Topics in ML, Lecture 3
Statistical Analysis of Empirical Risk Minimization

Lénaïc Chizat*

September 29, 2025

Ref: Chapters 1 and 4 in F. Bach "Learning Theory from First Principles" book.

1 Introduction

The purpose of today's lecture is to develop general tools to bound the *estimation error* of Empirical Risk Minimization (ERM).

- We consider the statistical learning framework of Lecture 1. Let $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i \in \{1, \dots, n\}$ be n iid samples. Given a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we study the performance of ERM over a class of predictors $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}$, given by

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) \quad \text{where} \quad \hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

- Let $f^*_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$ the best predictor in $\mathcal{F}$. The estimation error decomposes as

$$\underbrace{\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f)}_{\text{Estimation error}} = \big(\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\big) + \underbrace{\big(\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f^*_{\mathcal{F}})\big)}_{\leq 0} + \big(\hat{\mathcal{R}}(f^*_{\mathcal{F}}) - \mathcal{R}(f^*_{\mathcal{F}})\big) \tag{1}$$

$$\leq \sup_{f \in \mathcal{F}} \big(\mathcal{R}(f) - \hat{\mathcal{R}}(f)\big) + \sup_{f \in \mathcal{F}} \big(\hat{\mathcal{R}}(f) - \mathcal{R}(f)\big) \tag{2}$$

$$\leq 2 \sup_{f \in \mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \tag{3}$$

  Depending on the technique to bound the right hand side, it is sometimes best to re-start from (2) rather than (3). Such bounds (involving a supremum) are called *uniform concentration bounds*.

- If f was fixed – such as in the third term in Eq. (1) – then $\hat{\mathcal{R}}(f) - \mathcal{R}(f) = \frac{1}{n}\sum_{i=1}^n Z_i - \mathbf{E}[Z]$ with $Z_i = \ell(y_i, f(x_i))$ *independent* and identically distributed. Then by the Central Limit Theorem (CLT):

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbf{E}Z\right) \overset{law}{\to} \mathcal{N}(0, \sigma^2)$$

  where $\sigma^2 = \mathbf{E}[(Z_1 - \mathbf{E}[Z_1])^2]$ is the centered variance of Z_1.

- In this lecture we will study (i) non-asymptotic and (ii) uniform versions of this concentration phenomenon. We will consider general techniques and apply them as illustration to linear models with bounded norm and Lipschitz-losses for illustration.

- We will see both "expectation" and "high-probability" bounds. One can often translate one kind into the other in our setting (see exercises sheets 1 and 3).

*EPFL lenaic.chizat@epfl.ch

2 Non-asymptotic concentration bounds

Let us first investigate the first question of deriving non-asymptotic concentration bounds. A natural first step is to consider Gaussian random variables.

2.1 Concentration for Gaussian random variables

- Question: let $X_i \sim \mathcal{N}(0, \sigma^2)$ be independent. What is the probability that $\frac{1}{n}\sum_i X_i$ deviates from its mean 0 by at least a certain amount?
- It holds for $t \geq 0$,

$$\begin{aligned}\mathbf{P}[X_i \geq t] &= (2\pi\sigma^2)^{-1/2} \int_t^\infty e^{-x^2/(2\sigma^2)} \mathrm{d}x \\ &= (2\pi\sigma^2)^{-1/2} \int_0^\infty e^{-(x+t)^2/(2\sigma^2)} \mathrm{d}x \\ &= (2\pi\sigma^2)^{-1/2} e^{-t^2/(2\sigma^2)} \int_0^\infty e^{-x^2/(2\sigma^2)} e^{-xt/\sigma^2} \mathrm{d}x \\ &\leq \frac{1}{2} e^{-t^2/(2\sigma^2)}\end{aligned}$$

- Since $\frac{1}{n}\sum_i X_i$ is $\mathcal{N}(0, \sigma^2/n)$, it follows

$$\mathbf{P}\Big[\frac{1}{n}\sum_{i=1}^n X_i \geq t\Big] \leq \frac{1}{2} e^{-t^2 n/(2\sigma^2)}. \tag{4}$$

2.2 Sub-Gaussian random variables

In statistical learning, it is convenient to work with random variables (r.v.) that enjoy tail bounds such as (4): such r.v. are called sub-Gaussian. For convenience, we use a standard definition in terms of the moment generating function which easily implies bounds such as (4)[1] (the other implication is true as well, see [Wainwright, 2019, Thm. 2.6]).

