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*

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

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

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

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*

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

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).*

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 $i \in \{1, \dots, n\}$, and any $z_1, \dots, z_n, z_i' \in \mathcal{Z}$, we have*

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

Let $Z_1, \dots, Z_n$ be independent (and not necessarily identically distributed) random variables on $\mathcal{Z}$. Then, the random variable $U = f(Z_1, \dots, Z_n)$ 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 $U_i = \mathbf{E}_{Z_{i+1}', \dots, Z_n'} f(Z_1, ..., Z_i, Z_{i+1}', ..., Z_n')$. 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 $Z_1, \dots, Z_{n-1}$.

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 $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ 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 $|\ell(y, y')| \le \ell_\infty$. 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:

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

(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 $\hat{f} = \hat{f}(Z_1, \dots, Z_n)$ from the class $\mathcal{F}$ it holds, with probability at least $1 - \delta$, that*

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

where recall that

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

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*

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

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*.*

In the upper bounds (1) and (2) 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*

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