# STATISTICAL MACHINE LEARNING (WS2022/23)
# SEMINAR 2

**Assignment 1.** We are given a prediction strategy $h\colon \mathcal{X} \to \mathcal{Y} = \{1,\dots,Y\}$ assigning observations $x \in \mathcal{X}$ into one of $Y$ classes. Our task is to estimate the true risk $R(h) = \mathbb{E}_{(x,y)\sim p}\ell(y, h(x))$ where $\ell\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ is some application specific loss function. To this end, we collect a set of examples $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X}\times\mathcal{Y}) \mid i = 1,\dots,l\}$ drawn i.i.d. from the distribution $p(x,y)$ and compute the test risk

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i)) \;.$$

We want to construct the confidence interval such that

$$R(h) \in (R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon) \quad \text{holds with probability } \; \gamma \in (0,1) \tag{1}$$

The number of examples $l$, the precision parameter $\varepsilon$ and the confidence level $\gamma$ are three interdependent variables, i.e., fixing two of the variables allows to compute the third one.

**a)** Use the Hoeffding's inequality to derive a formula to compute $\varepsilon$ as a function of $l$ and $\gamma$ such that (1) holds.

**b)** Use the Hoeffding's inequality to derive a formula to compute $l$ as a function of $\varepsilon$ and $\gamma$ such that (1) holds.

**c)** Instantiate the formulas derived in a) and b) for the following loss functions:
(1) $\ell(y,y') = [\![y \neq y']\!]$
(2) $\ell(y,y') = |y - y'|$
(3) $\ell(y,y') = [\![|y - y'| \leq K]\!]$ where $K < Y$

**d)** Assume that use the loss $\ell(y,y') = [\![y \neq y']\!]$. Plot the precision $\varepsilon$ as a function of the number of examples $l \in \{10, 100, \dots, 100000\}$ for confidence levels $\gamma \in \{0.9, 0.95, 0.99\}$.

**e)** Assume that we use the loss $\ell(y,y') = [\![y \neq y']\!]$. What is the minimal number of examples $l$ we need to use to have a guarantee that the test risk will approximate the true risk with error $\pm 1\%$ at most?

**Assignment 2.** Let $\mathcal{X}$ be a set of input observations and $\mathcal{Y} = \mathcal{A}^n$ a set of sequences of length $n$ defined over a finite alphabet $\mathcal{A}$. Let $h\colon \mathcal{X} \to \mathcal{Y}$ be a prediction rule that for each $x \in \mathcal{X}$ returns a sequence $h(x) = (h_1(x),\dots,h_n(x))$. Assume that we want to measure the prediction accuracy of $h(x)$ by the expected Hamming distance $R(h) = \mathbb{E}_{(x,y_1,\ldots,y_n)\sim p}(\sum_{i=1}^n [\![h_i(x) \neq y_i]\!])$ where $p(x, y_1, \ldots, y_n)$ is a p.d.f. defined over $\mathcal{X} \times \mathcal{Y}$. As the distribution $p(x, y_1, \ldots, y_n)$ is unknown we estimate $R(h)$ by the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l} \sum_{j=1}^{l} \sum_{i=1}^{n} [\![y_i^j \neq h_i(x^j)]\!]$$

where $\mathcal{S}^l = \{(x^i, y_1^i, \ldots, y_n^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x, y_1, \ldots, y_n)$.

**a)** Assume that the sequence length is $n = 10$ and that we compute the test error from $l = 1000$ examples. Use the Hoeffding inequality to bound the probability that $R(h)$ will be in the interval $(R_{\mathcal{S}^l}(h) - 1, R_{\mathcal{S}^l}(h) + 1)$ ?

**b)** What is the minimal number of the test examples $l$ which we need to collect in order to guarantee that $R(h)$ is in the interval $(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon)$ with probability $\gamma$ at least? Write $l$ as a function of $\varepsilon$, $n$ and $\gamma$.

**Assignment 3.** Let $\mathcal{X} = [a, b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1) = p(x \mid y = -1)$ be uniform distributions on $\mathcal{X}$ and $p(y = +1) = 0.8$. Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns the strategy

$$h_m(x) = \begin{cases} y^j & \text{if } \; x = x^j \;\text{ for some }\; j \in \{1, \ldots, m\} \\ -1 & \text{otherwise} \end{cases}$$

**a)** Show that the empirical risk $R_{\mathcal{T}^m}(h_m) = \frac{1}{m} \sum_{i=1}^m \ell(y^i, h_m(x^i))$ equals 0 with probability 1 for any finite $m$.

**b)** Show that the expected risk $R(h_m) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h_m(x))$ equals 0.8 for any finite $m$.