Assignment 1 (Chebyshev). In this assignment, we will derive the Chebyshev inequality for the empirical risk. Let $X$ be a real valued random variable with expectation $\mu$ and finite variance $\nu$. The Chebyshev inequality asserts

$$\mathbb{P}( |X - \mu| > \varepsilon ) \leq \frac{\nu}{\varepsilon^2}.$$  

Let $X_i$, $i = 1, \ldots, m$ be independent, identically distributed random variables with expectation $\mu$ and finite variance $\nu$ and let $\bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_i$ be their empirical mean. Prove the inequality

$$\mathbb{P}( |\bar{X} - \mathbb{E} \bar{X}| > \varepsilon ) \leq \frac{\nu}{m\varepsilon^2}. \quad (1)$$

**Hint:** Recall the definition of the variance of a random variable. What is the variance of a sum of independent random variables?

Let us now consider a predictor $h: \mathcal{X} \to \mathcal{Y}$, and a loss $\ell(y, y')$. The risk of the predictor is denoted by $R(h)$ and its empirical risk on a test set $T_m = \{(x^j, y^j) \mid j = 1, \ldots, m\}$ is denoted by $R_{T_m}(h)$. Apply (1) to obtain the Chebyshev inequality for empirical risk in the lecture 2 slide 5.

Assignment 2 (Hoeffding). Next we prove the Hoeffding inequality for the empirical risk. Let $X_i$, $i = 1, \ldots, m$ be independent random variables bounded by the interval $[a,b]$, i.e. $a \leq X_i \leq b$. Let $\bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_i$ be their empirical mean. The Hoeffding inequality asserts that

$$\mathbb{P}( |\bar{X} - \mathbb{E} \bar{X}| > \varepsilon ) \leq 2 \exp\left(-\frac{2m \varepsilon^2}{(b-a)^2}\right).$$

As in the previous assignment, let us now consider a predictor $h: \mathcal{X} \to \mathcal{Y}$, and a loss $\ell(y, y')$. The risk of the predictor is denoted by $R(h)$ and its empirical risk on a test set $T_m = \{(x^j, y^j) \mid j = 1, \ldots, m\}$ is denoted by $R_{T_m}(h)$.

a) Prove that the generalisation error of $h$ can be bounded in probability by

$$\mathbb{P}\left( |R(h) - R_{T_m}(h)| > \varepsilon \right) < 2e^{-\frac{2m \varepsilon^2}{(b-a)^2}}, \quad (2)$$

where $\Delta \ell = \ell_{\text{max}} - \ell_{\text{min}}$.

b) Verify the value $m$ given in Example 1 of Lecture 2. for the special case of a binary classifier and the 0/1-loss.
Assignment 3 (Log Softmax). Consider a neural network with outputs $y_k$, $k = 1, \ldots, K$ representing posterior class probabilities. The last layer of this network is a softmax layer with output

$$y_k = \frac{e^{x_k}}{\sum_{\ell} e^{x_{\ell}}},$$

where $x_k$ are the outputs of the last linear layer and represent class scores. When learning such a network by maximising the log conditional likelihood, we have to consider log-probabilities

$$z_k = \log y_k = x_k - \log \sum_{\ell} e^{x_{\ell}}.$$

We will analyze the nonlinear part of the r.h.s., the log-sum-exp (aka smooth maximum) function:

$$f(x) = \log \sum_{\ell} e^{x_{\ell}}$$

(a) Prove that its gradient is given by $\nabla f(x) = y = \text{softmax}(x)$, i.e. by the vector of class probabilities. Conclude that the norm of the gradient is bounded by 1. This is a good property for gradient-based optimization. Also consider numerical stability of computing forward and backward of log softmax as a single operation versus the composition $\log \circ \text{softmax}$.

(b) Compute the second derivative of $f$ and show that it can be expressed as

$$\nabla^2 f(x) = \text{Diag}(y) - yy^T.$$

Prove that this symmetric matrix is positive semi-definite and conclude that $f(x)$ is a convex function. Note that the second derivative of log-sum-exp is the Jacobian of softmax.

Assignment 4 (Backprop). Given an operation with the output vector $y$ and the derivative of the loss w.r.t. $y$ – a row vector $J_y$, the "backprop" operation needs to compute derivatives w.r.t. all inputs. Compute the backprop of the following operations:

(a) $y = |x|$, where the absolute value is applied coordinate-wise to a vector $x$.

(b) $y = x + z$

c) $y = (x; z)$ — the concatenated vector of $x$ and $z$

d) Convolution in 1D: $y_i = \sum_k w_k x_{i-k} + b_i$. The inputs are: $w, x, b$. Ignore the index ranges for simplicity.