RWTH Aachen University **Winter 2023/24**

Instructor: Prof. Dr. Semih Çaycı
Teaching Assistant: Johannes Müller, M.Sc.

Mathematical Foundations of Deep Learning (11.80020)
Assignment 1

Due: Tuesday, Nov. 7th, till the beginning of class at 2pm via Moodle upload

Each problem is worth 4 points, there are 20 points on this sheet. Submission in pairs is possible.

Q1. (Union bound)

(a) Show that for arbitrary events (i.e., measurable sets) $A_1, A_2, \dots$ it holds that

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

(b) Use this to show that for a sequence of real random variables $X_1, \dots, X_n$ it holds that

$$\mathbb{P}\left(\max_{i=1,\dots,n} X_i > t\right) \leq \sum_{i=1}^{n} \mathbb{P}(X_i > t).$$

(c) Consider real σ^2-sub-Gaussian centered random variables $X_1, \dots, X_n$. Show that

$$\mathbb{P}\left(\max_{i=1,\dots,n} X_i > t\right) \leq n e^{-\frac{t^2}{2\sigma^2}}. \tag{1}$$

(d) Consider a bounded loss $\ell\colon \mathbb{Y} \times \mathbb{Y} \to [-B, B]$ for some $B \in \mathbb{R}_{\geq 0}$. Show that finite hypothesis classes are PAC-learnable with

$$n_0(\varepsilon, \delta) \leq \frac{8 \cdot B^2 \log(2|\mathcal{H}|/\delta)}{\varepsilon^2}.$$

Q2. (A maximal inequality) Consider σ^2-sub-Gaussian centered random variables $X_1, \dots, X_n$. Show that

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

and that

$$\mathbb{P}\left(\max_{i=1,\dots,n} X_i \geq \sigma(\sqrt{2\log n} + t)\right) \leq e^{-t\sqrt{2\log n} - \frac{t^2}{2}} \quad \text{for all } t \geq 0.$$

Hint: Consider $e^{\lambda\mathbb{E}[\max_i X_i]}$ and use Jensen's inequality. The tail bound (1) can be used.

Q3. (Tail bounds for a Gaussian random variable) Consider a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2

(a) Compute the centered logarithmic moment generating function $\widetilde{\varphi}_X$ of X.

(b) Use this to compute the centered moments $m_k := \mathbb{E}[(X - \mathbb{E}X)^k]$.

(c) Show that

$$\mathbb{P}(X - \mathbb{E}X > t) \leq e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \geq 0.$$

Q4. (Hoeffding vs Chernoff for Bernoulli variables) Consider a sequence of independent and identically Bernoulli variables $X_1, \ldots, X_n \in \{0, 1\}$ with parameter $p \in [0, 1]$, i.e., $\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0)$.

(a) Show that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p > t\right) \leq e^{-2nt^2} \quad \text{for } t \geq 0. \tag{2}$$

(b) Show that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p > t\right) \leq e^{-nD(p+t||p)} \quad \text{for } t \geq 0, \tag{3}$$

where

$$D(x||y) := x \log\left(\frac{x}{y}\right) + (1 - x) \log\left(\frac{1 - x}{1 - y}\right)$$

is the Kullback-Leibler-divergence.

(c) Show that (3) is tighter as (2). Are there choices of p, for which the two bounds agree?

Q5. (k-bit Perceptron) **Terminology:** We say that $m \in \mathbb{N}$ is a k-bit integer for $k \in \mathbb{N}$ if $m = \sum_{i=0}^{k-1} a_i 2^i$ for some $a_i \in \{0, 1\}$. We call a function $f\colon \mathbb{R}^d \to \{\pm 1\}$ a *k-bit perceptron* if

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{d} w_i x_i - b\right)$$

for some k-bit integers $w_1, \ldots, w_n, b \in \mathbb{N}$ and where

$$\operatorname{sgn}(x) := \begin{cases} 1 & \text{if } x \geq 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Problem: Let $S \subseteq \mathbb{R}^d \times \{0, 1\}$ denote a training set of n iid samples and consider the hypothesis class

$$\mathcal{H}_k = \left\{ f\colon \mathbb{R}^d \to \mathbb{R} : f \text{ is a } k\text{-bit perceptron} \right\}.$$

Let $\hat{f}_{\mathcal{H}_k}$ denote the empirical risk minimizer over $\mathcal{H}_k$ with respect to the sample loss $\ell(\hat{y}, y) = \mathbb{1}\{\hat{y} \neq y\}$ and denote the population risk by $\mathcal{R}$. Show that for any $\varepsilon, \delta \in (0, 1)$ it holds that

$$\mathbb{P}\left(\mathcal{R}(\hat{f}_{\mathcal{H}_k}) < \min_{f \in \mathcal{H}_k} \mathcal{R}(f) + \varepsilon\right) \geq 1 - \delta$$

whenever

$$n \geq \frac{2}{\varepsilon^2}\left(k(d + 1)\log 2 + \log\left(\frac{2}{\delta}\right)\right).$$

Note: The following are bonus problems worth 4 points per problem.

Q6. (Bonus problem: Moment vs Chernoff bounds) Suppose that $X \geq 0$, and that the moment generating function of X exists in an interval around zero. Given some $t > 0$ and an integer $k = 1, 2, \ldots$, show that

$$\inf_{k=0,1,2,\ldots} \frac{\mathbb{E}[|X|^k]}{t^k} \leq \inf_{\lambda>0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}.$$

Use this to derive a tail bound for X based on moments that improves Chernoff's bound.

Q7. (Bonus problem: Infinite hypothesis classes can be PAC-learnable) Consider a classification problem with $\mathbb{X} = \mathbb{R}^2$ and $\mathbb{Y} = \{0, 1\}$, and let $\mathcal{H} = \{h_r : r \in \mathbb{R}_{>0}\}$ be the hypothesis class, where $h_r(x) = \mathbb{1}_{\{\|x\|_2 \leq r\}}$ for $x \in \mathbb{X}$ and $r > 0$ and the 0-1 loss $\ell(\hat{y}, y) = \mathbb{1}_{\{\hat{y} \neq y\}}$. We call the problem *realizable in* $\mathcal{H}$ if $\mathcal{R}(h^*) = 0$ for some $h^* \in \mathcal{H}$. Prove that $\mathcal{H}$ is PAC-learnable assuming that the problem is realizable in $\mathcal{H}$ with sample complexity $n_0(\epsilon, \delta) \leq \lceil \log(1/\delta)/\epsilon \rceil$, i.e., show that there is learning algorithm $A = (A_n)_{n \in \mathbb{N}}$ such that

$$\mathbb{P}(\mathcal{R}(A_n(S_n)) \leq \varepsilon) \geq 1 - \delta \quad \text{for all } n \geq \lceil \log(1/\delta)/\epsilon \rceil. \tag{4}$$

Hint: For a given training set $S = \{(x_i, y_i) \in \mathbb{X} \times \mathbb{Y} : i = 1, 2, \ldots, n\}$, consider a prediction rule with the smallest circle containing all training points with label 1 as the decision boundary. Is this prediction rule an empirical risk minimizer?