Mathematics of Learning – Worksheet 10 - – Discussion on December 21th/22nd, 2023

- 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.

Exercise 1 [Descent directions, gradients, niveau lines]
Consider a function $f\colon \mathbb{R}^n \to \mathbb{R}$, to be continuously differentiable. We define a descent direction for f at a point $\tilde{x} \in \mathbb{R}^n$ as all vectors $\theta \in \mathbb{R}^n$, such that $\exists \rho > 0$ such that $f(\tilde{x} + \lambda\theta) < f(\tilde{x})$ for all $0 < \lambda < \rho$ (*"if we make a tiny (ρ) or an even smaller (λ) step in that direction (θ), the value of f gets smaller"*). Prove or disprove:

a) The set of descent directions for f at a point $\tilde{x}$ with $\nabla f(\tilde{x}) \neq 0$ is the set $\{\theta \in \mathbb{R}^n : \langle \nabla f(\tilde{x}), \theta \rangle < 0\}$.

b) For $\nabla f(\tilde{x}) \neq 0$, the minimization problem

$$\min_{\theta \in \mathbb{R}^n : ||\theta||_2 = 1} \lim_{\lambda \to 0} \frac{f(\tilde{x} + \lambda\theta) - f(\tilde{x})}{\lambda}$$

is solved by $-\frac{\nabla f(\tilde{x})}{||\nabla f(\tilde{x})||_2}$.

Solution. a) This is not true. Consider for example the function

$$\hat{f}\colon \mathbb{R}^2 \to \mathbb{R}, f(x) = -||x||_2^2 \text{ for all } x \in \mathbb{R}^2$$

at the point $x = (0,1)^T$. The gradient at this point is $(0,-2)^T$. Hence the set $\{\theta \in \mathbb{R}^2 : \langle \nabla f(\tilde{x}), \theta \rangle < 0\}$ is the set $\{\theta \in \mathbb{R}^2 : \theta_2 > 0\}$. If we look e.g. at the vector $(1,0)^T$ which is not in this set, but is a descent direction, since $f((0,1)^T + \lambda(1,0)^T) < f((0,1)^T)$ for all $\lambda > 0$, we see that the statement cannot be true.
But the set is a subset of the set of descent directions. To see this, (see part b)), we check that

$$\langle \nabla f(\tilde{x}), \theta \rangle = \lim_{\lambda \to 0} \frac{f(\tilde{x} + \lambda\theta) - f(\tilde{x})}{\lambda},$$

for the left to be negative, there exists $c < 0$, $\rho > 0$ such that for all $\lambda \leq \rho$ holds, that

$$f(\tilde{x} + \lambda\theta) - f(\tilde{x}) \leq c\lambda,$$

equivalent to θ being a descent direction since $f(\tilde{x} + \lambda\theta) < f(\tilde{x})$ for all $0 < \lambda < \rho$.

b) We know (or we read in the literature regarding multi-dimensional derivatives) that

$\lim_{\lambda \to 0} \frac{f(\tilde{x}+\lambda\theta)-f(\tilde{x})}{\lambda}$ is the definition of directional derivative of f at point $\tilde{x}$. We further know (or read in the literature) that directional derivatives can be calculated using the formula

$$\lim_{\lambda \to 0} \frac{f(\tilde{x}+\lambda\theta)-f(\tilde{x})}{\lambda} = \nabla f(\tilde{x})^T \theta.$$

We further exploit the Cauchy-Schwarz-Inequality:

$$-||v|| \cdot ||w|| \leq \langle v, w \rangle \leq ||v|| \cdot ||w||,$$

for nonzero vectors v, w, the left inequality holding with equality if and only if $w = av$ for an $a < 0$. To prove the statement, we first see that it holds for $\tilde{\theta} = -\frac{\nabla f(\tilde{x})}{||\nabla f(\tilde{x})||_2}$ that the norm is equal to 1. This means, that $\tilde{\theta}$ is feasible for the minimization problem. To prove the optimality, we use the Cauchy-Schwarz-Inequality. For arbitrary $\theta \in \mathbb{R}^n$ with $||\theta||_2 = 1$ it holds:

$$-||\nabla f(\tilde{x})||_2 = -||\nabla f(\tilde{x})||_2 \cdot ||\theta||_2 \leq \langle \nabla f(\tilde{x}), \theta \rangle.$$

Since $\tilde{\theta}$ is a negative multiple of $\nabla f(\tilde{x})$ the inequality becomes an equality for $\tilde{\theta}$. This implies:

$$\langle \nabla f(\tilde{x}), \tilde{\theta} \rangle = -||\nabla f(\tilde{x})||_2 \leq \langle \nabla f(\tilde{x}), \theta \rangle$$

for all θ with norm 1, and therefore $\tilde{\theta}$ is the unique minimum.

