Math of ML : Exercises 3*

September 29, 2025

In the first exercise, we explore some basic properties of sub-Gaussian random variables. Pay special attention to part 5, where we introduce an important technique of bounding maximum by a sum inside a logarithm.

Exercise 1 (On sub-Gaussian random variables)**.** *We say that a random variable X is* sub-Gaussian *with variance proxy σ^2 (also called σ^2-sub-Gaussian) if for any $\lambda \in \mathbb{R}$ we have*

$$\mathbf{E} \exp\left(\lambda(X - \mathbf{E}[X])\right) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Prove the following properties of sub-Gaussian random variables.

1. *A random variable X with normal distribution $N(\mu, \sigma^2)$ is σ^2-sub-Gaussian.*
2. *For any σ^2-sub-Gaussian random variable X we have $\mathrm{Var}(X) \leq \sqrt{2}\sigma^2$.*
3. *For any σ^2-sub-Gaussian random variable X and any $\delta \in (0, 1)$ we have*

$$\mathbf{P}\left(|X - \mathbf{E}X| \geq \sqrt{2\sigma^2 \log(2/\delta)}\right) \leq \delta.$$

4. *For $i = 1, \ldots, n$, let X_i be a σ_i^2-sub-Gaussian random variable. Assuming that the random variables $X_1, \ldots, X_n$ are independent, prove that for any $\lambda \in \mathbb{R}^n$, the linear combination $\sum_{i=1}^n \lambda_i X_i$ is τ^2-sub-Gaussian for some τ^2 that you should identify.*
5. *Let $X_1, \ldots, X_n$ be zero-mean σ^2-sub-Gaussian random variables (not necessarily independent). Prove that*

$$\mathbf{E}\left[\max_{i=1,\ldots,n} X_i\right] \leq \sqrt{2\log(n)\sigma^2}.$$

   Hint: *for any $\lambda > 0$ by Jensen's inequality we have $\mathbf{E}[X] \leq \frac{1}{\lambda}\log \mathbf{E}\exp(\lambda X)$. Apply Jensen's inequality to the random variable $\max_i X_i$ and bound $\max_i X_i$ by $\sum_i X_i$. The trick here is that we replace maximum by a sum* inside the logarithm*!*

6. *Let $X_1, \ldots, X_n$ be zero-mean σ^2-sub-Gaussian random variables (not necessarily independent). Prove that for any $\delta \in (0, 1)$ we have*

$$\mathbf{P}\left(\max_{i=1,\ldots,n} X_i \geq \sqrt{2\log(n/\delta)\sigma^2}\right) \leq \delta.$$

Solution 1. *1. Notice that $X - \mathbf{E}[X]$ is distributed as $\mathcal{N}(0, \sigma^2)$ random variable, whose moment generating function is exactly equal to $\exp(\lambda^2\sigma^2/2)$. Hence, X is σ^2-sub-Gaussian.*

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

2. *By definition of sub-Gaussianity, we have*

$$\mathbf{E}\exp(\lambda(X-\mathbf{E}X)) \le \exp(\lambda^2\sigma^2/2) = 1 + \frac{\lambda^2\sigma^2}{2} + \left(\exp(\lambda^2\sigma^2/2) - 1 - \frac{\lambda^2\sigma^2}{2}\right).$$

Expanding the exponent in the left hand and interchanging the infinite sum with expectation (e.g., by Dominated convergence theorem), we have

$$1 + \frac{1}{2}\lambda^2\mathrm{Var}(X) + \sum_{k=3}^{\infty}\frac{\lambda^k}{k!}\mathbf{E}(X-\mathbf{E}X)^k \le 1 + \frac{\lambda^2\sigma^2}{2} + \left(\exp(\lambda^2\sigma^2/2) - 1 - \frac{\lambda^2\sigma^2}{2}\right).$$

Hence, for any $\lambda > 0$ we have

$$\mathrm{Var}(X) - \sigma^2 \le \frac{2}{\lambda^2}\left(\exp(\lambda^2\sigma^2/2) - 1 - \frac{\lambda^2\sigma^2}{2}\right) - \frac{2}{\lambda^2}\sum_{k=3}^{\infty}\frac{\lambda^k}{k!}\mathbf{E}(X-\mathbf{E}X)^k$$

Taking the limit $\lambda \to 0_+$ on both sides, we conclude that

$$\mathrm{Var}(X) - \sigma^2 \le 0,$$

which is what we wanted to show.

