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

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

Solution: Recall that for $A \subseteq B$ we have $\mathbb{P}(A) \subseteq \mathbb{P}(B)$ and thus by the σ additivity of measures we have

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n \setminus \left(\cup_{i=1}^{n-1} A_i\right)\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n \setminus (\cup_{i=1}^{n-1} A_i)) \le \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

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

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

Solution: Using part (a) we estimate

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

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

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

Solution: By Hoeffding's inequality it holds that $\mathbb{P}(X_i > t) \le e^{-\frac{t^2}{2\sigma^2}}$ which in combination with (b) yields (1).

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

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

Solution: Let $\hat{f}_S$ denote the empirical risk minimizer over $\mathcal{H}$ with respect to the training set S and let $f^*_{\mathcal{H}}$ denote the minimizer of the population risk $\mathcal{R}$ over $\mathcal{H}$. Then the excess risk of the ERM can be bounded by

$$\mathcal{R}(\hat{f}_S)-\mathcal{R}^* = \mathcal{R}(\hat{f}_S)-\hat{\mathcal{R}}_S(\hat{f}_S)+\hat{\mathcal{R}}_S(\hat{f}_S)-\hat{\mathcal{R}}_S(f^*_{\mathcal{H}})+\hat{\mathcal{R}}_S(f^*_{\mathcal{H}})-\mathcal{R}(f^*_{\mathcal{H}}) \leq 2\max_{f\in\mathcal{H}}|\hat{\mathcal{R}}_S(f)-\mathcal{R}(f)|$$

and hence we aim to estimate the tails of the latter. Recall that by Hoeffding's inequality we have

$$\mathbb{P}(|\hat{\mathcal{R}}_S(f)-\mathcal{R}(f)|>\varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{2B^2}}$$

for any $f \in \mathcal{H}$. Using the union bound we can estimate

$$\begin{aligned}\mathbb{P}(\mathcal{R}(\hat{f}_S)-\mathcal{R}^*>\varepsilon) &\leq \mathbb{P}\left(\max_{f\in\mathcal{H}}|\hat{\mathcal{R}}_S(f)-\mathcal{R}(f)|>\frac{\varepsilon}{2}\right)\\ &\leq 2\cdot|\mathcal{H}|\cdot e^{-\frac{n\varepsilon^2}{2B^2}}.\end{aligned}$$

Solving $\delta = e^{-\frac{n\varepsilon^2}{2B^2}}$ for n yields $n_0(\varepsilon,\delta) \leq \frac{2B^2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2}$.
Remark: Note that we have used the independence of the samples here. If we don't use the independence we still have by Hoeffding's lemma that if $\mathbb{E}_S[\hat{\mathcal{R}}_S(f)] = \mathcal{R}(f)$ then

$$\mathbb{P}(|\hat{\mathcal{R}}_S(f)-\mathcal{R}(f))|>\varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{8B^2}}.$$

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.
Solution: For $\lambda \geq 0$ Jensen's inequality yields

$$e^{\lambda\mathbb{E}[\max_i X_i]} \leq \mathbb{E}[e^{\lambda\max_i X_i}] \leq \mathbb{E}\left[\sum_i e^{\lambda X_i}\right] \leq ne^{\frac{\sigma^2\lambda^2}{2}}$$

and hence

$$\mathbb{E}[\max_i X_i] \leq \lambda^{-1}\left(\log n+\frac{\lambda^2\sigma^2}{2}\right) = \frac{\log n}{\lambda}+\frac{\sigma^2\lambda}{2}.$$

Optimizing over λ (or simply setting $\lambda = \frac{\sqrt{2\log n}}{\sigma}$) yields

$$\mathbb{E}[\max_i X_i] \leq \sigma\sqrt{2\log n}.$$

Using (1) we estimate

$$\begin{aligned}\mathbb{P}\left(\max_{i=1,\dots,n} X_i \geq \sigma(\sqrt{2\log n}+t)\right) &\leq n\exp\left(-\frac{1}{2\sigma^2}\cdot\sigma^2(\sqrt{2\log n}+t)^2\right)\\ &= n\exp\left(-\log n - t\sqrt{2\log n}-\frac{t^2}{2}\right)\\ &= \exp\left(-t\sqrt{2\log n}-\frac{t^2}{2}\right).\end{aligned}$$

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.

