Exercise 2: Math Background

We use the following notations in this exercise:

- Scalars are denoted with lowercase letters. E.g. x, ϕ
- Vectors are denoted with bold lowercase letters. E.g. $\boldsymbol{x}, \boldsymbol{\phi}$
- Matrices are denoted with bold uppercase letters. E.g. $\boldsymbol{X}, \boldsymbol{\Sigma}$

1 Linear algebra

Tasks:

a) Let

$$f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} + \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} - \boldsymbol{C}\boldsymbol{y} + \boldsymbol{D}$$

with $\boldsymbol{x} \in \mathbb{R}^M, \boldsymbol{y} \in \mathbb{R}^N$, function $f : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}$.
Compute the dimensions of the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ for the function so that the mathematical expression is valid.

b) Let $\boldsymbol{x} \in \mathbb{R}^N, \boldsymbol{M} \in \mathbb{R}^{N \times N}$. Express the function $f(\boldsymbol{x}) = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j M_{ij}$ using only matrix-vector multiplications.

c) Suppose $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}$, where $\boldsymbol{V}$ is a vector space. $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$ and $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 1$. Prove that $\boldsymbol{u} = \boldsymbol{v}$.

Note: In this task we define the norm as $\|\boldsymbol{v}\| := \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}$, where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ is the inner product between two vectors.

2 Linear Least Square

In this exercise, we want to determine the gradients for a few simple functions, which will be helpful for the upcoming lectures.

Note: Remember the definition of a *gradient*: The gradient of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$, denoted by ∇f, is a vector-valued function that gives, geometrically, the rate and direction of the steepest ascent of f at each point in $\mathbb{R}^n$. The components of the gradient are the partial derivatives of f with respect to each coordinate axis, and are written as:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

where $x_1, x_2, \ldots, x_n$ are the coordinates of a point in $\mathbb{R}^n$.

a) For $\boldsymbol{x} \in \mathbb{R}^n$, let $f : \mathbb{R}^n \to \mathbb{R}$ with $f(\boldsymbol{x}) = \boldsymbol{b}^\top \boldsymbol{x}$ for some known vector $\boldsymbol{b} \in \mathbb{R}^n$. Determine the gradient of the function f.
Hint: Use that $f(\boldsymbol{x}) = \boldsymbol{b}^\top \boldsymbol{x} = \sum_{i=1}^n b_i x_i$.

b) Now consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ with $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ for a symmetric matrix $\boldsymbol{A} \in \mathbb{S}_n$. Determine the gradient of the function f.
Hint: A symmetric matrix $\boldsymbol{A} \in \mathbb{S}_n$ satisfies that $A_{ij} = A_{ji}$ for all $1 \leq i, j \leq n$.

c) Now let us go a step further and let us determine the derivative of the following function $f : \mathbb{R}^n \to \mathbb{R}$ with

$$f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.

3 Calculus - derivatives

a) Compute the derivatives for the following functions: $f_i : \mathbb{R} \to \mathbb{R}$, $i \in \{1, 2, 3\}$

- $f_1 : f_1(x) = (x^3 + x + 1)^2$
- $f_2 : f_2(x) = \frac{e^{2x}-1}{e^{2x}+1}$
- $f_3 : f_3(x) = (1 - x)\log(1 - x)$ (**Note**: In this course, $\log(x) = \log_e(x) = \ln(x)$)

b) For a function $f : \mathbb{R}^n \to \mathbb{R}$, the *gradient* is defined as $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})$. Calculate the gradients of the following functions: $f_i : \mathbb{R}^2 \to \mathbb{R}$, $i \in \{4, 5\}$

- $f_4 : f_4(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2^2$
- $f_5 : f_5(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2$

c) For a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the *Jacobian* is defined as

$$\mathbb{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Calculate the Jacobian matrix of the following functions: $f_i : \mathbb{R}^n \to \mathbb{R}^m$, $i \in \{6, 7\}$

- $f_6 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^2$, $f_6(r, \varphi) = (r\cos\varphi, r\sin\varphi)^\top$
- $f_7 : \mathbb{R} \to \mathbb{R}^2$, $f_7(t) = (r\cos t, r\sin t)^\top$

d) For a function $f : \mathbb{R}^n \to \mathbb{R}^n$ the divergence is defined as $\text{div} f = \sum_{i=1}^{N} \frac{\partial f_i}{\partial x_i}$. Calculate the divergence for the following functions: $f_i : \mathbb{R}^n \to \mathbb{R}^n$, $i \in \{8, 9\}$

