ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Chair for Mathematical Information Science

Thomas Allard
Sternwartstrasse 7
CH-8092 Zürich

Mathematics of Information
Uniform Laws of Large Numbers and Rademacher Complexity

This discussion session is largely based on [1, Chap. 2,4 and 5].

1 Introduction

In the present discussion session, we will investigate the connection between uniform laws of large numbers and Rademacher complexity. More specifically, we are interested in relating the Rademacher complexity of a function class $\mathcal{F}$

$$\mathcal{R}_n(\mathcal{F}) \coloneqq \mathbb{E}_{\varepsilon,X}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right]$$

to the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \coloneqq \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\big[f(X)\big]\right|,$$

and providing conditions on the Rademacher complexity that ensure a strong uniform law of large number, i.e., $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{a.s.}{\to} 0$. In particular, we will restrict our study to the case of uniformly bounded function classes, that is, function classes $\mathcal{F}$ such that there is a $b > 0$ satisfying $\|f\|_\infty \leqslant b$ for any $f \in \mathcal{F}$. In this special case, the following theorems, stated in the lecture, apply.

Theorem 1. For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $n \geqslant 1$ and any scalar $\delta \geqslant 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leqslant 2\mathcal{R}_n(\mathcal{F}) + \delta$$

with probability at least $1 - e^{-\frac{n\delta^2}{2b^2}}$.

Theorem 2. For any b-uniformly bounded class of functions $\mathcal{F}$, any positive integer $n \geqslant 1$ and any scalar $\delta \geqslant 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geqslant \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{2\sqrt{n}} - \delta$$

with probability at least $1 - e^{-\frac{n\delta^2}{2b^2}}$, where the expectation $\mathbb{E}[f(X)]$ is taken with respect to the probability $\mathbb{P}$.

An important consequence of these theorems is that, for a uniformly bounded class of functions, the Rademacher complexity converging to zero as n grows is a necessary and sufficient condition for a strong uniform law of large numbers. To see this, we apply Borel-Cantelli lemma (cf.

Lemma 6 in Appendix A) to the sequence of events $A_n \coloneqq \{\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 3\delta\}$ for a fixed $\delta > 0$, with

$$\begin{aligned}\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] &= \sum_{n\in\mathbb{N}} \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 3\delta\right] \\ &\leqslant \sum_{n=1}^{K} \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 3\delta\right] + \sum_{n=K+1}^{\infty} \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 2\mathcal{R}_n(\mathcal{F}) + \delta\right] \\ &\leqslant \sum_{n=1}^{K} \mathbb{P}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 3\delta\right] + \sum_{n=K+1}^{\infty} e^{-\frac{n\delta^2}{2b^2}} < \infty,\end{aligned}$$

where K is chosen such that $\mathcal{R}_n(\mathcal{F}) \leqslant \delta$, for all $n \geqslant K$. Borel-Cantelli lemma therefore yields that

$$\mathbb{P}\left[\limsup A_n\right] = \mathbb{P}\left[\limsup \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} > 3\delta\right] = 0.$$

Since we can choose δ arbitrarily small and since $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geqslant 0$, the last equation yields a strong uniform law of large numbers, namely

$$\mathbb{P}\left[\lim \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = 0\right] = 1.$$

The proofs of the theorems require some results on the concentration of random variables around their means that we present in Section 2. Section 3 and Section 4 go through the proofs of respectively Theorem 1 and Theorem 2. Section 5 introduces a new technique called *chaining* that will allow us to upper-bound the Rademacher complexity by means of metric entropy and VC dimension in Section 6.

2 Concentration Inequalities

The term *concentration inequality* refers to a variety of inequalities upper-bounding the probability that a random variable deviates from its mean by a given amount. The most classical concentration bounds are the laws of large numbers, which ensure that, with high probability, the average of independent random variables is close to their expectation. This section provides a brief introduction to the theory of concentration of measure; we refer the interested reader to [2] for a deeper analysis.

> **Remark 1.** Note that you have already encountered different types of concentration inequalities in the lecture (e.g., in the proof of the Johnson-Lindenstrauss lemma) as well as in previous discussion sessions (e.g., the Bernstein inequality).

The most elementary tail bound is *Markov's inequality*. Let X be a non-negative random variable with finite mean. Markov's inequality states that

$$\mathbb{P}[X \geqslant \delta] \leqslant \frac{\mathbb{E}[X]}{\delta}, \quad \forall \delta > 0. \tag{1}$$

For a random variable X of finite variance $Var(X)$, applying Markov's inequality to the random variable $(X - \mathbb{E}[X])^2$ gives *Chebyshev's inequality*:

$$\mathbb{P}\left[\left|X - \mathbb{E}[X]\right| \geqslant \delta\right] \leqslant \frac{Var(X)}{\delta^2}, \quad \forall \delta > 0.$$

This is a simple form of concentration inequality, guaranteeing that X is close to its mean with high probability when the variance is small.

A wide variety of concentration bounds can be obtained from Markov's inequality (1). In what follows, we are particularly interested in bounding random variables X for which the moment generating function $\mathbb{E}\big[e^{\lambda X}\big]$ is defined for all $\lambda \in \mathbb{R}$. Markov's inequality then gives

$$\mathbb{P}[X - \mu \geqslant \delta] = \mathbb{P}\big[e^{\lambda(X-\mu)} \geqslant e^{\lambda\delta}\big] \leqslant \frac{\mathbb{E}\big[e^{\lambda(X-\mu)}\big]}{e^{\lambda\delta}}, \quad \forall \delta > 0. \tag{2}$$

Minimizing the RHS of (2) over λ so as to obtain the tightest possible result yields the *Chernoff bound*:

$$\log \mathbb{P}[X - \mu \geqslant \delta] \leqslant \inf_{\lambda \in \mathbb{R}} \Big\{ \log \mathbb{E}\big[e^{\lambda(X-\mu)}\big] - \lambda\delta \Big\}. \tag{3}$$

Example 1 (Sub-Gaussian random variables)**.** A random variable X with finite expectation μ is said to be *sub-Gaussian* if there is a positive number σ such that