3. *We will apply Markov's inequality. For any $\lambda \in \mathbb{R}$ we have*

$$\begin{aligned}\mathbf{P}\left(X - \mathbf{E}X \ge \sqrt{2\sigma^2\log(2/\delta)}\right) &= \mathbf{P}\left(\exp(\lambda(X-\mathbf{E}X)) \ge \exp(\lambda\sqrt{2\sigma^2\log(2/\delta)})\right)\\ &\le \mathbf{E}\exp(\lambda(X-\mathbf{E}X))\exp(-\lambda\sqrt{2\sigma^2\log(2/\delta)})\\ &\le \exp(\lambda^2\sigma^2/2)\exp(-\lambda\sqrt{2\sigma^2\log(2/\delta)}).\end{aligned}$$

Optimizing over λ (i.e., taking $\lambda = \sqrt{2\log(2/\delta)}/\sigma$) yields

$$\mathbf{P}\left(X - \mathbf{E}X \ge \sqrt{2\sigma^2\log(2/\delta)}\right) \le \delta/2. \tag{1}$$

Similarly, we can prove that

$$\mathbf{P}\left(\mathbf{E}X - X \ge \sqrt{2\sigma^2\log(2/\delta)}\right) \le \delta/2. \tag{2}$$

Taking the union bound over the events (1) and (2) concludes the proof.

4. *For any $\lambda \in \mathbb{R}$, by independence of the random variables $X_1, \ldots, X_n$ we have*

$$\mathbf{E}\exp\left(\lambda\sum_{i=1}^n \lambda_i X_i\right) = \prod_{i=1}^n \mathbf{E}_{X_i}\exp(\lambda\lambda_i X_i).$$

Using the fact that X_i is σ_i^2-sub-Gaussian, it follows that

$$\begin{aligned}\mathbf{E}\exp\left(\lambda\sum_{i=1}^n \lambda_i X_i\right) &\le \prod_{i=1}^n \exp(\lambda^2\lambda_i^2\sigma_i^2/2)\\ &= \exp\left(\lambda^2\left(\sum_{i=1}^n \lambda_i^2\sigma_i^2\right)/2\right).\end{aligned}$$

Hence, the random variable $\sum_{i=1}\lambda_i^2\sigma_i^2$ is τ^2-sub-Gaussian, where

$$\tau^2 = \sum_{i=1}^n \lambda_i^2\sigma_i^2.$$

5. *Applying the hint, we have for any $\lambda > 0$:*

$$\begin{aligned}\mathbf{E}\max_{i=1,\dots,n} X_i &\le \frac{1}{\lambda}\log \mathbf{E}\exp\left(\lambda \max_{i=1,\dots,n} X_i\right)\\ &= \frac{1}{\lambda}\log \mathbf{E}\max_{i=1,\dots,n}\exp\left(\lambda X_i\right)\\ &\le \frac{1}{\lambda}\log \mathbf{E}\sum_{i=1,\dots,n}\exp\left(\lambda X_i\right)\\ &\le \frac{1}{\lambda}\log \sum_{i=1}^{n}\mathbf{E}\exp\left(\lambda X_i\right).\end{aligned}$$

Applying the sub-Gaussianity property for each of the summand inside the logarithm yields, for any $\lambda > 0$:

$$\mathbf{E}\max_{i=1,\dots,n} X_i \le \frac{1}{\lambda}\log\left(n\exp(\lambda^2\sigma^2/2)\right) = \frac{\log(n)}{\lambda} + \lambda\sigma^2/2.$$

Optimizing the above bound over $\lambda > 0$ yields the desired result.