Definition 2.1. *A random variable X with mean $\mu = \mathbf{E}[X]$ is sub-Gaussian with variance-proxy $\sigma^2 > 0$ if*

$$\mathbf{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Tail controls on sub-Gaussian random variables can be obtained using a technique called *Chernoff's bound* as follows. Let X be a centered σ^2-sub-Gaussian random variable. By Markov's inequality, for $\lambda \geq 0$,

$$\mathbf{P}[X \geq t] = \mathbf{P}[e^{\lambda X} \geq e^{\lambda t}] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{\lambda t}}.$$

Using the sub-Gaussian property, it follows

$$\begin{aligned}\mathbf{P}[X \geq t] &\leq \inf_{\lambda \geq 0} e^{\sigma^2\lambda^2/2 - \lambda t} \\ &\leq e^{-t^2/(2\sigma^2)}\end{aligned}$$

taking $\lambda = t/\sigma^2$ the minimizer of the quadratic in the exponent. This bound agrees, up to a factor $1/2$, with the bound for the Gaussian (indeed a $\mathcal{N}(0, \sigma^2)$ is σ^2-sub-Gaussian).

A key property of sub-Gaussian random variables is their behavior under averaging which is reminiscent of Eq. (4).

[1] without the 1/2 factor

Proposition 2.2 (Sum of sub-Gaussian random variables)**.** *If $(X_1, \dots, X_n)$ are σ^2-sub-Gaussian and independent, then $S_n = \frac{1}{n}\sum_i X_i$ is (σ^2/n)-sub-Gaussian.*

Proof.

$$\forall \lambda \in \mathbb{R}, \quad \mathbf{E}e^{\lambda(S_n - \mathbf{E}S_n)} = \prod_i \mathbf{E}e^{\lambda(X_i - \mathbf{E}[X_i])/n} \leq \prod_i e^{\lambda^2\sigma^2/(2n^2)} = \mathbf{E}e^{\frac{\lambda^2\sigma^2}{2n}}$$

□

- Combining this result with the tail bound, it follows

$$\mathbf{P}(S_n - \mathbf{E}S_n \geq t) \leq e^{-t^2 n/(2\sigma^2)}$$

  which is the bound we have for Gaussians (up to a factor $1/2$).

- Equivalently we have with probability at least $1 - \delta$ that

$$S_n - \mathbf{E}S_n \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}.$$

2.3 Bounded random variables

In this lecture, the main example of sub-Gaussian random variable is given by almost surely bounded random variables.

Lemma 2.3 (Hoeffding's lemma)**.** *If $X \in [a, b]$ a.s. then X is $(b-a)^2/4$-sub-Gaussian.*

For a guided proof see the exercises.

Theorem 2.4 (Hoeffding's inequality)**.** *Given iid $X_1, \dots, X_n$ with $X_i \in [a, b]$ a.s., then*

$$\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^n (X_i - \mathbf{E}X_i) \geq t\right] \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Proof. It follows directly from Proposition 2.2 and Chernoff's bound. □

Example 2.5. *For classification, and given a* fixed *f, setting $Z_i = 1_{f(X_i) \neq Y_i}$, it holds*

$$\mathbf{P}(\mathcal{R}(f) - \hat{\mathcal{R}}(f) \geq t) = \mathbf{P}\Big(\mathbf{E}Z_1 - \frac{1}{n}\sum Z_i \geq t\Big) \leq e^{-2nt^2}$$

or equivalently: with probability at least $1 - \delta$ it holds

$$\mathcal{R}(f) - \hat{\mathcal{R}}(f) \leq \sqrt{\frac{1}{2n}\log(1/\delta)}.$$

Now that we have tools to tackle non-asymptotic bounds, let us tackle the second challenge which is to derive *uniform* bounds.

