**Algorithmic Foundations of Learning**

# Problem Sheet 1

*Lecturer: Patrick Rebeschini* *Due date: 48 hours before class*

## 1.1 Slow Rate $1/\sqrt{n}$ (Question type: A)

1. Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mu = \mathbf{E}X_1$ and $\sigma^2 = \mathbf{Var}X_1$. Show that

$$\sqrt{\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right]} = \frac{\sigma}{\sqrt{n}}. \tag{1.1}$$

2. Let $Z$ be a real-valued random variable such that for any $u \geq 0$ we have $\mathbf{P}(|Z| \geq u) \leq 2e^{-\frac{u^2}{2\sigma^2}}$, for a positive constant $\sigma^2 > 0$ (this statement holds if $Z$ is Gaussian with mean 0 and variance $\sigma^2$, as we will prove later on in the course). Prove that for all $p \geq 1$ we have

$$(\mathbf{E}[|Z|^p])^{1/p} \leq 2\sqrt{\sigma^2}\sqrt{p}.$$

   *[Hint: Recall that for any non-negative random variable $X$ we have $\mathbf{E}X = \int_0^\infty \mathbf{P}(X > \varepsilon)\mathrm{d}\varepsilon$. Recall also that the Gamma function is defined as $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x$ for any $t > 0$. You might want to use that $\Gamma(t) \leq t^t$ for any $t > 0$ and that $p^{1/p} \leq 2$ for any $p > 0$]*

3. Let $X_1, X_2, \ldots$ be i.i.d. Gaussian random variables with mean $\mu$ and variance $\sigma^2$. Show that for any $p \geq 2$ we have

$$\left(\mathbf{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|^p\right]\right)^{1/p} \leq \frac{c_p}{\sqrt{n}}, \tag{1.2}$$

   where $c_p$ is a quantity dependent on $p$ (but independent of $n$) that you should state.

Along with the Central Limit Theorem, which is an asymptotic statement, the non-asymptotic statements (1.1) and (1.2) illustrate how the "slow" rate $1/\sqrt{n}$ permeates statistics. In machine learning, we will see that there are settings (essentially exploiting low noise or some convexity properties) where we can achieve the "fast" rate $1/n$, or be somewhere in between: $1/n^\alpha$ with $\alpha \in [1/2, 1]$.

It turns out that the inequality (1.2) holds not just for Gaussian random variables, but for a much bigger class of random variables called *sub-Gaussian*, which can be *defined* by the tail inequality $\mathbf{P}(|Z| \geq u) \leq 2e^{-\frac{u^2}{2\sigma^2}}$. Here, $\sigma^2$ is called variance proxy. Sub-Gaussian random variables will be properly introduced later on in the course and will play a crucial role throughout.

**Remark 1.1 (Monte Carlo)** *While the rate $1/\sqrt{n}$ is "slow", it is the very reason why Monte Carlo approximation methods are used to approximate integrals of the form $\int f(x)p(x)\mathrm{d}x$ for a given high-dimensional function $f : \mathbb{R}^d \to \mathbb{R}$ and density $p$. In fact, note that by the same arguments above, if we assume to have access to i.i.d. samples $X_1, \ldots, X_n$ from the density $p$, then the Monte Carlo estimate $\frac{1}{n}\sum_{i=1}^n f(X_i)$ yields*

$$\left(\mathbf{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int f(x)p(x)\mathrm{d}x\right|^p\right]\right)^{1/p} \leq \frac{c_p}{\sqrt{n}}.$$

*In this context, the Monte Carlo estimate is a good one as the rate* $1/\sqrt{n}$ *is* independent *of the dimension* $d$*. Note that deterministic "grid" methods to approximate the integral would give a dimension-dependent rate* $1/n^{1/d}$*, which is strictly worse than the Monte Carlo rate for* $d \geq 3$ *and yields the so-called curse of dimensionality: to have precision* $\varepsilon$ *one would need the number of samples* $n$ *to be of the order* $(1/\varepsilon)^d$*, which scales exponentially with* $d$*. Using Monte Carlo, one only needs* $(1/\varepsilon)^2$ *samples. In fact, there is active research on developing "quasi-Monte Carlo" methods to get rates close to the "fast" rate* $1/n$*.*

## 1.2 Bayes Decision Rules (Question type: A)

Consider the setting of Section 1.2 in the Lectures Notes. Show that the following different choices of loss functions yield the corresponding Bayes decision rule.

1. In regression, the choice $\phi(\hat{y}, y) = (\hat{y} - y)^2$ leads to $a^{\star\star}(x) = \mathbf{E}[Y|X = x]$.

   Hint: Prove that $\mathbf{E}\,\phi(a^{\star\star}(X), Y) \leq \mathbf{E}\,\phi(a(X), Y)$ for any $a \in \mathcal{B}$ by considering $\mathbf{E}[(a^{\star\star}(X) - a(X) + a(X) - Y)^2]$.

2. In classification, the choice $\phi(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y}$ leads to $a^{\star\star}(x) = \operatorname{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{P}(Y = \hat{y}|X = x)$.