6. *We may either proceed as in the previous part of this exercise, or use the (equivalent) approach of the union bound (here the union bound will play the role of bounding maximum by the sum). We have, by the union bound*

$$\begin{aligned}\mathbf{P}\left(\max_{i=1,\dots,n} X_i \ge \sqrt{2\log(n/\delta)\sigma^2}\right) &= \mathbf{P}\left(\cup_{i=1,\dots,n}\left\{X_i \ge \sqrt{2\log(n/\delta)\sigma^2}\right\}\right)\\ &\le \sum_{i=1}^{n}\mathbf{P}\left(\left\{X_i \ge \sqrt{2\log(n/\delta)\sigma^2}\right\}\right)\\ &\le \sum_{i=1}^{n}\delta/n = \delta,\end{aligned}$$

where the last step follows from (the proof of, particularly, see (1)) part 3 of this exercise.

The next two exercises showcase examples of non-trivial (and very useful!) sub-Gaussian random variables.

Exercise 2 (Hoeffding's Lemma)**.** *Let X be a random variable such that $a \le X - \mathbf{E}X \le b$. Let $\psi(\lambda) = \log \mathbf{E}\exp(\lambda(X - \mathbf{E}X))$. By Taylor's theorem, for any $\lambda \in \mathbb{R}$ there exists some $c_\lambda \in [-\lambda, \lambda]$ such that*

$$\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(c_\lambda). \tag{3}$$

1. *Compute $\psi(0)$ and $\psi'(0)$.*
2. *For any $c \in \mathbb{R}$, show that $\psi''(c)$ is equal to the variance of some random variable supported on $[a, b]$.*
3. *Using the fact that variance of a bounded random variables is bounded, deduce that X is $(b-a)^2/4$-sub-Gaussian (this result is known as Hoeffding's lemma).*

Solution 2. *Let $Y = X - \mathbf{E}X$.*

$$\psi'(\lambda) = \frac{\mathbf{E}[Y\exp(\lambda Y)]}{\mathbf{E}[\exp(\lambda Y)]} \tag{4}$$

and

$$\psi''(\lambda) = \frac{\mathbf{E}[Y^2\exp(\lambda Y))]}{\mathbf{E}[\exp(\lambda Y)]} - \frac{\mathbf{E}[Y\exp(\lambda Y)]^2}{\mathbf{E}[\exp(\lambda Y)]^2}.$$

1. *We have* $\psi(0) = \log 1 = 0$*; from* (4)*, we also have* $\psi'(0) = \mathbf{E}Y = \mathbf{E}[X - \mathbf{E}X] = 0$.

2. *Let* P *be the distribution of* Y*. Let* U_λ *be a random variable with distribution* Q_λ *defined by*

$$Q_\lambda(du) = \frac{\exp(\lambda u)}{\mathbf{E}[e^{\lambda Y}]} P(du).$$

Notice that U_λ *is supported on* $[a, b]$ *because* Y *is supported on* $[a, b]$*. Then, we have*

$$\begin{aligned}
\mathrm{Var}(U_\lambda) &= \mathbf{E}U_\lambda^2 - \mathbf{E}[U_\lambda]^2 \\
&= \int_a^b u^2 Q_\lambda(du) - \left(\int_a^b u Q_\lambda(du) \right)^2 \\
&= \int_a^b u^2 \frac{\exp(\lambda u)}{\mathbf{E}[e^{\lambda Y}]} P(du). - \left(\int_a^b u \frac{\exp(\lambda u)}{\mathbf{E}[e^{\lambda Y}]} P(du). \right)^2 \\
&= \psi''(\lambda).
\end{aligned}$$

3. *Using* (3) *with the two previous parts of this exercise, we have for any* $\lambda \in \mathbb{R}$

$$\psi(\lambda) = \frac{1}{2}\lambda^2 \mathrm{Var}(U_{c_\lambda}).$$

Since for any $\lambda \in \mathbb{R}$ *the random variable* U_{c_λ} *is supported on* $[a, b]$*, its variance is at most* $(b-a)^2/4$ *(to see this, note that for any random variable* U *supported on* $[a, b]$ *it holds that* $\mathrm{Var}(U) = \inf_{x \in \mathbb{R}} \mathbf{E}[(U - x)^2] \leq \mathbf{E}[(U - (b+a)/2)^2] \leq (b-a)^2/4$*). Hence,*

$$\psi(\lambda) \leq \frac{1}{2}\lambda^2 (b-a)^2/4,$$

which is what we wanted to show.