Solution: We can assume without loss of generality that $\mu = 0$ and compute

$$\begin{aligned} e^{\widetilde{\varphi}_X(\lambda)} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{\lambda x} e^{-\frac{x^2}{2\sigma^2}} \mathrm{d}x \\ &= e^{\frac{\sigma^2\lambda^2}{2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{-\frac{x^2 - 2\sigma^2\lambda x + \sigma^4\lambda^2}{2\sigma^2}} \mathrm{d}x \\ &= e^{\frac{\sigma^2\lambda^2}{2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{-\frac{(x-\sigma^2\lambda))^2}{2\sigma^2}} \mathrm{d}x \\ &= e^{\frac{\sigma^2\lambda^2}{2}}. \end{aligned}$$

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

Solution: Denoting the centered moment generating function $\widetilde{M}_X(\lambda) = e^{\widetilde{\varphi}_X}$ we have that $m_k = \widetilde{M}_X^{(k)}(0)$, where $\widetilde{M}_X^{(k)}$ denotes the k-th derivative of $\widetilde{M}_X$. We find that

$$\widetilde{M}_X^{(k)}(0) = \partial_\lambda^k e^{\frac{\sigma^2\lambda^2}{2}}|_{\lambda=0} = \partial_\lambda^k \sum_{n\in\mathbb{N}} \frac{\sigma^{2n}\lambda^{2n}}{2^n n!}|_{\lambda=0} = \begin{cases} \sigma^k \prod_{l=0}^{k/2-1}(k-2l-1) & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd.} \end{cases}$$

(c) Show that

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

Solution: By (a) a Gaussian random $X \sim \mathcal{N}(\mu, \sigma^2)$ variable is sub-Gaussian with parameter σ^2 and hence Chernoff's bound for sub-Gaussian random variables yields the claim.

Q4. (Hoeffding vs Chernoff for Bernoulli variables) Consider a sequence of independent and identically distributed 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

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

Solution: This is a direct consequence of Hoeffing's inequality.

(b) Show that

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

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.

Solution: Using Chernoff's bound we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p > t\right) \leq e^{-\widetilde{\varphi}_X^*(nt)}$$

for $X \coloneqq \sum_{i=1}^n X_i$. By the independence of X_i we have $\widetilde{\varphi}_X = \sum_{i=1}^n \widetilde{\varphi}_{X_i} = n\widetilde{\varphi}_{X_i}$, where in the last step we used that the variables X_i are identically distributed. Note that we have

$$\widetilde{\varphi}_X^*(nt) = \sup_{\lambda\in\mathbb{R}} \lambda nt - \widetilde{\varphi}_X(\lambda) = n\sup_{\lambda\in\mathbb{R}} \lambda t - \widetilde{\varphi}_{X_i}(\lambda) = n\widetilde{\varphi}_{X_i}^*(t).$$

Hence, it remains to show that

$$\widetilde{\varphi}_{X_i}^*(t) = D(p+t||p) = (p+t)\log\left(\frac{p+t}{p}\right) + (1-p-t)\log\left(\frac{1-p-t}{1-p}\right).$$

We compute

$$\widetilde{\varphi}_{X_i}(\lambda) = \log\mathbb{E}[e^{\lambda(X_i-p)}] = \log\left((1-p)e^{-\lambda p} + pe^{\lambda(1-p)}\right) = -\lambda p + \log(1-p+pe^{\lambda}).$$

In order to compute $\varphi_{X_i}^*(t)$ we solve

$$t = \partial_\lambda \widetilde{\varphi}_{X_i}(\lambda) = -p + \frac{pe^\lambda}{1-p+pe^\lambda}$$

for λ. This yields

$$(p+t)(1-p) = (p-(p+t)p)e^\lambda = p(1-p-t)e^\lambda$$

and consequently

$$\lambda^* = \log\left(\frac{(p+t)(1-p)}{p(1-p-t)}\right) = \log\left(\frac{p+t}{p}\right) + \log\left(\frac{1-p}{1-p-t}\right).$$

Inserting yields

$$\begin{aligned}
\widetilde{\varphi}_{X_i}^*(t) &= \lambda^* t + p\lambda^* - \log(1-p+pe^{\lambda^*}) \\
&= (p+t)\log\left(\frac{p+t}{p}\right) - (p+t)\log\left(\frac{1-p-t}{1-p}\right) - \log\left(1-p+p\cdot\frac{(1-p)(p+t)}{p(1-p-t)}\right) \\
&= (p+t)\log\left(\frac{p+t}{p}\right) - (p+t)\log\left(\frac{1-p-t}{1-p}\right) - \log\left(\frac{1-p}{1-p-t}\right) \\
&= (p+t)\log\left(\frac{p+t}{p}\right) + (1-p-t)\log\left(\frac{1-p-t}{1-p}\right).
\end{aligned}$$

