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

$A \in R^{M\times N}$ $B \in R^{M\times M}$ $C \in R^{1\times N}$ $D \in R^{1\times 1}$

$1\times M$ $N\times 1$ $1\times M$ $M\times 1$ $N\times 1$

a) Let $\boldsymbol{x} \in \mathbb{R}^{M}$ $^{\times 1}$, $\boldsymbol{y} \in \mathbb{R}^{N}$ $^{\times 1}$, function $f : \mathbb{R}^M \times \mathbb{R}^N \to \mathbb{R}$, $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} + \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} - \boldsymbol{C}\boldsymbol{y} + \boldsymbol{D}$. Compute the dimensions of the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ for the function so that the mathematical expression is valid.

$1\times N \times N\times N \times N\times 1$

b) Let $\boldsymbol{x} \in \mathbb{R}^{N}$ $^{\times 1}$, $\boldsymbol{M} \in \mathbb{R}^{N\times N}$. 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.

$f(x) = x^T \cdot M \cdot x$

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 $\boldsymbol{u} = \boldsymbol{v}$.

$||u-v||^2 = \langle u-v, u-v\rangle$
$= \langle u,u\rangle - \langle u,v\rangle - \langle v,u\rangle + \langle v,v\rangle$
$= ||u||^2 - 2\langle u,v\rangle + ||v||^2$
$= 0$

$\langle u \cdot v\rangle = u \cdot v = u^T \cdot v = ||u|| \cdot ||v|| \cdot \cos\theta = 1$

$\cos\theta = 1$ $\quad \theta = 0.$

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

$\nabla f = b_k.$

$\frac{\partial f(x)}{\partial x_k} = 2A_{kj}x_k$

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 $1 \le i, j \le n$.

$\nabla f = A_{ij}$

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

$$f(x) = ||Ax - b||_2^2 = (Ax - b)^\top (Ax - b)$$

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

$= ((Ax)^T - b^T)(Ax - b)$
$= x^T A^T A x - x^T A^T b - b^T \cdot A \cdot x + b^T \cdot b$

$\nabla_x f(x) = 2A^T A x - A^T b - b^T A$

Exercise 1.3

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

- $f_1 : f_1(x) = (x^3 + x + 1)^2$ $\quad 2(3x^2+1)(x^3+x+1)$
- $f_2 : f_2(x) = \frac{e^{2x}-1}{e^{2x}+1}$ $\quad = \frac{2e^{2x}\cdot(e^{2x}+1) - (e^{2x}-1)\cdot 2e^{2x}}{(e^{2x}+1)^2} = \frac{4e^{2x}}{(e^{2x}+1)^2}$
- $f_3 : f_3(x) = (1 - x)\log(1 - x)$

$= -\log(1-x) + 1$

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: $f_i : \mathbb{R}^2 \to \mathbb{R}$, $i \in \{4, 5\}$

- $f_4 : f_4(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2^2$ $\quad \nabla f = (x_1 \cdots x_n)$
- $f_5 : f_5(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x}||_2$ $\quad \nabla f = (\frac{1}{2} \cdots \frac{1}{2})$

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

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

$$J = \begin{pmatrix} \frac{f_1}{r} & \frac{f_1}{\varphi} \\ \frac{f_2}{r} & \frac{f_2}{\varphi} \end{pmatrix} = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}$$

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

- $f_6 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^2$, $f_6(r, \varphi) = (\overset{f_1}{r\cos\varphi}, \overset{f_2}{r\sin\varphi})^\top$ $= \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}$
- $f_7 : \mathbb{R} \to \mathbb{R}^2$, $f_7(t) = (r\cos t, r\sin t)^\top$ $\quad \begin{matrix} -r\sin t \\ r\cos t \end{matrix}$

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: $f_i : \mathbb{R}^n \to \mathbb{R}^n$, $i \in \{8, 9\}$

- $f_8 : \mathbb{R}^2 \to \mathbb{R}^2$, $f_8(x, y) = (-y, x)^\top$ $\quad \nabla f = \left(\frac{-y}{x} + \frac{x}{y}, \frac{-y}{x} + \frac{\partial x}{x} \right)$ $\text{div} f = 0$
- $f_9 : \mathbb{R}^2 \to \mathbb{R}^2$, $f_9(x, y) = (x, y)^\top$ $\quad = (1, 1)^\top$ $\quad f_9 = 2$

$(1, 1)^\top$

Exercise 1.4

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 $z = \begin{pmatrix} z_1 & z_2 & \ldots & z_n \end{pmatrix}^\top$. In the expanded form, we write:

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

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 $1 \le i, j \le n$. It is important to consider the two cases (1) $i = j$ and (2) $i \ne j$.

$$\frac{\partial \frac{e^{z_1}}{\sum_{k=1}^n e^{z_k}}}{\partial z_1} = \frac{e^{z_1} \cdot \sum_{k=1}^n e^{z_k} - e^{z_1} \cdot e^{z_1}}{\sum_{k=1}^n e^{z_k} \cdot \sum_{k=1}^n e^{z_k}} = \frac{e^{z_1} \cdot (\sum_{k=1}^n e^{z_k} - e^{z_1})}{\sum_{k=1}^n e^{z_k} \cdot \sum_{k=1}^n e^{z_k}} = \hat{y}_1 \cdot (1 - \hat{y}_1)$$

$$\frac{\cdot}{\partial z_2} = \frac{0 - e^{z_1} \cdot e^{z_2}}{\sum_{k=1}^n e^{z_k} \cdot \sum_{k=1}^n e^{z_k}} = -\hat{y}_1 \cdot \hat{y}_2$$

$$\Rightarrow \frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i \cdot (1 - \hat{y}_i) & i = j \\ -\hat{y}_i \cdot \hat{y}_j & i \ne j \end{cases}$$

Exercise 1.5

$\mathrm{Var}(XY) = E[X^2Y^2] - E[XY]^2$

$= E[X^2]\cdot E[Y^2] - (\underbrace{E[X]\cdot E[Y]}_{=0})^2$

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 $(E(x^2) - \underbrace{E(x)^2}_{=0})(E(y^2) - \underbrace{E(y)^2}_{=0})$

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

$= E(x^2)\cdot E(y^2)$

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.

$E(Z) = E\left(\frac{1}{\sigma}x - \frac{1}{\sigma}\mu\right)$

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

$= 0$

$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{1}{\sigma}x - \frac{1}{\sigma}\mu\right)$

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

$= 1$