Rk: Note that since a *and* b *only come into the result via* $|b - a|$*, the condition* $a \leq X - \mathbf{E}X \leq b$ *a.s. is effectively equivalent to saying that* X *belongs a.s. to some interval* I *of length* $|b - a|$*.*

Exercise 3 (Bounded differences/McDiarmid's inequality)**.** *Let* $f : \mathcal{Z}^n \to \mathbb{R}$ *be a function of bounded variation, that is, a function such that for any* $i \in \{1, \ldots, n\}$*, and any* $z_1, \ldots, z_n, z_i' \in \mathcal{Z}$*, we have*

$$|f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c.$$

Let $Z_1, \ldots, Z_n$ *be independent (and not necessarily identically distributed) random variables on* $\mathcal{Z}$*. Then, the random variable* $U = f(Z_1, \ldots, Z_n)$ *is sub-Gaussian with variance proxy* $nc^2/4$*.*

Hint: *The idea is to decompose* U *into a telescoping sum of* conditionally *independent random variables:*

$$U - \mathbf{E}U = (U_n - U_{n-1}) + ... + (U_2 - U_1) + (U_1 - U_0),$$

where $U_i = \mathbf{E}_{Z_{i+1}', \ldots, Z_n'} f(Z_1, ..., Z_i, Z_{i+1}', ..., Z_n')$*. Conclude the proof by a repeated application of Hoeffding's lemma. You may begin the proof by writing* $U - \mathbf{E}U = U_n - U_0 = (U_n - U_{n-1}) + (U_{n-1} - U_0)$ *and conditioning on the values of* $Z_1, \ldots, Z_{n-1}$*.*

Solution 3. *Observe that*

$$\mathbf{E}[\exp(\lambda(U - \mathbf{E}U))] = \mathbf{E}[\exp(\lambda(U_n - U_0))].$$

We claim that for any $i \in \{1, \dots, n\}$ *it holds that*

$$\mathbf{E}_{X_1,\dots,X_i}[\exp(\lambda(U_i - U_0))] \leq \exp\left(\frac{1}{2}\lambda^2 c^2/4\right) \mathbf{E}_{X_1,\dots,X_{i-1}}[\exp(\lambda(U_{i-1} - U_0))]. \tag{5}$$

We will now prove (5)*. We have*

$$\begin{aligned}
&\mathbf{E}_{X_1,\dots,X_i}[\exp(\lambda(U_i - U_0))] \\
&\quad = \mathbf{E}_{X_1,\dots,X_i}[\exp(\lambda(U_i - U_{i-1} + U_{i-1} - U_0))] \\
&\quad = \mathbf{E}_{X_1,\dots,X_{i-1}} \mathbf{E}_{X_i}[\exp(\lambda(U_i - U_{i-1} + U_{i-1} - U_0))|X_1, \dots, X_{i-1}] \\
&\quad = \mathbf{E}_{X_1,\dots,X_{i-1}} \exp(\lambda(U_{i-1} - U_0)) \mathbf{E}_{X_i}[\exp(\lambda(U_i - U_{i-1}))|X_1, \dots, X_{i-1}]. \qquad (6)
\end{aligned}$$

Now observe that

$$\begin{aligned}
&\mathbf{E}_{X_i}[\exp(\lambda(U_i - U_{i-1})|X_1, \dots, X_{i-1}] \\
&= \mathbf{E}_{X_i}[\exp(\lambda(U_i - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}])|X_1, \dots, X_{i-1}].
\end{aligned}$$