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

In particular, any Gaussian variable with variance σ^2 is sub-Gaussian with parameter σ. Combining the Chernoff bound (3) with the definition of sub-Gaussian random variables yields the following upper-deviation inequality

$$\log \mathbb{P}[X \geqslant \mu + \delta] \leqslant \inf_{\lambda \in \mathbb{R}} \Big\{ \frac{\sigma^2\lambda^2}{2} - \lambda\delta \Big\} = -\frac{\delta^2}{2\sigma^2}, \tag{4}$$

where the optimum of the quadratic function in λ has been found by differentiation. A similar argument applied to the sub-Gaussian random variable $-X$ yields the following lower-deviation inequality

$$\log \mathbb{P}[X \leqslant \mu - \delta] \leqslant -\frac{\delta^2}{2\sigma^2}. \tag{5}$$

Combining 4 and 5 through a union bound argument, we conclude that any sub-Gaussian variable satisfies the concentration inequality

$$\mathbb{P}[|X - \mu| \geqslant \delta] \leqslant 2e^{-\frac{\delta^2}{2\sigma^2}}. \tag{6}$$

We finalize this example by noting that sub-Gaussian random variables are very useful in practice as they extend concentration results of Gaussian variables to a broader class of random variables. For instance, it can be shown (cf. homework) that every random variable $X_{a,b}$ taking value in $[a, b]$ almost surely, with $-\infty < a < b < \infty$, is sub-Gaussian with parameter upper-bounded by $\frac{b-a}{2}$. In particular, we obtain from (6) that every bounded random variable $X_{a,b}$ satisfies the following concentration inequality:

$$\mathbb{P}[|X_{a,b} - \mu| \geqslant \delta] \leqslant 2e^{-\frac{2\delta^2}{(b-a)^2}}. \tag{7}$$

We now turn our attention to concentration inequalities for functions of independent random variables. Let $\{X_i\}_{i=1}^n$ be a sequence of independent random variables, and consider the random

variable $f(X) = f(X_1, \ldots, X_n)$ for some function $f \colon \mathbb{R}^n \to \mathbb{R}$. We are interested in obtaining bounds on the deviation of f from its mean. To this end, we consider the sequence of random variables $\{Y_k\}_{k=0}^n$ given by $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$ and

$$Y_k = \mathbb{E}[f(X)|X_1, \ldots, X_k], \quad \forall k = 1, \ldots, n-1,$$

where we assumed that the conditional expectations exist. Note that Y_0 is deterministic and the random variables Y_k will tend to exhibit more "fluctuations" as we move along the sequence from Y_0 to Y_n. Based on this intuition, the tail bounds are based on the following telescoping decomposition:

$$f(X) - \mathbb{E}[f(X)] = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{=:D_k},$$

in which the deviation $f(X) - \mathbb{E}[f(X)]$ is written as a sum of increments $D_k \coloneqq Y_k - Y_{k-1}$. The sequence $\{Y_k\}_{k=0}^n$ is an example of a martingale sequence, known as a *Doob martingale*, whereas the sequence $\{D_k\}_{k=1}^n$ is an example of a martingale difference sequence. Appendix B recalls basic definitions and properties about martingales.

The following lemma provides a concentration inequality for bounded martingale difference sequences.

Lemma 1 (Azuma-Hoeffding inequality)**.** Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^n$ be a martingale difference sequence for which there exist constants $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in [a_k, b_k]$ almost surely, for all $k = 1, \ldots, n$. Then

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geqslant \delta\right] \leqslant 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^n (b_k - a_k)^2}\right\} \quad \forall \delta \geqslant 0.$$

Proof. Since $D_k \in [a_k, b_k]$ almost surely, the conditional random variable $(D_k|\mathcal{F}_{k-1})$ also belongs to this interval almost surely, and hence is sub-Gaussian with parameter $\sigma_k = \frac{b_k - a_k}{2}$ (cf. Example 1). We therefore have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] &\overset{(*)}{=} \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right] \\ &\leqslant \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\right] e^{\lambda^2 \frac{(b_n - a_n)^2}{8}},\end{aligned}$$

where $(*)$ comes from the definition of conditional expectation (cf. Definition 3 in Appendix B). Iterating this argument, we get

$$\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] \leqslant e^{\lambda^2 \sum_{k=1}^n \frac{(b_k - a_k)^2}{8}},$$

which means that the random variable $\sum_{k=1}^n D_k$ is sub-Gaussian with parameter

$$\sigma \coloneqq \sqrt{\sum_{k=1}^{n} \frac{(b_k - a_k)^2}{4}}.$$

We conclude using the same technique as to obtain (7):

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geqslant \delta\right] \leqslant 2\exp\left\{-\frac{\delta^2}{2\sigma^2}\right\} = 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^n (b_k - a_k)^2}\right\}, \quad \forall \delta \geqslant 0.$$

□

An important application of Lemma 1 concerns functions that satisfy the bounded difference property. Let us first introduce some notation. Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, \ldots, n\}$, we define a new vector $x^{\backslash k} \in \mathbb{R}^n$ via

$$x_j^{\backslash k} \coloneqq \begin{cases} x_j, & \text{if } j \neq k, \\ x'_k, & \text{if } j = k. \end{cases} \tag{8}$$

With this notation in place, we say that $f \colon \mathbb{R}^n \to \mathbb{R}$ satisfies the *bounded difference property* with parameter L if, for each index $k \in \{1, \ldots, n\}$,

$$|f(x) - f(x^{\backslash k})| \leqslant L, \quad x, x' \in \mathbb{R}^n.$$

The following lemma gives a concentration inequality for functions satisfying the bounded difference property.

Lemma 2 (Bounded differences inequality)**.** Suppose that $f \colon \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property with parameter L and that the random vector $X = (X_1, \ldots, X_n)$ has independent components. Then,

$$\mathbb{P}\big[\big|f(X) - \mathbb{E}[f(X)]\big| \geqslant \delta\big] \leqslant 2e^{-\frac{2\delta^2}{nL^2}}, \quad \forall \delta \geqslant 0.$$

Proof. The proof of this lemma uses the Azuma-Hoeffding inequality. In order to use this result, we define the following martingale difference sequence

$$D_k \coloneqq \mathbb{E}\big[f(X)|X_1, \ldots, X_k\big] - \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}\big].$$

We claim that D_k lies in an interval of length at most L almost surely. In order to prove this claim, we introduce the random variables

$$A_k \coloneqq \inf_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] - \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}\big]$$

and

$$B_k \coloneqq \sup_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] - \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}\big].$$

On one hand, we have

$$D_k - A_k = \mathbb{E}\big[f(X)|X_1, \ldots, X_k\big] - \inf_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big],$$

so that $D_k \geqslant A_k$ almost surely. A similar argument shows that $D_k \leqslant B_k$ almost surely.

