Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne

Mathematics of Learning – Worksheet 12

- The exercise sheets will be uploaded every Monday. Solution sketches will be uploaded one week later.
- You can hand in your own solutions via StudOn and we correct them - this is not mandatory. Please hand in small groups of 2-3 students.
- For questions, please use the forum on StudOn since other students may have similar questions. If you have a more personal question about the exercises please send an email to ehsan.waiezi@fau.de or lars.weidner@fau.de respectively.

Basics [Expected Values, Variance, Moments of random variables.]
Given a probability space $(\Omega, \mathscr{A}, P)$ and any real-valued random variable $X : \Omega \to \mathbb{R}$, we say that a probability density function (PDF) f_X is associated to X, if for every measurable set $A \subset \mathbb{R}$, $P(X(\omega) \in A) = \int_A f_X(x)dx$.

a) Let X be an equally distributed random variable ove the interval $[-5, 5]$, i.e., the PDF is

$$f_X(x) = \begin{cases} \frac{1}{10}, \text{ if } x \in [-5,5] \\ 0 \text{ otherwise.} \end{cases}$$

Calculate the probability $P(X \in [-1, 2])$ and the probability $P(|X| \in [3, 5])$.

b) Let X be a random variable with the PDF

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda x} \text{ if } x \geq 0 \\ 0 \text{ otherwise.} \end{cases}$$

Calculate the probability $P(X \in [-1, 2])$ and the probability $P(X^2 \in [4, 9])$.

c) The expected value of a random variable X with associated PDF f_X can be calculated as

$$\int_{\mathbb{R}} x f_X(x) dx.$$

Calculate the expected values of the random variables from a) and b).

d) The k-th moment of a random variable X with associated PDF f_X is the expected value of X^k and can be calculated as

$$\int_{\mathbb{R}} x^k f_X dx.$$

Calculate the k-th moment for the random variables from a) and b) for $k = 2, 3$.

e) Investigate for which moments of random variables ($k \in \mathbb{N}$) the following holds: For given random variables X and Y, and scalars $\lambda, \mu \in \mathbb{R}$,

$$\mathbb{E}[(\lambda X + \mu Y)^k] = \lambda \mathbb{E}[X^k] + \mu \mathbb{E}[Y^k].$$

f) The Variance of a random variable is defined as $\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$. Prove or disprove: If $\mathbb{E}[X^2]$ is finite, then

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Calculate the variance of random variables of a) and b) afterwards.

Exercise 1 [Convergence of SGD for strongly convex functions].
The update scheme for stochastic gradient descent (SGD) is given by

$$\begin{aligned}
&(1) \quad \text{sample gradient estimator } g_k \\
&(2) \quad \theta_{k+1} \leftarrow \theta_k - \eta_k g_k, \\
&(3) \quad k \leftarrow k+1, \quad \text{go back to } (1),
\end{aligned}$$

where g_k is an unbiased gradient estimator of a loss function $\mathcal{L}$ with

$$\begin{aligned}
&\mathbb{E}[g_k] = \nabla \mathcal{L}(\theta_k), \\
&\mathbb{E}[\|g_k - \nabla \mathcal{L}(\theta_k)\|^2] \leq \sigma^2.
\end{aligned}$$

Assume that $\mathcal{L}$ is μ-strongly convex and L-smooth for constants $0 < \mu \leq L < \infty$, i.e., for all $\theta, \tilde{\theta}$ it holds

$$\mathcal{L}(\tilde{\theta}) + \langle \nabla \mathcal{L}(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{\mu}{2}\|\theta - \tilde{\theta}\|^2 \leq \mathcal{L}(\theta) \leq \mathcal{L}(\tilde{\theta}) + \langle \nabla \mathcal{L}(\tilde{\theta}), \theta - \tilde{\theta} \rangle + \frac{L}{2}\|\theta - \tilde{\theta}\|^2.$$

Assume that the step sizes η_k are such that

$$\lim_{k \to \infty} \eta_k = 0, \qquad \sum_{k=0}^{\infty} \eta_k = \infty.$$

Let θ^* denote the global minimum of $\mathcal{L}$ (you do not have to prove that this exists and is unique).

- Using strong convexity, show that the error $d_k := \mathbb{E}[\|\theta^k - \theta^*\|^2]$ satisfies the following recursive estimate:

  $$d_{k+1} \leq (1 - \eta_k \mu) d_k + \eta_k^2 \sigma^2 + \eta_k^2 \mathbb{E}[\|\nabla \mathcal{L}(\theta_k)\|^2].$$

  [Hint: Start with $d_{k+1} = \mathbb{E}[\|\theta^{k+1} - \theta^*\|^2]$, use the SGD update, and expand the square!]

- Use that $\mathcal{L}$ is L-smooth to show that

  $$d_{k+1} \leq \left(1 - \eta_k \mu \left(1 - \eta_k \frac{L^2}{\mu}\right)\right) d_k + \eta_k^2 \sigma^2.$$

  [Hint: Remember that $\nabla \mathcal{L}(\theta^*) = 0$ since θ^* is the global minimum of $\mathcal{L}$.]

- Argue that for $\eta_k < \frac{\mu}{L^2}$ there exists a constant $c > 0$ such that it holds

$$d_{k+1} \leq (1 - \eta_k c\mu) d_k + \eta_k^2 \sigma^2.$$

- Show that $\lim_{k\to\infty} d_k = 0$ if $\eta_k < \frac{1}{c\mu}$.

- Proof by induction that for step sizes of the form $\eta_k = \frac{\theta}{k}$ for suitable $\theta > 0$, there exists a constant $C > 0$ such that

$$d_k \leq \frac{C}{k}.$$

Exercise 2 [Implementation of an artificial neural network].
Implement and train a fully connected feedforward network with a sigmoidal activation function in each neuron for automatic recognition of handwritten digits from the popular MNIST database. You can use the provided code skeleton in the file `NeuralNetwork_MNIST_incomplete` uploaded on StudOn.

You can download the MNIST database named `mnist.pkl.gz` from StudOn. It contains vectorized images of handwritten digits of size 28×28 pixels together with a ground truth label, i.e., a digit in $\{0, \ldots, 9\}$.

We propose you to divide this implementation exercise into the following subtasks:

1. Initialize the artificial neural network with random weights and biases, e.g., normally distributed random variables

2. Implement the sigmoidal activation function and its derivative

3. Realize a feedforward pass, i.e., compute the output vector of the neural network for a given vectorized image

4. Optionally: implement a second version of feedforward pass, saving all intermediate results (you will need them for backprop.)

5. Implement a partitioning of the training data into randomized mini batches

6. Implement the backpropagation algorithm for a given mini batch

7. Realize a loop over multiple training epochs, where in each iteration the neural network is trained for all mini batches

Hint: If you get stuck for a while and need help, please use StudOn (or any kind of communication) to ask questions and help each other!