By the bounded variation assumption on the function f*, we have, conditionally on the values of* $X_1, \dots, X_{i-1}$*,*

$$\begin{aligned}
&U_i - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}] \\
&= \mathbf{E}_{X_{i+1},\dots,X_n}[U|X_1, \dots, X_i] - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}] \\
&\leq \mathbf{E}_{X_{i+1},\dots,X_n}\left[\sup_{x_i} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n)|X_1, \dots, X_i\right] - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}] \\
&= \mathbf{E}_{X_i,\dots,X_n}\left[\sup_{x_i} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n)|X_1, \dots, X_{i-1}\right] - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}] \\
&= \mathbf{E}_{X_i,\dots,X_n}\left[\sup_{x_i} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n) - f(X_1, \dots, X_n)|X_1, \dots, X_{i-1}\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
&U_i - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}] \\
&\geq \mathbf{E}_{X_i,\dots,X_n}\left[\inf_{x_i} f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_n) - f(X_1, \dots, X_n)|X_1, \dots, X_{i-1}\right].
\end{aligned}$$

Hence, conditionally on $X_1, \dots, X_{i-1}$*, the random variable* $U_i - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}]$ *lies in a bounded interval of length at most* c*. Therefore, by Hoeffding's lemma (cf. Exercise 2) we have*

$$\mathbf{E}_{X_i}[\exp(\lambda(U_i - \mathbf{E}_{X_i}[U_i|X_1, \dots, X_{i-1}])|X_1, \dots, X_{i-1}] \leq \exp\left(\frac{1}{2}\lambda^2 c^2/4\right)$$

Plugging the above into (6) *proves the inductive hypothesis* (5)*. Applying the inductive hypothesis* (5) *a total of* n *times yields*

$$\mathbf{E}\exp(\lambda(U - \mathbf{E}U)) \leq \exp\left(\frac{1}{2}\lambda^2(nc^2/4)\right),$$

thus proving that $U - \mathbf{E}U$ *is sub-Gaussian with variance proxy* $nc^2/4$*.*

In the following exercise, we demonstrate how the bounded differences inequality can be applied to obtain generalization error guarantees (applicable for any algorithm) and excess risk guarantees (applicable for empirical risk minimizers) that *hold with high probability*—while the lecture only covered an in-expectation guarantee.

Exercise 4 (High-probability generalization and excess risk guarantees)**.**

Let $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$ be i.i.d. random variables. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss function such that for any y, y' we have $|\ell(y, y')| \leq \ell_\infty$. For any function f, let $\mathcal{R}(f) = \mathbf{E}\ell(Y, f(X))$ and $\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i))$. Let $\mathcal{F}$ be an arbitrary class of predictors and let $\mathcal{L} = \{\ell_f : (x, y) \mapsto \ell(y, f(x)) : f \in \mathcal{F}\}$.

Using the bounded differences inequality proved in Exercise 3, show that for any $\delta \in (0, 1)$ the following deviation inequality holds:

$$\mathbf{P}\left(\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} \geq \mathbf{E}\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} + \ell_\infty\sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta. \tag{7}$$

(Hint: $U = \sup_{f\in\mathcal{F}}\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\}$ is a deterministic function of the datapoints $Z_1, ..., Z_n$.) Deduce the following two inequalities:

1. *For any statistical estimator that selects a predictor $\hat{f} = \hat{f}(Z_1, \ldots, Z_n)$ from the class $\mathcal{F}$ it holds, with probability at least $1 - \delta$, that*

$$\mathcal{R}(\hat{f}) \leq \hat{\mathcal{R}}(\hat{f}) + 2\mathrm{Rad}_n(\mathcal{L}) + \ell_\infty\sqrt{2\frac{\log(1/\delta)}{n}}, \tag{8}$$

where recall that

$$\mathrm{Rad}_n(\mathcal{L}) = \mathbf{E}_{Z_1,\ldots,Z_n}\mathbf{E}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i\ell(Y_i, f(X_i))\right].$$

2. *Let $\hat{f}^{(erm)} \in \mathrm{argmin}_{f\in\mathcal{F}}\hat{\mathcal{R}}(f)$ be any empirical risk minimizer among the functions in the class $\mathcal{F}$. Let $f^* \in \mathrm{argmin}_{f\in\mathcal{F}}\mathcal{R}(f)$ (for simplicity, we assume that such f^* exists). Prove that*

$$\mathcal{R}(\hat{f}^{(erm)}) \leq \mathcal{R}(f^*) + 2\mathrm{Rad}_n(\mathcal{L}) + 2\ell_\infty\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{9}$$