We now need to show that $B_k - A_k \leqslant L$ almost surely. Observe that by the independence of the $\{X_k\}_{k=1}^n$, we have for any $(x_1, \ldots, x_k)$

$$\mathbb{E}\big[f(X)|x_1, \ldots, x_k\big] = \mathbb{E}_{k+1}\big[f(x_1, \ldots, x_k, X_{k+1}^n)\big],$$

where $\mathbb{E}_{k+1}$ denotes the expectation over $X_{k+1}^n \coloneqq (X_{k+1}, \ldots, X_n)$. Consequently, we have

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] - \inf_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] \\ &= \sup_x \mathbb{E}_{k+1}\big[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n)\big] - \inf_y \mathbb{E}_{k+1}\big[f(X_1, \ldots, X_{k-1}, y, X_{k+1}^n)\big] \\ &\leqslant \sup_{x,y} \big\{\mathbb{E}_{k+1}\big[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n) - f(X_1, \ldots, X_{k-1}, y, X_{k+1}^n)\big]\big\} \\ &\leqslant L, \end{aligned}$$

using the bounded differences assumption. Thus the variable D_k lies within an interval of length at most L almost surely.

The proof is concluded by applying the Azuma-Hoeffding inequality (Lemma 1).

□

Remark 2. Under the same assumptions as in Lemma 2, a slight adaptation of the proof yields a one sided version of the result:

$$\mathbb{P}\big[f(X) - \mathbb{E}[f(X)] \geqslant \delta\big] \leqslant e^{-\frac{2\delta^2}{nL^2}}, \quad \forall \delta \geqslant 0,$$

or

$$\mathbb{P}\big[f(X) - \mathbb{E}[f(X)] \leqslant -\delta\big] \leqslant e^{-\frac{2\delta^2}{nL^2}}, \quad \forall \delta \geqslant 0.$$

3 Proof of Theorem 1

The proof of Theorem 1 is effected in two steps. The first step consists of showing that, for a sequence $\{X_i\}_{i=1}^n$ of i.i.d. random variables, the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \coloneqq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big] \right|$$

is sharply concentrated around its mean. The second step consists of showing that the mean can be upper bounded by the Rademacher complexity up to a constant pre-factor.

Proof. First step: In order to simplify the notation, it is convenient to define the recentered functions $\bar{f}(x) \coloneqq f(x) - \mathbb{E}[f(X)]$. Thinking of the samples as fixed for the moment, consider the function

$$G(x_1, \ldots, x_n) \coloneqq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right|.$$

We claim that G satisfies the bounded difference property required to apply Lemma 2. Since the function G is invariant to permutation of its coordinates, it suffices to bound the difference when the first coordinate x_1 is perturbed. Accordingly, we define the vector $y \in \mathbb{R}$ with $y_i = x_i$ for all $i = 2, \ldots, n$, and seek to bound the difference $|G(x) - G(y)|$. For any function f, we have

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \sup_{h \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{h}(y_i) \right| &\leqslant \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| \\
&\overset{(*)}{\leqslant} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) - \bar{f}(y_i) \right| + \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| \\
&= \frac{1}{n} \big| \bar{f}(x_1) - \bar{f}(y_1) \big| \\
&\overset{(**)}{\leqslant} \frac{2b}{n},
\end{aligned}$$

where $(*)$ holds by the triangle inequality and the final inequality $(**)$ comes from the uniform boundedness assumption $\|f\|_\infty \leqslant b$. Since the inequality holds for any function f, we may take

the supremum over $f \in \mathcal{F}$ on both sides to obtain

$$G(x) - G(y) \leqslant \frac{2b}{n}.$$

Repeating the same arguments with the roles of x and y reversed, we conclude that

$$|G(x) - G(y)| \leqslant \frac{2b}{n}.$$

Therefore, replacing the x_i with the X_i and taking the expectation, we have, from a one sided version of Lemma 2, with probability at least $1 - \exp\{-\frac{n\delta^2}{2b^2}\}$, that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leqslant \mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] + \delta, \tag{9}$$

for all $\delta \geqslant 0$.

Second step: It remains to show that $\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big]$ is upper-bounded by $2\mathcal{R}_n(\mathcal{F})$, which will be accomplished through a *symmetrization* argument. Specifically, let $\{Y_i\}_{i=1}^n$ be a second i.i.d. sequence sampled from $\mathbb{P}$, independent of $\{X_i\}_{i=1}^n$ and note that

$$\begin{aligned}
\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\Big(f(X_i) - \mathbb{E}_{Y_i}\big[f(Y_i)\big]\Big)\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n\big(f(X_i) - f(Y_i)\big)\right]\right|\right] \\
&\leqslant \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\big(f(X_i) - f(Y_i)\big)\right|\right].
\end{aligned}$$

Now, let $(\varepsilon_1, \ldots, \varepsilon_n)$ be an i.i.d. sequence of Rademacher random variables, i.e. such that ε_i takes values in $\{-1, 1\}$ equiprobably, independent of X and Y. For any function $f \in \mathcal{F}$, given that ε_i, X_i and Y_i are independent and that X_i and Y_i have the same distribution, the random vector with components $\varepsilon_i(f(X_i) - f(Y_i))$ has the same joint distribution as the random vector with components $f(X_i) - f(Y_i)$, hence

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\big(f(X_i) - f(Y_i)\big)\right|\right] &= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(X_i) - f(Y_i))\right|\right] \\
&\leqslant 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right] = 2\mathcal{R}_n(\mathcal{F}).
\end{aligned} \tag{10}$$

Combining (9) and (10) establishes the proof. □

4 Proof of Theorem 2

Proof. By a similar argument as in the first step of the proof of Theorem 1, we show that, with probability at least $1 - \exp\{-\frac{n\delta^2}{2b^2}\}$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geqslant \mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] - \delta,$$

for all $\delta \geqslant 0$.