## 1.3 Statements in Probability (Question type: B)

Let $\mathbf{P}(X < \varepsilon_1) \geq 1 - \delta_1$ and $\mathbf{P}(Y < \varepsilon_2) \geq 1 - \delta_2$. Show that $\mathbf{P}(X + Y < \varepsilon_1 + \varepsilon_2) \geq 1 - (\delta_1 + \delta_2)$.

## 1.4 Excess Risk, Prediction Error and Estimation Error with the Square Loss (Question type: B)

Consider a supervised learning setting where $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is a $d$-dimensional feature vector and $Y \in \mathbb{R}$ is its response. For any non-random predictor $a : \mathbb{R}^d \to \mathbb{R}$, consider the expected risk:

$$r(a) = \mathbf{E}[(a(X) - Y)^2].$$

1. Let $a^{\star\star} \in \operatorname{argmin} r(a)$. Show that the excess risk $r(a) - r(a^{\star\star})$ can be written as

$$\underbrace{r(a) - r(a^{\star\star})}_{\text{excess risk}} = \underbrace{\mathbf{E}[(a(X) - a^{\star\star}(X))^2]}_{\text{prediction error}}.$$

2. For any (possibly random) predictor $A : \mathbb{R}^d \to \mathbb{R}$, prove the following bias-variance decomposition:

$$\underbrace{\mathbf{E}\,r(A) - r(a^{\star\star})}_{\text{expected excess risk}} = \underbrace{\mathbf{E}\Big[\Big(\mathbf{E}[A(X)|X] - a^{\star\star}(X)\Big)^2\Big]}_{\text{expected squared bias}} + \underbrace{\mathbf{E}\,\mathbf{Var}[A(X)|X]}_{\text{expected variance}}.$$

Henceforth, consider the class of linear predictors given by $\mathcal{A} = \{a : a(x) = \langle x, w_a \rangle \text{ for some } w_a \in \mathbb{R}^d\}$, where $\langle x, w_a \rangle = x^\top w_a$ denotes the inner product between $x$ and $w_a$. Assume $Y = a^{\star\star}(X) + \xi$, where $a^{\star\star}(X) = \langle X, w_{a^{\star\star}} \rangle$ and $\xi$ is an independent random variable with mean 0.

3. Show that the irreducible risk is $r(a^{\star\star}) = \mathbf{Var}(\xi)$.
4. Prove that for any $a \in \mathcal{A}$ defined as $a(x) = \langle x, w_a \rangle$ the following equivalence holds:

$$r(a) - r(a^{\star\star}) = (w_a - w_{a^{\star\star}})^\top \mathbf{m}(w_a - w_{a^{\star\star}}),$$

   where $\mathbf{m}$ is a matrix that you should specify.
5. What does the previous result imply when $X$ is isotropic, i.e., $\mathbf{E}[X_i X_j] = \mathbf{1}_{i=j}$?

## 1.5 Examples of Rademacher Complexity (Question type: B)

Compute the Rademacher complexity $\texttt{Rad}(\mathcal{T})$ of the following sets.

1. Let $\mathcal{T} = \{t\}$ for a given $t \in \mathbb{R}^n$.
2. Let $\mathcal{T} = \{(1,3), (-2,3)\} \subseteq \mathbb{R}^2$.
3. Let $\mathcal{T}$ be the subset of $\{-1, 0, 1\}^n$ containing all elements with $k$-sparse components, i.e., with exactly $k$ components different than 0.
4. For $c > 0$, let $\mathcal{T}$ be the subset of $[-c, c]^n$ containing all elements with $k$-sparse components, i.e., with exactly $k$ components different than 0.

## 1.6 Rademacher Complexity of a Finite Set (Question type: B)

Prove Massart's lemma, Lemma 2.9 in the Lecture Notes.

Hint: Follow the proof strategy of Proposition 2.2 to prove $\texttt{Rad}(\mathcal{T}) \le \max_{t\in\mathcal{T}} \|t\|_2 \sqrt{2\log|\mathcal{T}|}/n$ and use the properties of the Rademacher complexity under scalar translations, Proposition 2.6.

## 1.7 Contraction Property of Rademacher Complexity (Question type: C)

Prove Talagrand's lemma, Lemma 2.10 in the Lecture Notes.

Hint:

1. First, prove that for any set $\mathcal{T} \subseteq \mathbb{R}^2$ and for any 1-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ we have

$$\sup_{t\in\mathcal{T}}(t_1 + f(t_2)) + \sup_{t\in\mathcal{T}}(t_1 - f(t_2)) \le \sup_{t\in\mathcal{T}}(t_1 + t_2) + \sup_{t\in\mathcal{T}}(t_1 - t_2).$$

2. Second, for any $k \in \{1, \dots, n\}$, by conditioning on $\{\Omega_1, \dots, \Omega_{k-1}, \Omega_{k+1}, \dots, \Omega_n\}$ prove that

$$\mathbf{E}\sup_{t\in\mathcal{T}}\sum_{i=1}^n \Omega_i f_i(t_i) \le \mathbf{E}\sup_{t\in\mathcal{T}}\left(\sum_{i\neq k}\Omega_i f_i(t_i) + \Omega_k t_k\right).$$