Assignment 1. Let the observation $x \in \mathcal{X} = \mathbb{R}^n$ and the hidden state $y \in \mathcal{Y} = \{+1, -1\}$ be generated by a multivariate normal distribution

$$p(x, y) = p(y) \frac{1}{(2\pi)^{\frac{n}{2}} \det(C_y)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_y)^T C_y^{-1} (x-\mu_y)}$$

where $\mu_y \in \mathbb{R}^n$, $y \in \mathcal{Y}$, are mean vectors, $C_y \in \mathbb{R}^{n \times n}$, $y \in \mathcal{Y}$, are covariance matrices and $p(y)$ is a prior probability. Assume that the model parameters are unknown and we want to learn a strategy $h : \mathcal{X} \rightarrow \mathcal{Y}$ which minimizes the probability of misclassification. To this end we use a learning algorithm $A : \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$ which returns a strategy $h$ from the class $\mathcal{H} = \{h(x) = \text{sign}(\langle w, x \rangle + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R}\}$ containing all linear classifiers.

a) What is the approximation error in case that $C_+ = C_-$?

b) Is the approximation error going to increase or decrease if $C_+ \neq C_-$?

Assignment 2. We are given a set $\mathcal{H} = \{h_i : \mathcal{X} \rightarrow \{1, \ldots, 100\} \mid i = 1, \ldots, 1000\}$ containing 1000 strategies each predicting a biological age $y \in \{1, \ldots, 100\}$ from an image $x \in \mathcal{X}$ capturing a human face. The quality of a single strategy is measured by the expected absolute deviation between the predicted age and the true age

$$R_{\text{MAE}}(h) = \mathbb{E}_{(x,y) \sim p}|y - h(x)|,$$

where the expectation is computed w.r.t. an unknown distribution $p(x, y)$. The empirical estimate of $R_{\text{MAE}}(h)$ reads

$$R_T(h) = \frac{1}{m} \sum_{i=1}^{m} |y^i - h(x^i)|$$

where $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ is a set of examples drawn from i.i.d. random variables with the same unknown $p(x, y)$. Let $h_m \in \text{Arg \ min}_{h \in \mathcal{H}} R_T(h)$ be a strategy with the minimal empirical risk.

a) What is the minimal $\varepsilon > 0$ which allows you to claim that the expected risk $R_{\text{MAE}}(h_m)$ is in the interval $(R_T(h_m) - \varepsilon, R_T(h_m) + \varepsilon)$ with probability 95% at least?

b) What is the minimal number of the training examples $m$ which guarantees that $R_{\text{MAE}}(h_m)$ is in the interval $(R_T(h_m) - 1, R_T(h_m) + 1)$ with probability 95% at least?
**Assignment 3.** Assume we want to learn a strategy \( h: \mathcal{X} \to \mathcal{Y} \) minimizing the expectation \( R(h) = \mathbb{E}_{(x,y) \sim p} \ell(y, h(x)) \) of a loss \( \ell: \mathcal{Y} \times \mathcal{Y} \to [a,b] \) w.r.t. to some distribution \( p(x, y) \). We use the ERM algorithm to select \( h_m \in \text{Arg min}_{h \in \mathcal{H}} R_T(h) \) from the class \( \mathcal{H} = \{ h_i: \mathcal{X} \to \mathcal{Y} \mid i = 1, \ldots, H \} \) containing \( H \) strategies. Let \( h_\mathcal{H} \in \text{arg min}_{i=1,\ldots,H} R(h_i) \) be the best strategy in the class \( \mathcal{H} \). Let \( \varepsilon > 0 \) and \( \gamma \in (0, 1) \) be fixed.

Derive a formula to compute the minimal number of training examples \( m \) such that
\[
\mathbb{P} \left( R(h_m) - R(h_\mathcal{H}) < \varepsilon \right) \geq \gamma,
\]
i.e. probability of having the estimation error \( R(h_m) - R(h_\mathcal{H}) \) less than \( \varepsilon \) is at least \( \gamma \).

**Hint:** use the results from Slides 8 and 10 of Lecture 3.

**Assignment 4.** Let \( \mathcal{H} \subseteq \{+1, -1\}^\mathcal{X} \) be a hypothesis class with VC dimension \( d < \infty \) and \( T^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m \) a training set drawn from i.i.d. random variables with distribution \( p(x, y) \). Then, the following inequality holds for any \( \varepsilon > 0 \),
\[
\mathbb{P} \left( \sup_{h \in \mathcal{H}} \left| R^{0/1}_0(h) - R^{0/1}_{T^m}(h) \right| \geq \varepsilon \right) \leq 4 \left( \frac{2 e m}{d} \right)^d e^{-\frac{m \varepsilon^2}{8}},
\]
where \( R^{0/1}_0(h) = \mathbb{E}_{(x,y) \sim p}(\mathbb{I}[y \neq h(x)]) \) and \( R^{0/1}_{T^m}(h) = \frac{1}{m} \sum_{i=1}^m [y^i \neq h(x^i)] \).

Show that this implies the ULLN for the class of strategies \( \mathcal{H} \).