It hence remains to show that $\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big]$ is lower-bounded by $\frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}}$. Once again, we use a symmetrization argument.

$$
\begin{aligned}
\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big]\right|\right] \\
&= \frac{1}{2}\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big]\right|\right] + \frac{1}{2}\mathbb{E}_Y\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}\big[f(Y)\big]\right|\right] \\
&\overset{(i)}{\geqslant} \frac{1}{2}\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left\{\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big]\right| + \left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}\big[f(Y)\big]\right|\right\}\right] \\
&\overset{(ii)}{\geqslant} \frac{1}{2}\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right] \\
&\overset{(iii)}{=} \frac{1}{2}\mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f(X_i) - f(Y_i)\big)\right|\right] \\
&\overset{(iv)}{\geqslant} \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f(X_i) - \mathbb{E}[f(Y)]\big)\right|\right],
\end{aligned}
$$

where (i) is due to the sum of sup being greater or equal to the sup of the sum, (ii) is by the triangle inequality, (iii) comes from the observation that the variables $\varepsilon_i\big(f(X_i) - f(Y_i)\big)$ and $f(X_i) - f(Y_i)$ are identically distributed and (iv) is obtained by interchanging the expectation and the supremum. We further note that, combining the triangle inequality with the fact that a sum of sup is greater or equal to the sup of the sum, we get

$$
\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f(X_i) - \mathbb{E}[f(X)]\big)\right|\right] \geqslant \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] - \mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}[f(X)]\right|\right].
$$

The first term $\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$ equals the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ and the second term $\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}[f(X)]\right|\right]$ can be upper-bounded as follows:

$$
\begin{aligned}
\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}[f(X)]\right|\right] &= \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\,\mathbb{E}_{\varepsilon}\left[\left|\sum_{i=1}^n \varepsilon_i\right|\right] \\
&= \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\,\mathbb{E}_{\varepsilon}\left[\sqrt{\left(\sum_{i=1}^n \varepsilon_i - \mathbb{E}_{\varepsilon}\left[\sum_{i=1}^n \varepsilon_i\right]\right)^2}\right] \\
&\overset{(*)}{\leqslant} \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\sqrt{Var_{\varepsilon}\left[\sum_{i=1}^n \varepsilon_i\right]} \\
&\overset{(**)}{=} \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\sqrt{\sum_{i=1}^n Var_{\varepsilon_i}[\varepsilon_i]} = \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{\sqrt{n}},
\end{aligned}
$$

where $(*)$ is Jensen's inequality (cf. Lemma 7 in Appendix A) and $(**)$ holds because the variance of the sum of independent random variables is the sum of the variances of these random variable. Combining all the results yields the claim.

□

5 Dudley's entropy integral

Given a uniformly bounded function class $\mathcal{F}$, we know from Theorem 1 and Theorem 2 that the Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}) \coloneqq \mathbb{E}_{\varepsilon,X}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right]$$

going to zero as n grows provides a necessary and sufficient condition for the class $\mathcal{F}$ to be Glivenko-Cantelli. Therefore, we are now interested in upper-bounding the Rademacher complexity. To do so, we start by providing an upper-bound on the maximum of sub-Gaussian random variables. Namely, we prove the following lemma.

Lemma 3. For $n \geqslant 2$, let $\{X_i\}_{i=1}^n$ be a set of zero-mean random variables, each sub-Gaussian with parameter σ. Then

$$\mathbb{E}\left[\max_{i=1,\dots,n}|X_i|\right] \leqslant 2\sigma\sqrt{\log n}.$$

Note that the random variables X_i are not assumed to be independent.

Proof. By Jensen's inequality inequality and using the sub-Gaussian assumption on the X_i, we obtain, for all $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\left[\max_{i=1,\dots,n}X_i\right] &\leqslant \frac{1}{\lambda}\log\mathbb{E}\left[\exp\left\{\lambda\max_{i=1,\dots,n}X_i\right\}\right]\\ &\leqslant \frac{1}{\lambda}\log\sum_{i=1}^{n}\mathbb{E}\left[\exp\left\{\lambda X_i\right\}\right]\\ &\leqslant \frac{\log n}{\lambda}+\frac{\lambda\sigma^2}{2}.\end{aligned}$$

Optimizing over the choice of λ, we obtain

$$\mathbb{E}\left[\max_{i=1,\dots,n}X_i\right] \leqslant \sigma\sqrt{2\log n}.$$

We apply this result to the set $\{Y_j\}_{j=1}^{2n}$ of zero-mean random variables, each sub-Gaussian with parameter σ defined in the following way:

$$Y_j \coloneqq \begin{cases} X_i & \text{if } j = 2i,\\ -X_i & \text{if } j = 2i-1,\end{cases}$$

which gives the desired result

$$\mathbb{E}\left[\max_{i=1,\dots,n}|X_i|\right] = \mathbb{E}\left[\max_{j=1,\dots,2n}Y_j\right] \leqslant \sigma\sqrt{2\log 2n} \leqslant 2\sigma\sqrt{\log n}.$$

□

Remark 3. For a function class $\mathcal{F}$ with polynomial discrimination of order ν, Lemma 3 already provides an upper bound on the Rademacher complexity. Indeed, the Rademacher complexity can be rewritten

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{\varepsilon,X}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_{\varepsilon,X}\left[\max_{\theta\in\mathcal{F}(X_1^n)}|\langle\varepsilon,\theta\rangle|\right],\end{aligned}$$

where, by assumption, the set $\mathcal{F}(X_1^n)$ contains at most $(n+1)^\nu$ elements. To be able to apply Lemma 3, we need to first show that, for any $\theta \in \mathcal{F}(X_1^n)$, the random variable $\langle\varepsilon,\theta\rangle$ is sub-Gaussian with parameter $\sigma \coloneqq \sup_{\theta\in\mathcal{F}(X_1^n)}\sqrt{\sum_{i=1}^n\theta_i^2}$:

$$\begin{aligned}\mathbb{E}_\varepsilon\left[e^{\lambda\langle\varepsilon,\theta\rangle}\right] &= \prod_{i=1}^n \mathbb{E}_{\varepsilon_i}\left[e^{\lambda\varepsilon_i\theta_i}\right] \\ &= \prod_{i=1}^n \frac{e^{\lambda\theta_i}+e^{-\lambda\theta_i}}{2} \\ &\leqslant e^{\frac{\lambda^2}{2}\sum_{i=1}^n\theta_i^2} \\ &\leqslant e^{\frac{\lambda^2}{2}\left\{\sup_{\theta\in\mathcal{F}(X_1^n)}\sum_{i=1}^n\theta_i^2\right\}},\end{aligned}$$

where we used the identity $e^x + e^{-x} \leqslant 2e^{\frac{x^2}{2}}$ for all $x \in \mathbb{R}$ (the verification is left as an exercise). Applying Lemma 3 provides the following upper-bound on the Rademacher complexity already stated in the lecture

$$\mathcal{R}_n(\mathcal{F}) \leqslant 2D_n\sqrt{\frac{\nu\log(n+1)}{n}},$$

with $D_n \coloneqq \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\sqrt{\frac{\sum_{i=1}^n f(X_i)^2}{n}}\right]$. In the case where $\mathcal{F}$ is b-uniformly bounded, this upper bounded reduces to

$$\mathcal{R}_n(\mathcal{F}) \leqslant 2b\sqrt{\frac{\nu\log(n+1)}{n}} \xrightarrow[n\to\infty]{} 0. \tag{11}$$

We shall see in Section 6 that a more careful analysis yields a faster rate of convergence.

