Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne Winter term 23/24

Mathematics of Learning – Worksheet 12 - – Discussion on Jan 18th/19th, 2024

- 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 $P(|X| \in [3, 5])$.

Solution. We have to calculate the integrals

$$\int_{-1}^{2} \frac{1}{10} dx = 0.3$$

and

$$P(|X| \in [3,5]) = P(X \in [-5,-3]) + P(X \in [3,5]) = \int_{-5}^{-3} \frac{1}{10} dx + \int_{3}^{5} \frac{1}{10} dx = 0.2 + 0.2 = 0.4.$$

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 $P(X^2 \in [4, 9])$.

Solution. We note that $\frac{d}{dx}(-e^{-\lambda x}) = \lambda e^{-\lambda x}$. We have to calculate the integrals

$$\int_{-1}^{0} 0 dx + \int_{0}^{2} \lambda \cdot e^{-\lambda x} dx = [-e^{-\lambda x}]_0^2 = 1 - e^{-2\lambda}.$$

and

$$P(X^2 \in [4,9]) = P(X \in [2,3]) = \int_2^3 \lambda \cdot e^{-\lambda x} dx = [-e^{-\lambda x}]_2^3 = e^{-2\lambda} - e^{-3\lambda}.$$

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).

Solution. We have to calculate for the first random variable

$$\int_{-5}^{5} \frac{1}{10} x dx = 0$$

and for the second we use partial integration to get

$$\int_0^\infty \lambda x e^{-\lambda x} dx = [-xe^{-\lambda x}]_0^\infty - \int_0^\infty -e^{-\lambda x} dx = 0 + \int_0^\infty e^{-\lambda x} dx = [-\frac{1}{\lambda} e^{-\lambda x}]_0^\infty = \frac{1}{\lambda}.$$

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 $k = 2, 3$.

Solution. For the first random variable, we calculate the integrals

$$\int_{-5}^{5} \frac{1}{10} x^2 dx = [\frac{1}{10} \cdot \frac{1}{3} x^3]_{-5}^5 = \frac{25}{3}$$

and

$$\int_{-5}^{5} \frac{1}{10} x^3 dx = [\frac{1}{10} \cdot \frac{1}{4} x^4]_{-5}^5 = 0.$$

For the second random variable, we calculate

$$\int_0^\infty \lambda x^2 e^{-\lambda x} dx = [-x^2 e^{-\lambda x}]_0^\infty - \int_0^\infty -2xe^{-\lambda x} dx = 0 + \int_0^\infty 2xe^{-\lambda x} dx = \frac{2}{\lambda} \int_0^\infty \lambda x e^{-\lambda x} dx \overset{c)}{=} \frac{2}{\lambda^2}.$$

and

$$\int_0^\infty \lambda x^3 e^{-\lambda x} dx = [-x^3 e^{-\lambda x}]_0^\infty - \int_0^\infty -3x^2 e^{-\lambda x} dx = 0 + \int_0^\infty 3x^2 e^{-\lambda x} dx = \frac{3}{\lambda} \int_0^\infty \lambda x^2 e^{-\lambda x} dx = \frac{6}{\lambda^3}.$$

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

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

Solution. For $k = 1$, the equation holds, since the expected value can also be calculated integrating over the basic set of the probability space, this is Ω.

$$\mathbb{E}[(\lambda X + \mu Y)] = \int_\Omega \lambda X(\omega) + \mu Y(\omega) d\omega = \lambda \int_\Omega X(\omega) d\omega + \mu \int_\Omega Y(\omega) d\omega = \lambda \mathbb{E}[X] + \mu \mathbb{E}[Y].$$

For higher moments this does not hold in general, look for example at constant random variables $X = 1$ and $Y = 2$. We get $3^k \neq 1^k + 2^k$ for $k \neq 1$.

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.

Solution. We do the following easy calculation, using the results from e):

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] - \mathbb{E}[X]^2] = \mathbb{E}[X^2] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}[X]^2 =$$

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

For the random variable of a) we obtain:

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \overset{c)\ d)}{=} \frac{25}{3} - 0^2 = \frac{25}{3}$$

and for b):

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \overset{c)\ d)}{=} \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

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:

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

  [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

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

  [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

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

- 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

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

Solution.

First item We want to show:

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

For this we compute

$$\begin{aligned}
d_{k+1} &= \mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] = \mathbb{E}[\|\theta_k - \eta_k g_k - \theta^*\|^2] \\
&= \mathbb{E}[\|\theta_k - \theta^*\|^2 - 2\eta_k \langle g_k, \theta_k - \theta^* \rangle + \eta_k^2 \|g_k\|^2] \\
&= d_k + 2\eta_k \mathbb{E}[\langle \nabla \mathcal{L}(\theta_k), \theta^* - \theta_k \rangle] + \eta_k^2 \mathbb{E}[\|g_k\|^2].
\end{aligned}$$

From the strong convexity if follows

$$\langle \nabla \mathcal{L}(\theta_k), \theta^* - \theta_k \rangle \le \mathcal{L}(\theta^*) - \mathcal{L}(\theta_k) - \frac{\mu}{2} \|\theta_k - \theta^*\|^2 \le -\frac{\mu}{2} \|\theta_k - \theta^*\|^2$$

and the variance bound implies

$$\mathbb{E}[\|g_k\|^2] \le \sigma^2 + \|\nabla \mathcal{L}(\theta_k)\|^2.$$

Plugging both estimates in yields

$$d_{k+1} \le (1 - \eta_k \mu) d_k + \eta_k^2 \sigma^2 + \eta_k^2 \|\nabla \mathcal{L}(\theta_k)\|^2.$$