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

$$\begin{aligned}
\partial_t^2 D(p+t||p) &= \partial_t^2\left((p+t)\log\left(\frac{p+t}{p}\right) + (1-p-t)\log\left(\frac{1-p-t}{1-p}\right)\right) \\
&= \partial_t\left(1+\log\left(\frac{p+t}{p}\right) + 1 + \log\left(\frac{1-p-t}{1-p}\right)\right) \\
&= \frac{1}{p+t} + \frac{1}{1-p-t} = \frac{1}{(p+t)(1-p-t)} \geq 4 = \partial_t^2(2t^2).
\end{aligned}$$

Since $D(p+0||p) = 0 = 2 \cdot 0^2$ this implies $D(p+t||p) \geq 2t^2$ and consequently shows that the Chernoff bound is tighter. Note that $\partial_t^2 D(p+t||p) = 4$ if and only if $p+t = 1/2$. This shows that if $p = 1/2$ we have $D(p+t||p) = 2t^2 + O(t^3)$ for $t \to \mathbf{0}$. However, this also shows that $D(p+t||p) > 2t^2$ for all $t > 0$.

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

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

Solution: We want to apply the general PAC-learnability result for finite hypothesis classes and hence compute the cardinality $|\mathcal{H}_k| = 2^{(d+1)k}$ of the set of k-bit perceptrons. Now the statements follows directly from Q1 (d).

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 $t > 0$, show that

$$\inf_{k=0,1,2,\ldots} \frac{\mathbb{E}[X^k]}{t^k} \leq \inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}. \tag{4}$$

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

Solution: We set

$$c := \inf_{k=0,1,2} \frac{\mathbb{E}[X^k]}{t^k}$$

then in particular $\mathbb{E}[X^k] \geq ct^k$. Now we estimate

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[\sum_{k \in \mathbb{N}} \frac{\lambda^k X^k}{k!}\right] = \sum_{k \in \mathbb{N}} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \geq c \sum_{k \in \mathbb{N}} \frac{\lambda^k t^k}{k!} = e^{\lambda t}.$$

Devining by $e^{\lambda t}$ and taking the infimum over λ yields (4).

For $t \geq 0$ we can use Markov's inequality to estimate

$$\mathbb{P}(X > t) = \mathbb{P}(X^k > t^k) \leq \frac{\mathbb{E}[X^k]}{t^k}.$$

Taking the infimum over k this yields

$$\mathbb{P}(X > t) \leq \inf_{k=0,1,2,\dots} \frac{\mathbb{E}[X^k]}{t^k}$$

which is an improvement of Chernoff's bound by (4).

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

$$\mathbb{P}(\mathcal{R}(A_n(S_n)) \leq \varepsilon) \geq 1 - \delta \quad \text{for all } n \geq \lceil \log(1/\delta)/\epsilon \rceil. \tag{5}$$

Hint: For a given training set $S = \{(x_i, y_i) \in \mathbb{X} \times \mathbb{Y} : i = 1, 2, \dots, n\}$, 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?

Solution: Let us denote the realizing hypothesis by $h^* = h_{r^*}$. Consider the learning algorithm $S \mapsto h_{r_S}$, where

$$r = r_S := \max\{\|x_i\| : y_i = 1, i = 1, \dots, n\}.$$

Note that by the realizability assumption h_{r_S} achieves zero empirical risk and hence is an empirical risk minimizer (note that the ERM is not unique in this case). Further, by the realizability assumption it holds that

$$\mathcal{R}(h_{r_S}) = P(r_S < \|x\| \leq r^*) = P(\overline{B_{r^*}} \setminus \overline{B_{r_S}}),$$

where $B_r = \{x : \|x\| < r\}$. Set

$$r_\varepsilon := \sup\{r > 0 : \mathbb{P}(\overline{B_{r^*}} \setminus B_r) > \varepsilon\},$$

then $P(\overline{B_{r^*}} \setminus \overline{B_{r_\varepsilon}}) \leq \varepsilon$ and $P(\overline{B_{r^*}} \setminus B_{r_\varepsilon}) \geq \varepsilon$ and thus $P(B_{r_\varepsilon}) \leq 1 - \varepsilon$. In particular, $\mathcal{R}(h_{r_S}) = P(\overline{B_{r^*}} \setminus \overline{B_{r_S}}) > \varepsilon$ implies $r_S < r_\varepsilon$. Note that

$$\mathbb{P}(r_S < r) = \prod_{i=1}^{n} P(x_i < r) = P(B_r)^n.$$

Now, we have

$$\mathbb{P}(\mathcal{R}(h_{r_S}) > \varepsilon) \leq \mathbb{P}(r_S \leq r_\varepsilon) \leq P(B_{r_\varepsilon})^n \leq (1 - \varepsilon)^n \leq e^{-n\varepsilon}.$$

Solving $\delta = e^{-n\varepsilon}$ for n yields the claim.