The result of Lemma 3 can actually be refined yielding the so-called *Dudley's entropy integral bound*. We consider a metric space $(\mathbb{T},\rho)$ of diameter $D \coloneqq \sup_{\theta_1,\theta_2\in\mathbb{T}}\rho(\theta_1,\theta_2)$ with ε-covering number $N(\varepsilon;\mathbb{T},\rho)$. The main result of this section reads as follows:

Lemma 4 (Dudley's entropy integral bound)**.** Let $\{X_\theta\}_{\theta\in\mathbb{T}}$ be a zero-mean random variable, such that $X_{\theta_1} - X_{\theta_2}$ is sub-Gaussian with parameter $\rho(\theta_1,\theta_2)$ for any $\theta_1,\theta_2 \in \mathbb{T}$. Then, we have

$$\mathbb{E}\Big[\sup_{\theta_1,\theta_2\in\mathbb{T}}\big|X_{\theta_1} - X_{\theta_2}\big|\Big] \leqslant 32\int_0^D \sqrt{\log N\big(t;\mathbb{T},\rho\big)}\,dt.$$

Proof. We fix a $\delta \in (0; D)$ and consider a δ-cover $\mathbb{C}_\delta = \{\theta_1, \ldots, \theta_N\}$ of the metric space $(\mathbb{T}, \rho)$. By definition of the cover $\mathbb{C}_\delta$, for any $\theta \in \mathbb{T}$, we can find some $\theta_i \in \mathbb{C}_\delta$ such that $\rho(\theta, \theta_i) \leqslant \delta$. Therefore

$$\begin{aligned}\big|X_\theta - X_{\theta_1}\big| &\leqslant \big|X_\theta - X_{\theta_i}\big| + \big|X_{\theta_i} - X_{\theta_1}\big| \\ &\leqslant \sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ \rho\gamma,\gamma') \leqslant \delta}} \big|X_\gamma - X_{\gamma'}\big| + \max_{i=1,\ldots,N} \big|X_{\theta_i} - X_{\theta_1}\big|.\end{aligned}$$

Given some other arbitrary $\tilde{\theta} \in \mathbb{T}$, the same upper-bound holds for $\big|X_{\theta_1} - X_{\tilde{\theta}}\big|$, so that adding together the bounds, we obtain

$$\sup_{\theta,\tilde{\theta} \in \mathbb{T}} \big|X_\theta - X_{\tilde{\theta}}\big| \leqslant 2 \sup_{\substack{\gamma,\gamma' \in \mathbb{T} \\ \rho(\gamma,\gamma') \leqslant \delta}} \big|X_\gamma - X_{\gamma'}\big| + 2 \max_{i=1,\ldots,N} \big|X_{\theta_i} - X_{\theta_1}\big|.$$

Define, for each integer $m = 1, \ldots, L$, a minimal $\varepsilon_m \coloneqq D2^{-m}$-cover $\mathbb{C}_m$ of $\mathbb{C}_\delta$ in the metric ρ. Note that any element of $\mathbb{T}$ is allowed in the cover. Since $\mathbb{C}_\delta$ is a subset of $\mathbb{T}$, each set has its cardinality upper-bounded as $|\mathbb{C}_m| \leqslant N(\varepsilon_m; \mathbb{T}, \rho)$. Since $\mathbb{C}_\delta$ is finite, there exists some integer L for which $\mathbb{C}_L = \mathbb{C}_\delta$. For each $m = 1, \ldots, L$, define the mapping $\pi_m \colon \mathbb{C}_\delta \to \mathbb{C}_m$ via

$$\pi_m(\theta) = \operatorname*{argmin}_{\beta \in \mathbb{C}_m} \rho(\theta, \beta),$$

so that $\pi_m(\theta)$ is the best approximation of $\theta \in \mathbb{C}_\delta$ from the set $\mathbb{C}_m$. Using this notation, we can decompose the random variable X_θ into a sum of increments in terms of an associated sequence $(\gamma_1, \ldots, \gamma_L)$, where we define $\gamma_L \coloneqq \theta$ and $\gamma_{m-1} \coloneqq \pi_{m-1}(\gamma_m)$ recursively for $m = L, \ldots, 2$. By construction, we then have the chaining relation

$$X_\theta - X_{\gamma_1} = \sum_{m=2}^{L} X_{\gamma_m} - X_{\gamma_{m-1}},$$

and hence

$$\big|X_\theta - X_{\gamma_1}\big| = \sum_{m=2}^{L} \max_{\beta \in \mathbb{C}_m} \big|X_\beta - X_{\pi_{m-1}(\beta)}\big|,$$

We have decomposed the difference between X_θ and the final element X_{γ_1}in its associated chain as a sum of increments. Given any other $\tilde{\theta} \in \mathbb{C}_\delta$, we can define the chain $(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_L)$, and then derive an analogous bound for the increment $\big|X_{\tilde{\theta}} - X_{\tilde{\gamma}_1}\big|$. By applying the triangle inequality, we obain

$$\big|X_\theta - X_{\tilde{\theta}}\big| \leqslant \big|X_\theta - X_{\gamma_1}\big| + \big|X_{\tilde{\theta}} - X_{\tilde{\gamma}_1}\big| + \big|X_{\gamma_1} - X_{\tilde{\gamma}_1}\big|.$$

Taking the maximum over $\theta, \tilde{\theta} \in \mathbb{C}_\delta$ on the left-hand side and using our upper-bounds on the right-hand side, we obtain

$$\max_{\theta,\tilde{\theta} \in \mathbb{C}_\delta} \big|X_\theta - X_{\tilde{\theta}}\big| \leqslant 2 \sum_{m=2}^{L} \max_{\beta \in \mathbb{C}_m} \big|X_\beta - X_{\pi_{m-1}(\beta)}\big| + \max_{\gamma,\tilde{\gamma} \in \mathbb{C}_1} \big|X_{\gamma_1} - X_{\tilde{\gamma}_1}\big|.$$