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

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

Since $\nabla \mathcal{L}$ is Lipschitz continuous, it holds

$$\mathbb{E}[\|\nabla \mathcal{L}(\theta_k)\|^2] = \mathbb{E}[\|\nabla \mathcal{L}(\theta_k) - \nabla \mathcal{L}(\theta^*)\|^2] \leq L^2 \mathbb{E}[\|\theta_k - \theta^*\|^2] = L^2 d_k.$$

Hence it follows

$$\begin{aligned} d_{k+1} &\leq (1 - \eta_k \mu) d_k + \eta_k^2 \sigma^2 + \eta_k^2 \|\nabla \mathcal{L}(\theta_k)\|^2 \leq (1 - \eta_k \mu) d_k + \eta_k^2 \sigma^2 + L^2 d_k \eta_k^2 \\ &= \left(1 - \eta_k \mu \left(1 - \eta_k \frac{L^2}{\mu}\right)\right) d_k + \eta_k^2 \sigma^2. \end{aligned}$$

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

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

In this case it holds

$$1 - \eta_k \frac{L^2}{\mu} \geq c > 0$$

and therefore

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

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

Let $\varepsilon > 0$ be arbitrary. Then it holds that

$$\begin{aligned} d_{k+1} - \varepsilon &\leq (1 - \eta_k \mu c)(d_k - \varepsilon) - \eta_k \mu c \varepsilon + \eta_k^2 \sigma^2 \\ &= (1 - \eta_k \mu c)(d_k - \varepsilon) + \eta_k (\eta_k \sigma^2 - \mu c \varepsilon) \\ &\leq (1 - \eta_k \mu c)(d_k - \varepsilon), \end{aligned}$$

where the last inequality holds for k large enough, using that $\eta_k \to 0$.

Iterating this, we obtain for any $n \in \mathbb{N}$:

$$d_{k+n} - \varepsilon \leq \prod_{i=k}^{k+n-1} (1 - \eta_i \mu c)(d_k - \varepsilon).$$

Finally, since $\eta_i < \frac{1}{\mu c}$ and one has the inequality $\log(1 - x) \leq -x$ for $x < 1$, it holds

$$\begin{aligned} \prod_{i=k}^{k+n-1} (1 - \eta_i \mu c) &= \exp\left(\log\left(\prod_{i=k}^{k+n-1} (1 - \eta_i \mu c)\right)\right) = \exp\left(\sum_{i=k}^{k+n-1} \log(1 - \eta_i \mu c)\right) \\ &\leq \exp\left(-\mu c \sum_{i=k}^{k+n-1} \eta_i)\right) \to 0, \quad n \to \infty. \end{aligned}$$

Hence, we get $\lim_{k \to \infty} d_k \leq \varepsilon$ and since ε was arbitrary, we have shown the claim.

Fifth item Prove 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

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

Let $\theta < \frac{1}{c\mu}$ and $C := \max(\frac{\theta^2\sigma^2}{\theta\mu c - 1}, d_1)$. Then we claim that $d_k \leq C/k$ for all $k \in \mathbb{N}$.

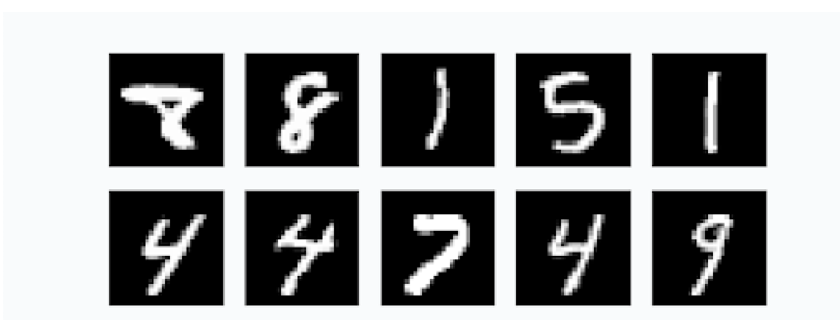
To start the induction for $k = 1$: $d_1 = d_1/1 \leq C/k$.

Assume now that $d_k \leq C/k$ for some $k \in \mathbb{N}$. Then it holds

$$\begin{aligned}
d_{k+1} &\leq \left(1 - \frac{\theta\mu c}{k}\right) d_k + \frac{\theta^2}{k^2}\sigma^2 \\
&\leq \left(1 - \frac{\theta\mu c}{k}\right) \frac{C}{k} + \frac{\theta^2}{k^2}\sigma^2 \\
&= \frac{C}{k} - \frac{\theta C\mu c}{k^2} + \frac{\theta^2\sigma^2}{k^2} \\
&= \frac{C}{k+1}\left(\frac{k+1}{k} - \theta\mu c\frac{k+1}{k^2} + \frac{\theta^2\sigma^2}{C}\frac{k+1}{k^2}\right) \\
&\leq \frac{C}{k+1}\left(\frac{k+1}{k} - \theta\mu c\frac{k+1}{k^2} + \frac{\theta^2\sigma^2}{\theta^2\sigma^2}(\theta\mu c - 1)\frac{k+1}{k^2}\right) \\
&= \frac{C}{k+1}\left(\frac{k+1}{k} - \frac{k+1}{k^2}\right) \\
&= \frac{C}{k+1}\left(1 - \frac{1}{k^2}\right) \\
&\leq \frac{C}{k+1}.
\end{aligned}$$

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!

Solution. See the python code in StudOn.