Exercise 2 [Derivative of logistic activation function].
Let $\psi \colon \mathbb{R} \to \mathbb{R}$ the logistic activation function for a perceptron defined as

$$\psi(t) \ := \ \frac{1}{1+e^{-t}}$$

Show that the derivative ψ' of ψ can be computed as:

$$\psi'(t) = \psi(t)(1-\psi(t)).$$

Solution. We first determine the derivative ψ' of ψ using the quotient rule:

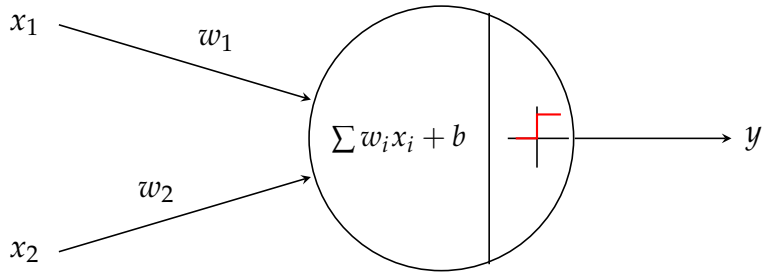
$$\psi'(t) = \frac{d}{dt}\psi(t) = \frac{d}{dt}\left(\frac{1}{1+e^{-t}}\right) = \frac{e^{-t}}{(1+e^{-t})^2}.$$

We can rewrite this term by adding zero as follows:

$$\begin{aligned}\psi'(t) = \frac{e^{-t}}{(1+e^{-t})^2} &= \frac{1}{1+e^{-t}} \cdot \frac{e^{-t}+(1-1)}{1+e^{-t}} = \frac{1}{1+e^{-t}} \cdot \left(\frac{1+e^{-t}}{1+e^{-t}} - \frac{1}{1+e^{-t}}\right) \\ &= \frac{1}{1+e^{-t}} \cdot \left(1 - \frac{1}{1+e^{-t}}\right) = \psi(t)(1-\psi(t)).\end{aligned}$$

Exercise 3 [Implementation of perceptrons for binary logic functions].

Let $f_\theta \colon \mathbb{R}^2 \to \mathbb{R}$ be a parametrized map realized by the following binary perceptron that maps two inputs $\vec{x} = (x_1, x_2)$ to an output y:



Here, $\theta \in \mathbb{R}^3$ is the vector of free parameters with $\theta := (w_1, w_2, b)$, where w_1, w_2 are the weights of the respective inputs and b is the bias of the perceptron. We assume that the activation function of the perceptron is the *Heavyside step function* $H\colon \mathbb{R} \to \{0,1\}$ defined as :

$$H(x) := \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0. \end{cases}$$

Implement a Python function `perceptron(x, theta)` that return a output `y`, which is either 0 or 1. Use this function to implement a family of perceptrons, which realize the following binary logic functions:

AND		
x_1	x_2	y
0	0	0
1	0	0
0	1	0
1	1	1

OR		
x_1	x_2	y
0	0	0
1	0	1
0	1	1
1	1	1

XOR		
x_1	x_2	y
0	0	0
1	0	1
0	1	1
1	1	0

NAND		
x_1	x_2	y
0	0	1
1	0	1
0	1	1
1	1	0

NOR		
x_1	x_2	y
0	0	1
1	0	0
0	1	0
1	1	0

Hint: One of the binary logic functions cannot be realized by a simple perceptron. Explain the reasons for this and suggest an alternative realization.

Solution. See the python file on StudOn (the explanation why XOR cannot be expressed as a perceptron is also there.).