We first upper-bound the maximum over $\mathbb{C}_1$, which has $N(D/2; \mathbb{T}, \rho)$ elements. By assumption, the increment is sub-Gaussian with parameter at most $\rho(\gamma, \tilde{\gamma}) \leqslant D$. Consequently, we have

$$\mathbb{E}\Big[\max_{\gamma,\tilde{\gamma} \in \mathbb{C}_1} \big|X_{\gamma_1} - X_{\tilde{\gamma}_1}\big|\Big] \leqslant 2D\sqrt{\log N\Big(\frac{D}{2}; \mathbb{T}, \rho\Big)}.$$

Similarly, for each $m = 2, \dots, L$, the set $\mathbb{C}_m$ has $N(D2^{-m}; \mathbb{T}, \rho)$ elements, and, moreover, $\max_{\beta \in \mathbb{C}_m} \rho\big(\beta, \pi_{m-1}(\beta)\big) \leqslant D2^{-(m-1)}$, whence

$$\mathbb{E}\Big[\max_{\beta \in \mathbb{C}_m} \big| X_\beta - X_{\pi_{m-1}(\beta)} \big| \Big] \leqslant 2D2^{-(m-1)} \sqrt{\log N\big(D2^{-m}; \mathbb{T}, \rho\big)}.$$

Combining the pieces, we conclude that

$$\mathbb{E}\Big[\max_{\gamma, \tilde{\gamma} \in \mathbb{C}_\delta} \big| X_{\gamma_1} - X_{\tilde{\gamma}_1} \big| \Big] \leqslant 4 \sum_{m=1}^{L} D2^{-(m-1)} \sqrt{\log N\big(D2^{-m}; \mathbb{T}, \rho\big)}.$$

Since the metric entropy $\log N(t; \mathbb{T}, \rho)$ is non-increasing in t, we have

$$D2^{-(m-1)} \sqrt{\log N\big(D2^{-m}; \mathbb{T}, \rho\big)} \leqslant 4 \int_{D2^{-(m+1)}}^{D2^{-m}} \sqrt{\log N\big(u; \mathbb{T}, \rho\big)}\, du,$$

and hence

$$\mathbb{E}\Big[\max_{\theta, \tilde{\theta} \in \mathbb{C}_\delta} \big| X_\theta - X_{\tilde{\theta}} \big| \Big] \leqslant 32 \int_{\frac{\delta}{4}}^{D} \sqrt{\log N\big(u; \mathbb{T}, \rho\big)}\, du.$$

We conclude the proof by taking the limit $\delta \to 0^+$.

□

The method used in the proof is known as *chaining*. It essentially consists in upper-bounding the supremum of sub-Gaussian processes by decomposing it into a sum of maxima over finite sets that are successively refined. These maxima are upper-bounded by a function of the metric entropy through Lemma 3. For a deeper analysis of chaining and its applications, we refer the interested reader to [3, Chapter 11].

Remark 4. Note that we have actually proven a stronger version of the Dudley entropic bound:

$$\mathbb{E}\Big[\sup_{\theta_1, \theta_2 \in \mathbb{T}} \big| X_{\theta_1} - X_{\theta_2} \big| \Big] \leqslant 2\mathbb{E}\Big[\sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho(\gamma, \gamma') \leqslant \delta}} \big| X_\gamma - X_{\gamma'} \big| \Big] + 32 \int_{\frac{\delta}{4}}^{D} \sqrt{\log N\big(t; \mathbb{T}, \rho\big)}\, dt.$$

Even though the additional flexibility to choose $\delta \in [0; D]$ can be useful in certain problems, taking $\delta = 0$ is sufficient for our purpose.

In fact, there exists a lot of different versions of Dudley's entropic bound in the literature. First of all, it is possible, with a slight adaptation of the proof, to derive a better bound for the quantity $\mathbb{E}\Big[\sup_{\theta_1, \theta_2 \in \mathbb{T}} \big(X_{\theta_1} - X_{\theta_2} \big) \Big]$, i.e. without taking the absolute values of the difference of the random variables. This type of bound also provides a bound on the supremum of the process using the zero-mean assumption:

$$\mathbb{E}\Big[\sup_{\theta \in \mathbb{T}} X_\theta \Big] = \mathbb{E}\Big[\sup_{\theta \in \mathbb{T}} X_\theta - X_{\theta_0} \Big] \leqslant \mathbb{E}\Big[\sup_{\theta_1, \theta_2 \in \mathbb{T}} \big(X_{\theta_1} - X_{\theta_2} \big) \Big].$$

The constant 32 in the bound can also be refined by a more careful analysis.

Finally, it is also possible to derive a weaker bound, by directly bounding the maximum in (5), without refining it beforehand:

$$\mathbb{E}\Big[\sup_{\theta_1, \theta_2 \in \mathbb{T}} \big| X_{\theta_1} - X_{\theta_2} \big| \Big] \leqslant 2\mathbb{E}\Big[\sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho(\gamma, \gamma') \leqslant \delta}} \big| X_\gamma - X_{\gamma'} \big| \Big] + 4D \sqrt{\log N\big(t; \mathbb{T}, \rho\big)}.$$

> Although this bound is worse than the bound provided in Lemma 4, it can turn out to be easier to handle in practical applications.

Dudley's entropic bound turns out to be particularly useful in order to provide some upper-bound on the Rademacher complexity. This statement is made precise in the next section.

6 Bounding the Rademacher Complexity

In the present section, we upper-bound the Rademacher complexity using metric entropy and VC dimension arguments. In particular, we show that, if the VC dimension of the class $\mathcal{F}$ is finite, then $\mathcal{R}_n(\mathcal{F}) \to 0$. In this section, we use Dudley's integral to bound the Rademacher complexity of a b-uniformly bounded class of functions $\mathcal{F}$ of finite VC dimension.