- $f_8 : \mathbb{R}^2 \to \mathbb{R}^2$, $f_8(x, y) = (-y, x)^\top$
- $f_9 : \mathbb{R}^2 \to \mathbb{R}^2$, $f_9(x, y) = (x, y)^\top$

4 Sigmoid derivative

In this question we will derive the derivative of the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

As seen in lecture 02, the sigmoid function is a popular activation function used in machine learning, which maps any input value to a value between 0 and 1. In logistic regression, the sigmoid function is used to map the output of the regression algorithm to a probability between 0 and 1, which can be interpreted as the probability of an input belonging to a particular class. This probability is then used to make a binary decision about whether the input belongs to the class or not.

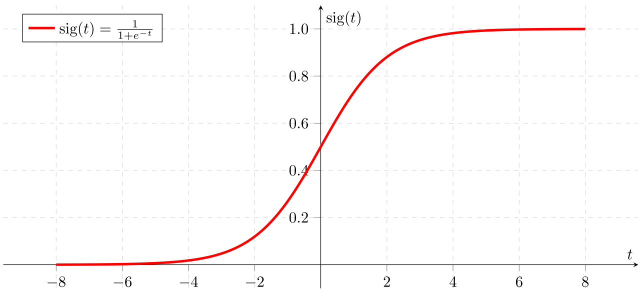

Figure 1: The sigmoid function

a) Find the derivative of the sigmoid function: $\frac{\partial \sigma(x)}{\partial x}$

b) Show that the derivative expression that you've found in the previous task could be represented with the sigmoid function iteslf, i.e.:

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1 - \sigma(x))$$

Hint: $e^{-x} = e^{-x} + 1 - 1$

5 Softmax derivative

In this exercise, we want to take a look at the softmax function, which is a common activation function in neural networks in order to normalize the output of a network to a probability distribution over predicted output classes. We will discuss the softmax function later in this lecture in more detail.

The softmax function $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

for $1 \leq i \leq n$ and $z = \begin{pmatrix} z_1 & z_2 & \dots & z_n \end{pmatrix} \top$. In the expanded form, we write:

$$\hat{y} = \sigma(z_1, z_2, \dots z_n) = \left[\frac{e^{z_1}}{\sum_{k=1}^{n} e^{z_k}}, \frac{e^{z_2}}{\sum_{k=1}^{n} e^{z_k}}, \cdots, \frac{e^{z_n}}{\sum_{k=1}^{n} e^{z_k}} \right].$$

Determine the derivative of the softmax function.

Hint: Deriving $\sigma(z)$ with respect to z will lead to $n \times n$ partial derivatives, i.e. $\frac{\partial \sigma(z)_i}{\partial z_j}$ for $1 \leq i, j \leq n$. It is important to consider the two cases (1) $i = j$ and (2) $i \neq j$

6 Probability

a) **Variance**.

We say that two random variables X, Y are independent if and only if the joint cumulative distribution function $F_{X,Y}(x, y)$ satisfies

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

In the case of independence, the following property holds for these variables: Let f, g be two real-valued functions defined on the codomains of X, Y, respectively. Then

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)].$$

Assume that X, Y are two random variables that are independent and identical distributed (i.i.d.) with $X, Y \sim \mathcal{N}(0, \sigma^2)$. Prove that

$$\mathrm{Var}(XY) = \mathrm{Var}(X)\mathrm{Var}(Y)$$

Remember this property, as it will play an important role at a later point of the lecture, when we take a look at the initialization of the weights of a neural network (Xavier initialization).

b) **Normal distribution**.

Remark: The family of random variables that are normally distributed is closed under linear transformation, that means if X is normally distributed, then for every $a, b \in \mathbb{R}$ the random variable $aX + b$ is normally distributed.

For this exercise, assume that the random variable X is normally distributed with mean μ and variance σ^2, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $Z = \frac{X-\mu}{\sigma}$. From the remark, we know that Z is again normally distributed. Determine the mean and the variance of the random variable Z.