

Technical University Munich Informatics



Introduction to Deep Learning (IN 2346)

Prof. Dr. Angela Dai

Exercise 2: Math Background

Exercise 1.1

Notation. We use the following notations in this exercise:

- Scalars are denoted with lowercase letters. E.g. x, ϕ
- AERMAN BERMAN CERIAIV • Vectors are denoted with bold lowercase letters. E.g. x, ϕ • Matrices are denoted with bold uppercase letters. E.g. X, Σ
- a) Let $x \in \mathbb{R}^{M, y} \in \mathbb{R}^{N, f}$ function $f : \mathbb{R}^{M} \times \mathbb{R}^{N} \to \mathbb{R}$, $f(x, y) = x^{\top} A y + x^{\top} B x C y + D$. Compute the dimensions of the matrices A, B, C, D for the function so that the mathematical
- 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 $u, v \in V$, where V is a vector space. ||u|| = ||v|| = 1 and $\langle u, v \rangle = 1$. Prove that $u = v. ||u|| = ||v|| = 1 \text{ and } \langle u, v \rangle = 1.$ $u = v. ||u - v||^2 = \langle u - v \rangle \langle u - v \rangle - \langle u - v \rangle \langle u - v \rangle \langle u - v \rangle = \langle u - v \rangle \langle u - v \rangle$ Exercise 1.2

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

- a) For $x \in \mathbb{R}^n$, let $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x) = b^{\top}x$ for some known vector $b \in \mathbb{R}^n$. Determine the gradient of the function f. Har = 2 AK; XK Hint: Use that $f(x) = b^{\mathsf{T}} x = \sum_{i=1}^{n} b_i x_i$.
- b) Now consider the quadratic function $f: \mathbb{R}^n \to \mathbb{R}$ with $f(x) = x^{\top} Ax$ for a symmetric matrix $A \in \mathbb{S}_n$. Determine the gradient of the function f. Hint: A symmetric matrix $A \in \mathbb{S}_n$ 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

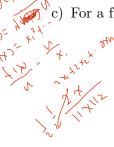
- a) Compute the derivatives for the following functions: $f_i: \mathbb{R} \to \mathbb{R}, i \in \{1, 2, 3\}$ $f_1: f_1(x) = (x^3 + x + 1)^2$ $f_2: f_2(x) = \frac{e^{2x} 1}{e^{2x} + 1}$ = $2e^{\frac{i\gamma}{2} \cdot (e^{\gamma x} + 1)} (e^{\frac{i\gamma}{2} \cdot 1}) \cdot 2e^{\frac{i\gamma}{2}}$ = $4e^{\frac{i\gamma}{2} \cdot 1}$
 - $f_3: f_3(x) = (1-x)\log(1-x)$



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

$$f_4: f_4(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2^2 \qquad \forall f = (\chi_1 \cdot - \chi_1)$$

- $f_5:f_5(x)=rac{1}{2}||x||_2$ $\forall t \in \left(\begin{array}{ccc} 1 & 1 & 1 \\ 2 & 1 & 1 \end{array}\right)$
- c) For a function $f: \mathbb{R}^n \to \mathbb{R}^m$, the *Jacobian* is defined as



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

d) For a function $f: \mathbb{R}^n \to \mathbb{R}^n$ the divergence is defined as $\operatorname{div} f = \sum_{i=1}^N \frac{\partial f_i}{\partial x_i}$. Calculate the

divergence for the following functions:
$$f_i: \mathbb{R}^n \to \mathbb{R}^n, i \in \{8,9\}$$

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

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

= $(-y,x)^\top$

= $(-y,x)^\top$

= $(-y,x)^\top$

Exercise 1.4

 $(1,1)^{7}$ 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

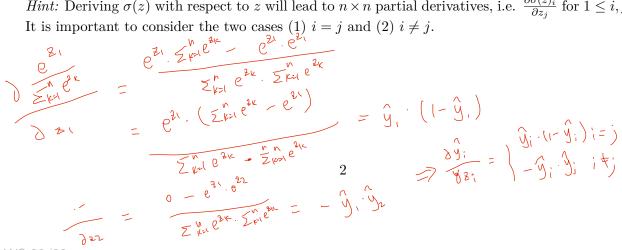
$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$$

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

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

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. $\frac{\partial \sigma(z)_i}{\partial z_i}$ for $1 \le i, j \le n$.



Var (xx) = E [x²]. E(²) - (E[x]. E(y])

= E [x²]. E(²) - (E[x]. E(y]) Exercise 1.5

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 $(\chi^2) - \xi(\chi) = (\chi^2) - \xi(\chi) =$

$$Var(XY) = Var(X)Var(Y)$$
. $\Rightarrow ((\chi^1) \cdot ((\chi^1) \cdot ((\chi^1) \cdot (\chi^1)) \cdot ((\chi^1) \cdot (\chi^1)$

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.

$$E(z) = E\left(\frac{1}{6}x - \frac{1}{6}h\right)$$

$$= \frac{1}{6}E(x) - \frac{h}{6}$$

$$= 0$$

$$Var(z) = Var\left(\frac{1}{6}x - \frac{1}{6}h\right)$$

$$= \frac{1}{6^2} \cdot Var(x)$$

$$= 0$$