Let's first consider some fixed sample $x_1, \dots, x_n$. We define the zero-mean random variable

$$Z_f \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(x_i),$$

and consider the stochastic process $\{Z_f \mid f \in \mathcal{F}\}$. We verify that the increment $Z_f - Z_g$ is sub-Gaussian with parameter $\|f-g\|_{\mathbb{P}_n} \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-g(x_i))^2}$:

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(Z_f - Z_g)}\right] &= \prod_{i=1}^{n} \mathbb{E}_{\varepsilon_i}\left[e^{\frac{\lambda}{\sqrt{n}}\varepsilon_i(f(x_i)-g(x_i))}\right] \\
&\leqslant \prod_{i=1}^{n} \frac{1}{2}\left[e^{\frac{\lambda}{\sqrt{n}}(f(x_i)-g(x_i))} + e^{-\frac{\lambda}{\sqrt{n}}(f(x_i)-g(x_i))}\right] \\
&\leqslant e^{\frac{\lambda^2}{2n}\sum_{i=1}^{n}(f(x_i)-g(x_i))^2} = e^{\frac{\lambda^2}{2}\|f-g\|_{\mathbb{P}_n}^2},
\end{aligned}$$

where we again used the identity $e^x + e^{-x} \leqslant 2e^{\frac{x^2}{2}}$ for all $x \in \mathbb{R}$.

Let's define $\mathcal{F}_0 \coloneqq \mathcal{F} \cup \{0\}$. By Dudley's integral, we have

$$\begin{aligned}
\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right] &= \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|Z_f\right|\right] \\
&= \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}_0}\left|Z_f\right|\right] \\
&= \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}_0}\left|Z_f - Z_0\right|\right] \\
&\leqslant \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon\left[\sup_{f,g\in\mathcal{F}}\left|Z_f - Z_g\right|\right] \\
&\leqslant \frac{32}{\sqrt{n}}\int_0^{2b}\sqrt{\log N\big(t;\mathcal{F}_0,\|\cdot\|_{\mathbb{P}_n}\big)}\,dt \\
&\leqslant \frac{K}{\sqrt{n}}\int_0^{2b}\sqrt{\log N\big(t;\mathcal{F},\|\cdot\|_{\mathbb{P}_n}\big)}\,dt,
\end{aligned}$$

where we used the fact that $\sup_{f,g\in\mathcal{F}} \|f-g\|_{\mathbb{P}_n} \leqslant 2b$ and K is a universal constant.

We now provide a bound on the metric entropy using the VC dimension ν of the function class $\mathcal{F}$.

Lemma 5 ([4, Theorem 2.6.7]). For a b-uniformly bounded class of functions of finite VC-dimension ν, there is a universal constant C such that

$$N(t;\mathcal{F},\|\cdot\|_{\mathbb{P}_n}) \leqslant C\nu(16e)^{\nu}\left(\frac{b}{t}\right)^{2\nu}, \quad \forall t\in(0;2b).$$

Substituting in the metric entropic bound, we find that there are universal constants c_0 and c_1, depending on b but not on ν and n, such that

$$\begin{aligned}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right] &\leqslant c_0\sqrt{\frac{\nu}{n}}\left\{1+\int_0^{2b}\sqrt{\log b-\log t}\,dt\right\}\\ &\leqslant c_1\sqrt{\frac{\nu}{n}}.\end{aligned}$$

We conclude by taking the expectation over the samples:

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\varepsilon,X}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right] \leqslant c_1\sqrt{\frac{\nu}{n}} \xrightarrow[n\to\infty]{} 0.$$

> **Remark 5.** This result is to be contrasted with the bound obtained in (11). Since, by Sauer-Shelah's lemma, a function class of finite VC dimension ν has a polynomial discrimination of degree at most ν, we observe that employing chaining methods improved the convergence rate by a factor $\sqrt{\log(n+1)}$.

References

[1] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[2] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[3] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[4] A. W. van der Vaart and J. A. Wellner, *Weak Convergence*. New York, NY: Springer New York, 1996, pp. 16–28. [Online]. Available: https://doi.org/10.1007/978-1-4757-2545-2_3

A Auxiliary results

A.1 Probability theory

Definition 1 (Probability space)**.** A probability space $(\Omega, \mathcal{G}, \mathbb{P})$ is a triple consisting of a sample space Ω, a σ-algebra $\mathcal{G}$ containing all the events and a probability measure $\mathbb{P}\colon \mathcal{G} \to [0,1]$.

A probability space is a special case of a measured space, for which the total mass of the measure is 1.

Lemma 6 (Borel-Cantelli)**.** Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of random events on a probability space $(\Omega, \mathcal{G}, \mathbb{P})$ such that $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] < \infty$. Then

$$\mathbb{P}[\limsup A_n] = 0,$$

where $\limsup A_n \coloneqq \bigcap_{n=0}^{\infty} \left(\bigcup_{k\geqslant n}^{\infty} A_k \right)$.

Proof. Since $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] < \infty$, we have

$$\mathbb{E}\left[\sum_{n\in\mathbb{N}} \mathbb{1}_{A_n}\right] = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n] < \infty.$$

It implies that $\sum_{n\in\mathbb{N}} \mathbb{1}_{A_n}$ is almost surely finite, or, equivalently,

$$\mathbb{P}[\limsup A_n] = 0.$$

□

A.2 Convex functions

Definition 2 (Convex function)**.** A function $f\colon \mathbb{R} \to \mathbb{R}$ is convex on the real line $\mathbb{R}$ if, for any $x, y \in \mathbb{R}$ and $\lambda \in (0,1)$, we have

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y).$$

If the inequality in the definition above is a strict inequality, then we say that f is strictly convex. A function f such that $-f$ is convex is called concave.

A powerful tool to manipulate convex functions is *Jensen's inequality*:

Lemma 7 (Jensen's inequality)**.** Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space, let φ be a convex function and X a real-valued random variable. Then, we have

$$\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)].$$

Note that there exists various forms of Jensen's inequality in the literature. We have stated here a version in a probabilistic setting that will be useful for our purpose.

> **Example 2.** We consider two examples of convex functions that appears frequently in our analysis.