Solution 4. *Define*

$$U = F(Z_1, \ldots, Z_n) = \sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\}.$$

We will apply the McDiarmid's inequality to the random variable U defined above. Before doing so, we will prove that f satisfies bounded differences property. Indeed, fix $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$ and let z_i, z_i' be arbitrary. Let $\hat{\mathcal{R}}$ denote the empirical risk function supported on the samples $Z_1, \ldots, Z_{i-1}, z_i, Z_{i+1}$ and let $\hat{\mathcal{R}}'$ denote the empirical risk function supported on the samples $Z_1, \ldots, Z_{i-1}, z_i', Z_{i+1}$. Assume, without loss of generality, the existence of functions $f, f' \in \mathcal{F}$ such that

$$f \in argmax_{f\in\mathcal{F}}\mathcal{R}(f) - \hat{\mathcal{R}}(f) \quad \text{and} \quad f' \in argmax_{f\in\mathcal{F}}\mathcal{R}(f) - \hat{\mathcal{R}}'(f).$$

Then, we have

$$\begin{aligned}
&F(Z_1, \ldots, Z_{i-1}, z_i, Z_{i+1}, \ldots, Z_n) - F(Z_1, \ldots, Z_{i-1}, z_i', Z_{i+1}, \ldots, Z_n) \\
&\leq \hat{\mathcal{R}}(f) - \hat{\mathcal{R}}'(f) \\
&= \frac{1}{n}\left(\ell(y_i, f(x_i)) - \ell(y_i', f(x_i'))\right) \\
&\leq \frac{2}{n}\ell_\infty.
\end{aligned}$$

Similarly, we have

$$F(Z_1,\dots,Z_{i-1},z_i,Z_{i+1},\dots,Z_n)-F(Z_1,\dots,Z_{i-1},z_i',Z_{i+1},\dots,Z_n)\geq -\frac{2}{n}\ell_\infty.$$

Hence, the random variable U satisfies the bounded differences property with $c=\frac{2}{n}\ell_\infty$. Applying McDiarmid's inequality to U (cf. Exercise 3 and Exercise 1 part 3, specifically (1)) yields the desired bound (7).

It remains to do the final two parts of this exercise.

1. *This part follows almost immediately by (7). Indeed, by the inequality (7), for any $\delta\in(0,1)$, with probability at least $1-\delta$ it holds that*

$$\mathcal{R}(\hat{f})-\hat{\mathcal{R}}(\hat{f})\leq\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f)-\hat{\mathcal{R}}(f)\right\}\leq\mathbf{E}\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f)-\hat{\mathcal{R}}(f)\right\}+\ell_\infty\sqrt{2\frac{\log(1/\delta)}{n}}. \tag{10}$$

   We have seen in the lectures that the following deterministic inequality always holds (proved by symmetrizing the empirical process):

$$\mathbf{E}\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f)-\hat{\mathcal{R}}(f)\right\}\leq 2\mathrm{Rad}_n(\mathcal{L}).$$

   Plugging in the above inequality into (10) completes the proof.

2. *In this exercise, we need to exploit the fact that $\hat{f}^{(erm)}$ returns not any function, but a function that minimizes the empirical risk. Indeed, the following deterministic inequality holds*

$$\begin{aligned}&\mathcal{R}(\hat{f}^{(erm)})-\mathcal{R}(f^*)\\&=\underbrace{(\mathcal{R}(\hat{f}^{(erm)})-\hat{\mathcal{R}}(\hat{f}^{(erm)}))}_{\text{uniform convergence}}+\underbrace{(\hat{\mathcal{R}}(\hat{f}^{(erm)})-\hat{\mathcal{R}}(f^*))}_{\leq 0\ \text{(specific to erm)}}+\underbrace{(\hat{\mathcal{R}}(f^*))-\mathcal{R}(f^*))}_{\text{concentration for a fixed function}}\\&\leq\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f)-\hat{\mathcal{R}}(f)\right\}+\hat{\mathcal{R}}(f^*)-\mathcal{R}(f^*).\end{aligned} \tag{11}$$