3 Analysis via union bounds

3.1 Finite hypothesis class

The first simple case is that of a finite hypothesis class $\mathcal{F} = \{f_1, \dots, f_m\}$ of cardinality $|\mathcal{F}| = m$, and we assume that the loss function is bounded between 0 and ℓ_∞.

By Hoeffding's inequality, for each $f_i \in \mathcal{F}$,

$$\mathbf{P}\big(|\mathcal{R}(f_i) - \hat{\mathcal{R}}(f_i)| \geq t\big) \leq 2\exp\left(-\frac{2nt^2}{\ell_\infty^2}\right).$$

where the factor 2 in front of the exponential comes the fact that here we have stated a two-sided bound. By a union bound, it follows

$$\begin{aligned}\mathbf{P}\Big(\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \geq t\Big) &= \mathbf{P}\Big(\bigcup_{f\in\mathcal{F}} \{|\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \geq t\}\Big) \\ &\leq 2|\mathcal{F}|\exp\left(-\frac{2nt^2}{\ell_\infty^2}\right).\end{aligned}$$

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

$$\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \leq \sqrt{\frac{\ell_\infty^2 \log(2|\mathcal{F}|/\delta)}{2n}}$$

In case one wishes an expectation bound instead of a high-probability bound, then we should use a bound on the expectation of the maximum of sub-Gaussian random variables (see exercises).

3.2 Infinite classes via covering numbers

This argument can easily be extended to infinite classes of functions $\mathcal{F}$ by a discretization argument. Suppose that for $\epsilon > 0$, it is possible to cover $\mathcal{F}$ with $m(\epsilon)$ balls of radius ϵ (say, in sup-norm). The quantity $m(\epsilon)$ is called the covering number of $\mathcal{F}$. Then assuming ℓ is G-Lipschitz in its second argument

$$\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)| \leq G\epsilon + \sup_{f\in\{f_1,\dots,f_{m(\epsilon)}\}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)|.$$

For certain classes of functions, one can upper bound $m(\epsilon)$ and then optimize over ϵ to derive probability or expectation bounds. See [Wainwright, 2019, Sec. 5.1] for more on these techniques and examples of covering number bounds, of see F. Bach's book.

4 Analysis via Rademacher complexity

Let us now present a tool which does not rely on discretization and will lead to the most interesting bounds in this class.

4.1 Rademacher complexity bounds

Let us slightly rephrase our problem: consider $\mathcal{H} = \{h : (x, y) \mapsto \ell(y, f(x)),\ f \in \mathcal{F}\}$ and let $z = (x, y)$. We have

$$\sup_{f\in\mathcal{F}} \hat{\mathcal{R}}(f) - \mathcal{R}(f) = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) - \mathbf{E}\ell(y_i, f(x_i)) \tag{5}$$

$$= \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbf{E}h(z) \tag{6}$$

Definition 4.1 (Rademacher complexity). *Given $\epsilon_1, \dots, \epsilon_n$ independent Rademacher random variables (i.e. uniform on $\{-1, +1\}$), the Rademacher complexity of $\mathcal{H}$ is defined as*

$$\mathrm{Rad}_n(\mathcal{H}) \coloneqq \mathbf{E}_{\epsilon_i, z_i}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(z_i)\Big].$$

This quantity measures the ability of functions in $\mathcal{H}$ to correlate/align with random (Rademacher) labels on a sample of size n. It is a measure of the size/capacity of $\mathcal{H}$. The next result shows that controlling the Rademacher complexity of $\mathcal{H}$ is enough to control the generalization gap.

Lemma 4.2 (Symmetrization lemma).

$$\mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbf{E}h(z)\Big] \le 2\,\mathrm{Rad}_n(\mathcal{H}).$$

Note that it also holds $\mathbf{E}\big[\sup_{h\in\mathcal{H}} \mathbf{E}h(z) - \frac{1}{n}\sum h(z_i)\big] \le 2\,\mathrm{Rad}_n(\mathcal{H})$.