(a) The square function

$$\varphi\colon x \in \mathbb{R} \mapsto x^2 \in \mathbb{R}$$

is convex. Applying Jensen's inequality yields

$$\mathbb{E}[X]^2 \leqslant \mathbb{E}\left[X^2\right],$$

or, equivalently,

$$\mathbb{E}[X] \leqslant \sqrt{\mathbb{E}\left[X^2\right]}.$$

(b) The exponential function

$$\varphi\colon x \in \mathbb{R} \mapsto e^x \in \mathbb{R}$$

is convex. Applying Jensen's inequality yields

$$e^{\mathbb{E}[X]} \leqslant \mathbb{E}\left[e^X\right],$$

or, equivalently,

$$\mathbb{E}[X] \leqslant \log \mathbb{E}\left[e^X\right].$$

B Martingales

Definition 3 (Conditional expectation)**.** Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space. Let $\mathcal{G}' \subset \mathcal{G}$ be a sub-σ-algebra of $\mathcal{G}$ and X an integrable random variable. Then there exists a random variable Z, $\mathcal{G}'$ measurable and integrable, such that, for all bounded random variable U that is $\mathcal{G}'$-measurable

$$\mathbb{E}[XU] = \mathbb{E}[ZU].$$

Such a random variable Z is called the conditional expectation of X with respect to $\mathcal{G}'$ and is written

$$Z = \mathbb{E}[X \,|\, \mathcal{G}'].$$

When $\mathcal{G}' = \sigma(X_1, \ldots, X_n)$ is the σ-algebra generated by a sequence of random variables $\{X_k\}_{k=1}^n$ and X a random variable, then we write $\mathbb{E}[X \,|\, X_1, \ldots, X_n]$ for $\mathbb{E}[X \,|\, \mathcal{G}']$.

Lemma 8. The conditional expectation satisfies the following properties:

- Linearity:
$$\mathbb{E}[aX + bY \,|\, \mathcal{G}'] = a\mathbb{E}[X \,|\, \mathcal{G}'] + b\mathbb{E}[Y \,|\, \mathcal{G}']. \tag{12}$$
- Pulling out known factors:
$$\mathbb{E}[XY \,|\, \mathcal{G}'] = X\mathbb{E}[Y \,|\, \mathcal{G}'], \tag{13}$$
if X is $\mathcal{G}'$-measurable.
- Tower property:
$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}_2] \,|\, \mathcal{G}_1] = \mathbb{E}[X \,|\, \mathcal{G}_1], \quad \text{for } \mathcal{G}_1 \subset \mathcal{G}_2. \tag{14}$$
- Law of total expectation:
$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}']] = \mathbb{E}[X]. \tag{15}$$

Definition 4 (Filtration)**.** A filtration $\{\mathcal{G}_k\}_{k=1}^\infty$ is a sequence of σ-algebras satisfying $\mathcal{G}_k \subseteq \mathcal{G}_{k+1}$, for all $k \geqslant 1$.

A sequence of random variables $\{X_k\}_{k=1}^\infty$ is said to be *adapted* to the filtration $\{\mathcal{G}_k\}_{k=1}^\infty$ if X_k is measurable with respect to $\mathcal{G}_k$, for all $k \geqslant 1$.

Definition 5 (Martingale)**.** Given a sequence of random variables $\{X_k\}_{k=1}^\infty$ adapted to the filtration $\{\mathcal{G}_k\}_{k=1}^\infty$, the sequence of pairs $\{(X_k, \mathcal{G}_k)\}_{k=1}^\infty$ is a martingale if, for all $k \geqslant 1$,

$$\mathbb{E}[|X_k|] < \infty \quad \text{and} \quad \mathbb{E}[X_{k+1} \,|\, \mathcal{G}_k] = X_k.$$

When the sequence is a martingale with respect to itself (i.e., with $\mathcal{G}_k = \sigma(X_1, \ldots, X_k)$), then we simply write that $\{X_k\}_{k=1}^\infty$ is a martingale.

Example 3 (Doob Martingales)**.** Consider a sequence of real-valued independent random variables $\{X_k\}_{k=1}^n$ as well as the random variable $f(X) \coloneqq f(X_1, \ldots, X_n)$, for some function $f\colon \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}[|f(X)|] < \infty$. We construct a sequence of random variables $\{Y_k\}_{k=0}^n$ defined as

$$Y_k = \begin{cases} \mathbb{E}[f(X)] & \text{for } k = 0, \\ f(X) & \text{for } k = n, \\ \mathbb{E}[f(X) \,|\, X_1, \ldots, X_k] & \text{for } k = 1 \ldots, n-1. \end{cases}$$

Such a sequence of random variables is known as a *Doob martingale*. Considering the filtration

$$\mathcal{G}_k = \begin{cases} \{\emptyset, \Omega\} & \text{for } k = 0, \\ \sigma(X_1, \ldots, X_k) & \text{for } k = 1, \ldots, n, \end{cases}$$

we verify that $\{(Y_k, \mathcal{G}_k)\}_{k=0}^n$ actually defines a martingale sequence:

$$\mathbb{E}[|Y_k|] = \mathbb{E}[|\mathbb{E}[f(X) \,|\, \mathcal{G}_k]|] \leqslant \mathbb{E}[\mathbb{E}[|f(X)| \,|\, \mathcal{G}_k]] = \mathbb{E}[|f(X)|] < \infty$$

and

$$\mathbb{E}[Y_{k+1} \,|\, \mathcal{G}_k] = \mathbb{E}[\mathbb{E}[f(X) \,|\, \mathcal{G}_{k+1}] \,|\, \mathcal{G}_k] = \mathbb{E}[f(X) \,|\, \mathcal{G}_k] = Y_k$$

by the tower property (14).

A closely related notion is that of *martingale difference sequence*.

Definition 6 (Martingale difference sequence)**.** An adapted sequence $\{(D_k, \mathcal{G}_k)\}_{k=1}^\infty$ is a martingale difference sequence if, for all $k \geqslant 1$,

$$\mathbb{E}[|D_k|] < \infty \quad \text{and} \quad \mathbb{E}[D_{k+1} \,|\, \mathcal{G}_k] = 0.$$

As suggested by their name, such difference sequences arise in a natural way from martingales. In particular, given a martingale $\{(X_k, \mathcal{G}_k)\}_{k=0}^\infty$, let us define $D_k \coloneqq X_k - X_{k-1}$, for $k \geqslant 1$. We then have

$$\begin{aligned} \mathbb{E}[D_{k+1} \,|\, \mathcal{G}_k] &= \mathbb{E}[X_{k+1} \,|\, \mathcal{G}_k] - \mathbb{E}[X_k \,|\, \mathcal{G}_k] \\ &= X_k - X_k = 0, \end{aligned}$$

where we have used the definition of a martingale (Definition 5), the linearity of conditional expectation (12) and the fact that X_k is measurable with respect to $\mathcal{G}_k$.