Homework Week 1

Mathematics of deep learning
MASH & IASD 2025

Lecturer: Bruno Loureiro, bruno.loureiro@di.ens.fr

Instructions: This homework is **due on Monday 20/01/2025**. Please send your solutions in a PDF file named HW1_Nom_Prenom.pdf to the above address with the subject "*[MATHSDL2025] Homework 1*". Formats accepted: LaTeX or a **readable** scan of hand-written solutions.

Exercise 1. Concentration inequalities

(a) *(Markov's inequality)* Let $X \geq 0$ denote a non-negative random variable. Show thar, for any $t > 0$:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \tag{1}$$

(b) *(Chernoff's bound)* Let $X \geq 0$ be a real random variable. Using Markov's inequality, show that for all $C \in \mathbb{R}$ and $t > 0$:

$$\mathbb{P}(X \geq c) \leq \mathbb{E}\left[e^{tX}\right] e^{-ct} \tag{2}$$

Give an example of a probability distribution which has exponential tails.

(c) *(Hoeffding's inequality)* Let $X_1, \dots, X_n$ denote n i.i.d. bounded random variables such that $\mathbb{E}[X_i] = 0$ and $|X| \leq C$. Using Chernoff's inequality and Hoeffding's lemma 1, show that for all $t > 0$:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq e^{-\frac{t^2}{2nC^2}} \tag{3}$$

Give an example of a probability distribution that has doubly exponential tails. How is this result related to the CLT?

Lemma 1 (Hoeffding's lemma)**.** Let $X \in [a, b]$ be a bounded random variable. Then, for all $t > 0$:

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq e^{\frac{t^2(a-b)^2}{8}} \tag{4}$$

Exercise 2.

Consider a supervised learning problem with training data $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$ that we assume is sampled i.i.d. from a distribution p. Let $\mathcal{H} = \{f_\theta : \mathcal{X} \to \mathcal{Y} : \theta \in \Theta\}$ denote a parametric hypothesis class, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ a loss function, which we assume is uniformly bounded by a constant $B > 0$.

(a) Which loss functions we discussed in class satisfy this assumption and which do not?

(b) Write down the definition of the population $R(\theta)$ and empirical $\hat{R}(\theta; \mathcal{D})$ risks.

(c) Using Hoeffding's inequality in eq. (3), show that for any fixed $\theta \in \Theta$ and all $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$|R(\theta) - \hat{R}(\theta; \mathcal{D})| \leq B\sqrt{\frac{\log 1/\delta}{2n}} \tag{5}$$

(d) What are the consequences of this upper bound on the number of samples required in order to achieve a small generalisation gap $\epsilon > 0$?