Proof. Let $D = (z_1, \dots, z_n)$ and let $D' = (z'_1, \dots, z'_n)$ another set of independent samples. Since $\mathbf{E}[h(z)] = \mathbf{E}[h(z'_i)|D]$, we have

$$\begin{aligned}
\mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbf{E}h(z)\Big] &= \mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbf{E}[h(z'_i)|D]\Big] \\
&= \mathbf{E}\Big[\sup_{h\in\mathcal{H}} \mathbf{E}\Big[\frac{1}{n}\sum \big(h(z_i) - h(z'_i)\big)|D\Big]\Big]
\end{aligned}$$

Using the fact that the expectation of a sup is larger than the sup of the expectation, it follows

$$\begin{aligned}
\mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbf{E}h(z)\Big] &\le \mathbf{E}\Big[\mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \big(h(z_i) - h(z'_i)\big)|D\Big]\Big] \\
&\le \mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \big(h(z_i) - h(z'_i)\big)\Big]
\end{aligned}$$

Using the fact that z_i and z'_i have the same distribution, it follows

$$\begin{aligned}
\mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbf{E}h(z)\Big] &\le \mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \epsilon_i\big(h(z_i) - h(z'_i)\big)\Big] \\
&\le \mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \epsilon_i h(z_i)\Big] + \mathbf{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \epsilon_i h(z'_i)\Big] \\
&\le 2\,\mathrm{Rad}_n(\mathcal{H}).
\end{aligned}$$

□

Lemma 4.3 (Contraction principle). *If the loss function is G-Lipschitz in its second variable, then*

$$\mathrm{Rad}_n(\mathcal{H}) \le G\,\mathrm{Rad}_n(\mathcal{F})$$

Combining all these results, it follows that $\mathrm{Rad}_n(\mathcal{F})$ gives a bound on the estimation error *in expectation*. More precisely, using (1), we have for the ERM estimator $\hat{f}$:

$$\mathbf{E}[\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f)] \le 2\mathbf{E}[\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \hat{\mathcal{R}}(f)|] \le 4G\,\mathrm{Rad}_n(\mathcal{F}). \tag{7}$$

To also deduce high-probability bounds once we have an expectation bound, we can use MacDiarmid's inequality (wkipdedia link.

4.2 Application to linear predictors

Consider the class of predictors

$$\mathcal{F} = \{f : x \mapsto \theta^\top \phi(x),\ \|\theta\|_2 \leq D\}.$$

and assume that $\mathbf{E}\|\phi(x_i)\|_2^2 \leq R^2$.

$$\begin{aligned}
\mathrm{Rad}_n(\mathcal{F}) &= \mathbf{E}_{\epsilon_i, x_i} \sup_{\|\theta\|_2 \leq D} \frac{1}{n} \sum \epsilon_i \theta^\top \phi(x_i) \\
&= \mathbf{E}\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \phi(x_i)\|_2 \cdot D \\
&\leq \sqrt{\mathbf{E}\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \phi(x_i)\|_2^2} \cdot D \\
&= \sqrt{\frac{1}{n^2} \sum_{i=1}^n \mathbf{E}\epsilon_i^2 \|\phi(x_i)\|_2^2} \cdot D \\
&= \frac{DR}{\sqrt{n}}
\end{aligned}$$

Note that we used only one inequality (Jensen's). In fact the bound is tight up to a universal constant, by Khintchine's inequality. Let us summarize our findings.

Proposition 4.4. *Let ℓ be a G-Lipschitz continuous loss function, linear prediction functions $\mathcal{F} = \{f : x \mapsto \theta^\top \phi(x),\ \|\theta\|_2 \leq D\}$ where $\mathbf{E}\|\phi(x)\|_2^2 \leq R^2$. Let $\hat{f}$ be a minimizer of the empirical risk, then*

$$\mathbf{E}\mathcal{R}(\hat{f}) \leq \inf_{\|\theta\|_2 \leq D} \mathcal{R}(f_\theta) + \frac{4GDR}{\sqrt{n}}.$$

Compared to last week's bound, this one holds in the *random design setting*. For least squares, we can get finer results with more specific tools, but the Rademacher complexity approach works also for more complex models as we will see in the next lectures.

References

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.