \operatorname*{arg\,min}_{\theta \in \mathcal{M}} R_n(\theta), \tag{1.1}$$

where

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

is known as the empirical risk of θ and

$$\mathcal{M} \subset \Theta,$$

is called hypothesis class or model. The role of $\mathcal{M}$ is fundamental in this framework and its choice should leverage the statistician's knowledge of the problem at hand, *i.e.*, the available information on the unknown distribution of the data.

2 Estimation-Approximation tradeoff

The role of model $\mathcal{M}$ is made clear by observing that the excess risk of an empirical risk minimizer $\hat{\theta}_n$ decomposes as

$$\begin{aligned} \mathcal{E}(\hat{\theta}_n) &= R(\hat{\theta}_n) - R^\star \\ &= (R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta)) + \inf_{\theta \in \mathcal{M}} (R(\theta) - R^\star). \end{aligned} \tag{2.1}$$

In this decomposition, known as the estimation-approximation decomposition, the two terms on the right hand-side of (2.1) show opposite behaviors in terms of the model $\mathcal{M}$.

The first term,

$$R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta),$$

is random, referred to as the estimation error, and quantifies the performance of $\hat{\theta}_n$ compared to the best possible (deterministic) predictor in $\mathcal{M}$. Roughly speaking, the estimation error tends to get larger (and the optimization problem (1.1) more difficult to solve) as the complexity of $\mathcal{M}$ increases. Hence, from this point of view, one should favor a simple or small model $\mathcal{M}$.

The second term,

$$\inf_{\theta \in \mathcal{M}} (R(\theta) - R^\star)$$

is deterministic, non-negative and referred to as the approximation error. Note that, while $\mathcal{M} \subset \Theta$ may be much smaller than Θ, it may be that there exists $\theta \in \mathcal{M}$ such that

$$R(\theta) = R^\star,$$

in which case the approximation error is 0. More generally, the approximation error accounts for the approximation properties of $\mathcal{M}$ relative to the set of elements $\theta \in \Theta$ solving $R(\theta) = R^\star$. Contrary to the estimation error, this term tends to get smaller as the complexity or size of $\mathcal{M}$ gets larger. Specifying a small $\mathcal{M}$ for which the approximation error is small is a problem relative to both approximation theory and the statistician's expertise.

3 Risk bounds for finite classes

In this section, we focus on bounding the estimation error

$$R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta),$$

in the simple setting where the model $\mathcal{M}$ is composed of a finite number of elements.

3.1 A general result

We start with a technical lemma.

Lemma 3.1. *Suppose that X, Y are two sub-gaussian random variables (not necessarily independent) with respective variance proxys σ_X^2 and σ_Y^2. Then, $X - Y$ is sub-gaussian with variance proxy at most $2\sigma_X^2 + 2\sigma_Y^2$.*

Proof. Exercise. □

Theorem 3.2. *Suppose that, for all $\theta \in \mathcal{M}$, the random variable $\ell(\theta, Z)$ is sub-gaussian with variance proxy at most σ^2. Then, for all $n \geq 1$ and all $\delta \in (0,1)$,*

$$R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta) \leq \sqrt{\frac{8\sigma^2}{n} \ln\left(\frac{|\mathcal{M}|}{\delta}\right)},$$

with probability at least $1 - \delta$.

Proof. We divide the proof in three steps.

Step 1. In this first step, we show how to bound the estimation error by the uniform deviation between the risk and the empirical risk on the class $\mathcal{M}$. Introduce

$$\bar{\theta} \in \arg\min_{\theta \in \mathcal{M}} R(\theta),$$

and denote,

$$\bar{R}(\theta) := R(\theta) - R(\bar{\theta}) \quad \text{and} \quad \bar{R}_n(\theta) := R_n(\theta) - R_n(\bar{\theta}).$$

Now observe that since,

$$\bar{R}_n(\hat{\theta}_n) \leq 0,$$

we get

$$\begin{aligned} R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta) &= \bar{R}(\hat{\theta}_n) \\ &\leq \bar{R}(\hat{\theta}_n) - \bar{R}_n(\hat{\theta}_n). \end{aligned}$$

In particular, we deduce that

$$R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta) \leq \max_{\theta \in \mathcal{M}} (\bar{R}(\theta) - \bar{R}_n(\theta)).$$

Step 2. Now we combine the first step, and the union bound, to deduce that, for all $t > 0$,

$$\begin{aligned} \mathbb{P}(R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta) > t) &\leq \mathbb{P}(\max_{\theta \in \mathcal{M}} (\bar{R}(\theta) - \bar{R}_n(\theta)) > t) \\ &= |\mathcal{M}| \max_{\theta \in \Theta} \mathbb{P}(\bar{R}(\theta) - \bar{R}_n(\theta) > t). \end{aligned}$$