The random variables $\ell(Y_i,f^*(X_i))$ are contained in the interval $[-\ell_\infty,\ell_\infty]$, hence, by Hoeffding's lemma, the random variable $\ell(Y_i,f^*(X_i))$ is ℓ_∞^2 sub-Gaussian. By Exercise 1 part 4 it follows that $\hat{\mathcal{R}}(f^*)-\mathcal{R}(f^*)$ is ℓ_∞^2/n sub-Gaussian, and hence, we have

$$\mathbf{P}\left(\hat{\mathcal{R}}(f^*)-\mathcal{R}(f^*)\geq\ell_\infty\sqrt{\frac{2\log(2/\delta)}{n}}\right)\leq\delta/2. \tag{12}$$

By the previous parts of this exercise, we also have

$$\mathbf{P}\left(\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f)-\hat{\mathcal{R}}(f)\right\}\geq 2\mathrm{Rad}_n(\mathcal{L})+\ell_\infty\sqrt{\frac{2\log(2/\delta)}{n}}\right)\leq\delta/2. \tag{13}$$

By the union bound over the events (12) and (13), the decomposition (11) yields, with probability at least $1-\delta$:

$$\mathcal{R}(\hat{f}^{(erm)})-\hat{\mathcal{R}}(f^*)\leq 2\mathrm{Rad}_n(\mathcal{L})+2\ell_\infty\sqrt{\frac{2\log(2/\delta)}{n}},$$

which is what we wanted to show.

Exercise 5 (Rademacher complexity of a set of points)**.** *For a subset $S \subset \mathbb{R}^n$, we define the unnormalized Rademacher complexity as*

$$\mathrm{URad}(S) := \mathbf{E}_\varepsilon \sup_{u \in S} \varepsilon^\top u$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is a vector of independent Rademacher random variables (each taking value $+1$ or -1 with probability $1/2$).

1. *What is the link between the definition in class of* $\mathrm{Rad}(\mathcal{F})$ *for a hypothesis class $\mathcal{F}$ and* URad*?*
2. *Compute* $\mathrm{URad}(\{u\})$ *for an arbitrary $u \in \mathbb{R}^n$*
3. *Compute* $\mathrm{URad}(HC)$ *where $HC = \{-1, +1\}^n$ is the unit hypercube*
4. *Give an upper bound on* $\mathrm{URad}(\{\mathbf{1}, -\mathbf{1}\})$*, where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all entries equal to 1.*

Solution 5.

1. *We have* $\mathrm{Rad}(\mathcal{F}) = \frac{1}{n}\mathbf{E}_{x_1,\dots,x_n}\mathrm{URad}(\{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\})$.
2. $\mathbf{E}_\varepsilon \varepsilon^\top u = \sum_i u_i \mathbf{E}_{\varepsilon_i} \varepsilon_i = 0$.
3. *For any realization of the random signs ε, the choice $\varepsilon \in HC$ maximizes the supremum* $\sup_{u \in HC} \varepsilon^\top u$*. Thus,* $\mathrm{URad}(HC) = n$.
4. *By Jensen's inequality, we have* $\mathrm{URad}(\{\mathbf{1}, -\mathbf{1}\}) = \mathbf{E}|\sum_{i=1}^n \varepsilon_i| \le \sqrt{\mathbf{E}|\sum_{i=1}^n \varepsilon_i|^2} = \sqrt{n}$.

In the upper bounds (8) and (9) we pay for the Rademacher complexity of the "loss class" $\mathcal{L}$. Whenever the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is L-Lipschitz in its second argument, that is, whenever, for any $y, y_1, y_2 \in \mathcal{Y}$ it holds that

$$|\ell(y, y_1) - \ell(y, y_2)| \le L|y_1 - y_2|,$$

we can pay, up to factor L, for the complexity of the class of predictors $\mathcal{F}$. Showing how to achieve this is the purpose of the next exercise.