Step 3. Observe that, for all $\theta \in \mathcal{M}$, we have

$$\bar{R}(\theta) = \mathbb{E}[\ell(\theta, Z) - \mathbb{E}\ell(\bar{\theta}, Z)],$$

and

$$\bar{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\ell(\theta, Z_i) - \ell(\bar{\theta}, Z_i)).$$

The variables

$$\ell(\theta, Z_i) - \ell(\bar{\theta}, Z_i), 1 \leq i \leq n,$$

are independent and, according to Lemma 3.1, they are sub-gaussian with variance proxy at most $4\sigma^2$. Hence, applying Hoeffding's inequality, we conclude that

$$\mathbb{P}(\bar{R}(\theta) - \bar{R}_n(\theta) > t) \leq \exp\left(-\frac{nt^2}{8\sigma^2}\right).$$

Combining this observation with the result of Step 2 we get finally that, for all $t > 0$,

$$\mathbb{P}(R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta) > t) \leq |\mathcal{M}| \exp\left(-\frac{nt^2}{8\sigma^2}\right).$$

Selecting any $\delta \in (0, 1)$, selecting $t > 0$ such that

$$\delta := |\mathcal{M}| \exp\left(-\frac{nt^2}{8\sigma^2}\right),$$

and expressing t in terms of δ, it appears that this statement is equivalent to the desired result. □

Corollary 3.3. *Under the assumptions of the previous result, we have for all $n \geq 1$,*

$$\mathbb{E}[R(\hat{\theta}_n)] - \inf_{\theta \in \mathcal{M}} R(\theta) \leq \sqrt{\frac{8\sigma^2 \ln(e|\mathcal{M}|)}{n}}.$$

Proof. Exercise. □

3.2 Faster rates for strongly convex losses

In this paragraph, we show how the previous result can be greatly improved under additional assumptions on the loss function ℓ. We start by mentioning an auxiliary result.

Theorem 3.4 (Bernstein's inequality)**.** *Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables taking values in a bounded interval $[-b, b]$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1] \geq t\right) \leq \exp\left(-\frac{nt^2}{2\text{var}(X_1) + \frac{2bt}{3}}\right).$$

It is an easy exercise to observe that, under the same assumptions, Bernstein's inequality improves upon the bound of Hoeffding's inequality for all $t \in (0, b]$ (the only relevant range of t's in this context) provided the X_i's have a small variance and more precisely if

$$\text{var}(X_1) \leq \frac{2b^2}{3}.$$

In the sequel, we suppose that Θ is a convex subset of a normed vector space equipped with norm $\|.\|$.

Theorem 3.5. *Suppose that model $\mathcal{M}$ is well specified, i.e., that there exists $\theta^* \in \mathcal{M}$ such that $R(\theta^*) = R^*$. Suppose in addition that there exists $b, L, \alpha > 0$ such that the following assumptions hold:*

(1) *For all $\theta \in \mathcal{M}$,*

$$\mathbb{P}(0 \le \ell(\theta, Z) \le b) = 1,$$

(2) *For all $z \in \mathcal{Z}$, for all $\theta, \theta' \in \Theta$,*

$$|\ell(\theta, z) - \ell(\theta', z)| \le L\|\theta - \theta'\|,$$

(3) *For all $z \in \mathcal{Z}$, the map $\theta \in \Theta \mapsto \ell(\theta, z)$ is α-convex.*

Then, for all $n \ge 1$ and all $\delta \in (0, 1)$,

$$R(\hat{\theta}_n) - \inf_{\theta \in \mathcal{M}} R(\theta) \le \max\left\{\frac{L^2}{\alpha}, \frac{b}{3}\right\} \frac{4}{n} \ln\left(\frac{|\mathcal{M}|}{\delta}\right),$$

with probability at least $1 - \delta$.

Proof. We divide the proof in several steps.

Step 1. As in the proof of Theorem 3.2, denote

$$\bar{R}(\theta) := R(\theta) - R(\theta^*) \quad \text{and} \quad \bar{R}_n(\theta) := R_n(\theta) - R_n(\theta^*).$$

Bernstein's inequality implies that, for all $t > 0$ and all $\theta \in \mathcal{M}$,

$$\mathbb{P}(\bar{R}(\theta) - \bar{R}_n(\theta) > t) \le \exp\left(-\frac{nt^2}{v(\theta) + \frac{2bt}{3}}\right),$$

where

$$v(\theta) := \mathrm{Var}(\ell(\theta, Z) - \ell(\theta^*, Z)).$$

Using assumption (2), we get

$$\begin{aligned} v(\theta) &\le \mathbb{E}[(\ell(\theta, Z) - \ell(\theta^*, Z))^2] \\ &\le L^2\|\theta - \theta^*\|^2. \end{aligned}$$

Assumption (3) implies in addition that the risk function is α-convex which implies, according to Lecture 4, that

$$\frac{\alpha}{2}\|\theta - \theta^*\|^2 \le \bar{R}(\theta).$$

Combining the two previous observations, we deduce that,

$$v(\theta) \le \frac{2L^2}{\alpha}\bar{R}(\theta).$$

As a result, for all $t > 0$,

$$\begin{aligned}\mathbb{P}(\bar{R}(\theta) - \bar{R}_n(\theta) > t) &\le \exp\left(-\frac{nt^2}{\frac{2L^2}{\alpha}\bar{R}(\theta) + \frac{2bt}{3}}\right) \\ &\le \exp\left(-\frac{nt^2}{\max\{\frac{4L^2}{\alpha}\bar{R}(\theta), \frac{4bt}{3}\}}\right) \\ &= \exp\left(-\min\left\{\frac{\alpha nt^2}{4L^2\bar{R}(\theta)}, \frac{3nt}{4b}\right\}\right).\end{aligned}$$

It is an easy exercise to check that the above inequality implies that, for all $\delta \in (0,1)$,

$$\bar{R}(\theta) - \bar{R}_n(\theta) > \max\left\{2L\sqrt{\frac{\bar{R}(\theta)}{\alpha n}\ln\left(\frac{1}{\delta}\right)}, \frac{4b}{3n}\ln\left(\frac{1}{\delta}\right)\right\}, \tag{3.1}$$

with probability at most δ.

Step 2. Lets number the elements of $\mathcal{M}$ as

$$\mathcal{M} = \{\theta_1, \dots, \theta_m\}.$$

Observe that the inequality of the theorem, i.e.,

$$\bar{R}(\hat{\theta}_n) \le \max\left\{\frac{L^2}{\alpha}, \frac{b}{3}\right\}\frac{4}{n}\ln\left(\frac{m}{\delta}\right),$$

is equivalent to

$$\bar{R}(\hat{\theta}_n) \le \max\left\{2L\sqrt{\frac{\bar{R}(\hat{\theta}_n)}{\alpha n}\ln\left(\frac{m}{\delta}\right)}, \frac{4b}{3n}\ln\left(\frac{m}{\delta}\right)\right\}.$$

As a result, using the fact that $\bar{R}_n(\hat{\theta}_n) \le 0$,

$$\begin{aligned}&\mathbb{P}\left(\bar{R}(\hat{\theta}_n) > \max\left\{\frac{L^2}{\alpha}, \frac{b}{3}\right\}\frac{4}{n}\ln\left(\frac{m}{\delta}\right)\right) \\ &= \mathbb{P}\left(\bar{R}(\hat{\theta}_n) > \max\left\{2L\sqrt{\frac{\bar{R}(\hat{\theta}_n)}{\alpha n}\ln\left(\frac{m}{\delta}\right)}, \frac{4b}{3n}\ln\left(\frac{m}{\delta}\right)\right\}\right) \\ &\le \mathbb{P}\left(\bar{R}(\hat{\theta}_n) - \bar{R}_n(\hat{\theta}_n) > \max\left\{2L\sqrt{\frac{\bar{R}(\hat{\theta}_n)}{\alpha n}\ln\left(\frac{m}{\delta}\right)}, \frac{4b}{3n}\ln\left(\frac{m}{\delta}\right)\right\}\right) \\ &= \sum_{j=1}^{m}\mathbb{P}\left(\hat{\theta}_n = \theta_j, \bar{R}(\theta_j) - \bar{R}_n(\theta_j) > \max\left\{2L\sqrt{\frac{\bar{R}(\theta_j)}{\alpha n}\ln\left(\frac{m}{\delta}\right)}, \frac{4b}{3n}\ln\left(\frac{m}{\delta}\right)\right\}\right) \\ &\le m\max_{1\le j\le m}\mathbb{P}\left(\bar{R}(\theta_j) - \bar{R}_n(\theta_j) > \max\left\{2L\sqrt{\frac{\bar{R}(\theta_j)}{\alpha n}\ln\left(\frac{m}{\delta}\right)}, \frac{4b}{3n}\ln\left(\frac{m}{\delta}\right)\right\}\right) \\ &\le \delta,\end{aligned}$$

where the last inequality follows from Step 1. □