Exercise 6 ($(\star)$ Talagrand's contraction principle)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^n$ be a set of points and let $\phi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz function. Define $\phi \circ V = \{(\phi(v_1), \dots, \phi(v_n))^\top : v \in \mathcal{V}\} \subseteq \mathbb{R}^n$. Prove that*

$$\mathrm{URad}(\phi \circ \mathcal{V}) \le L\mathrm{URad}(\mathcal{V}).$$

Hint: *Observe that*

$$\begin{aligned}\mathrm{URad}(\phi \circ \mathcal{V}) &= \mathbf{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\mathbf{E}_{\varepsilon_1}\left[\sup_{v \in \mathcal{V}}\left\{\varepsilon_1\phi(v_1) + \sum_{i=2}^n \varepsilon_i \phi(v_i)\right\} | \varepsilon_2, \dots, \varepsilon_n\right]\right] \\ &= \mathbf{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\frac{1}{2}\sup_{v \in \mathcal{V}}\left\{\phi(v_1) + \sum_{i=2}^n \varepsilon_i \phi(v_i)\right\} + \frac{1}{2}\sup_{v \in \mathcal{V}}\left\{-\phi(v_1) + \sum_{i=2}^n \varepsilon_i \phi(v_i)\right\}\right].\end{aligned}$$

Using the L-Lipschitzness property of ϕ, deduce that

$$\mathrm{URad}(\phi \circ \mathcal{V}) \le \mathbf{E}_\varepsilon \sup_{v \in \mathcal{V}}\{L\varepsilon_1 v_1 + \sum_{i=2}^n \varepsilon_i \phi(v_i)\}.$$

Solution 6. *Following the hint, we have*

$$\begin{aligned}
\mathrm{URad}(\phi \circ \mathcal{V}) &= \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v\in\mathcal{V}}\left\{\phi(v_1)+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}+\frac{1}{2}\sup_{v\in\mathcal{V}}\left\{-\phi(v_1)+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}\right] \\
&= \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v\in\mathcal{V}}\left\{\phi(v_1)+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}+\frac{1}{2}\sup_{u\in\mathcal{V}}\left\{-\phi(u_1)+\sum_{i=2}^{n}\varepsilon_i\phi(u_i)\right\}\right] \\
&= \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v,u\in\mathcal{V}}\left\{\phi(v_1)-\phi(u_1)+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)+\sum_{i=2}^{n}\varepsilon_i\phi(u_i)\right\}\right] \\
&\leq \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v,u\in\mathcal{V}}\left\{L|v_1-u_1|+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)+\sum_{i=2}^{n}\varepsilon_i\phi(u_i)\right\}\right],
\end{aligned}$$

where the last step follows via the Lipschitz property of the function ϕ. Next, we observe that $L|v_1-u_1|$ may be replaced by $L(v_1-u_1)$ in the above supremum because we are allowed to interchange the roles of u and v. Hence,

$$\begin{aligned}
\mathrm{URad}(\phi \circ \mathcal{V}) &\leq \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v,u\in\mathcal{V}}\left\{L|v_1-u_1|+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)+\sum_{i=2}^{n}\varepsilon_i\phi(u_i)\right\}\right] \\
&= \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v,u\in\mathcal{V}}\left\{Lv_1-Lu_1+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)+\sum_{i=2}^{n}\varepsilon_i\phi(u_i)\right\}\right] \\
&= \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\left[\frac{1}{2}\sup_{v\in\mathcal{V}}\left\{Lv_1+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}+\frac{1}{2}\sup_{v\in\mathcal{V}}\left\{-Lu_1+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}\right] \\
&= \mathbf{E}_{\varepsilon_2,\ldots,\varepsilon_n}\mathbf{E}_{\varepsilon_1}\left[\sup_{v\in\mathcal{V}}\left\{L\varepsilon_1 v_1+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}|\varepsilon_2,\ldots,\varepsilon_n\right] \\
&= \mathbf{E}_{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n}\sup_{v\in\mathcal{V}}\left\{L\varepsilon_1 v_1+\sum_{i=2}^{n}\varepsilon_i\phi(v_i)\right\}.
\end{aligned}$$

Repeating the same argument for the variables with index $i=2,\ldots,n$ yields

$$\mathrm{URad}(\phi \circ \mathcal{V}) \leq L\mathrm{URad}(\mathcal{V}),$$

which finishes our proof.