Exercise 2: Math Background (Solution)

1 Linear algebra

a) $\boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{B} \in \mathbb{R}^{M\times M}, \boldsymbol{C} \in \mathbb{R}^{1\times N}, \boldsymbol{D} \in \mathbb{R}^{1\times 1}$.

b) $f(\boldsymbol{x}) = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j M_{ij} = \sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j M_{ij} = \sum_{i=1}^{N} x_i (\boldsymbol{M}\cdot\boldsymbol{x})_i = \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}$.

c) Proof: Consider $\|\boldsymbol{u}-\boldsymbol{v}\|^2$, we have:

$$\begin{aligned}
\|\boldsymbol{u}-\boldsymbol{v}\|^2 &= \langle \boldsymbol{u}-\boldsymbol{v}, \boldsymbol{u}-\boldsymbol{v}\rangle \\
&= \langle \boldsymbol{u},\boldsymbol{u}\rangle - \langle \boldsymbol{u},\boldsymbol{v}\rangle - \langle \boldsymbol{v},\boldsymbol{u}\rangle + \langle \boldsymbol{v},\boldsymbol{v}\rangle \\
&= \|\boldsymbol{u}\|^2 - 2\langle \boldsymbol{u},\boldsymbol{v}\rangle + \|\boldsymbol{v}\|^2 \\
&= 0
\end{aligned}$$

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

* $||x||$ refers to the L2-norm $||x||_2$, unless stated otherwise.

2 Linear Least Square

a) By definition of the gradient, we need to determine $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$. For $1 \leq k \leq n$, we have

$$\frac{\partial f(\boldsymbol{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n} b_i x_i\right) = \sum_{i=1}^{n} \frac{\partial}{\partial x_k}(b_i x_i) = \sum_{i=1}^{n} \delta_{ik} b_i = b_k.$$

The Kronecker delta is defined as follows: $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$

Hence, we obtain $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \boldsymbol{b}$.

b) To determine the gradient of the function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A}$ is a symmetric matrix in $\mathbb{S}_n$, we can use the definition of the gradient:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]$$

We start by computing the partial derivative of f with respect to x_i.

$$\begin{aligned}\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}(\boldsymbol{x}^\top \cdot (\boldsymbol{A}\boldsymbol{x})) = \frac{\partial \boldsymbol{x}^\top}{\partial x_i} \cdot (\boldsymbol{A}\boldsymbol{x}) + \boldsymbol{x}^\top \cdot \frac{\partial (\boldsymbol{A}\boldsymbol{x})}{\partial x_i} = \boldsymbol{e_i}^\top \cdot (\boldsymbol{A}\boldsymbol{x}) + \boldsymbol{x}^\top \cdot \boldsymbol{A}\boldsymbol{e_i} \\ = \sum_j A_{ij}x_j + \sum_j A_{ij}x_j = 2\sum_j A_{ij}x_j = 2(\boldsymbol{A}\boldsymbol{x})_i\end{aligned}$$

where $\boldsymbol{e_i}$ is the standard basis vector in the i'th direction (1 at the i'th, and all other entries are 0's).

Thus, the gradient of f is:

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = [2(\boldsymbol{A}\boldsymbol{x})_1, 2(\boldsymbol{A}\boldsymbol{x})_2, \ldots, 2(\boldsymbol{A}\boldsymbol{x})_n] = 2\boldsymbol{A}\boldsymbol{x}$$

Therefore, the gradient of the quadratic function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ is $\frac{\partial f}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$.

c) Let us first rewrite the expression:

$$\begin{aligned}f(\boldsymbol{x}) &= \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \\ &= (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= ((\boldsymbol{A}\boldsymbol{x})^\top - \boldsymbol{b}^\top)(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= (\boldsymbol{x}^\top \boldsymbol{A}^\top - \boldsymbol{b}^\top)(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b} - \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{b} \\ &= \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b} + \boldsymbol{b}^\top \boldsymbol{b}.\end{aligned}$$

Note that $\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b} = \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}$, because both result with a scalar. Since if $s \in \mathbb{R} \rightarrow s^\top = s \rightarrow \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b} = (\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b})^\top = \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}$.

Thus, by using part a) $\rightarrow \frac{\partial \boldsymbol{b}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{b}$ and b) $\rightarrow \frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$, we obtain:

$$\begin{aligned}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) &= \nabla_{\boldsymbol{x}}(\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b} + \boldsymbol{b}^\top \boldsymbol{b}) = \nabla_{\boldsymbol{x}}\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \nabla_{\boldsymbol{x}} 2\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{b} + 0 \\ &= 2\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^\top \boldsymbol{b} = 2\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\end{aligned}$$

.

3 Calculus - derivatives

a) The derivatives are:

- $f_1'(x) = \left[(x^3+x+1)^2\right]' = 2(x^3+x+1)(x^3+x+1)' = 2(x^3+x+1)(3x^2+1)$
- $f_2'(x) = \left[\frac{e^{2x}-1}{e^{2x}+1}\right]' = \frac{(e^{2x}-1)'(e^{2x}+1)-(e^{2x}-1)(e^{2x}+1)'}{(e^{2x}+1)^2} = \frac{2e^{2x}(e^{2x}+1)-(e^{2x}-1)2e^{2x}}{(e^{2x}+1)^2} = \frac{4e^{2x}}{(e^{2x}+1)^2}$
- $f_3'(x) =$

$$\begin{aligned}
&= \left[(1-x)\log(1-x)\right]' \\
&= \log(1-x)\cdot(1-x)' + (1-x)\cdot\log'(1-x) \\
&= -\log(1-x) + (1-x)\cdot\frac{\partial\log(y)}{y}\cdot\frac{y}{x} = -\log(1-x) + (1-x)\cdot\frac{1}{1-x}\cdot(1-x)' \\
&= -\log(1-x) - 1
\end{aligned}$$

b) The gradients are:

- $\nabla f_4 = \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{1}{2}\|\boldsymbol{x}\|_2^2\right) = \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{x}\right) = \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{1}{2}\boldsymbol{x}^\top I\boldsymbol{x}\right) = \frac{1}{2}\cdot 2Ix = \mathbf{x}$
- $\nabla f_5 = \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{1}{2}\|\boldsymbol{x}\|_2\right) = \frac{\partial}{\partial \boldsymbol{x}}\left(\frac{1}{2}\sqrt{\boldsymbol{x}^\top\boldsymbol{x}}\right) = \frac{1}{2}\cdot\frac{1}{2}(\boldsymbol{x}^\top\boldsymbol{x})^{-\frac{1}{2}}\cdot\frac{\partial(\boldsymbol{x}^\top\boldsymbol{x})}{x} = \frac{1}{2}\cdot\frac{1}{2}(\boldsymbol{x}^\top\boldsymbol{x})^{-\frac{1}{2}}\cdot 2I\boldsymbol{x} = \frac{1}{2}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$

c) The Jacobians are:

- $J_{f_6} = \begin{bmatrix} \frac{\partial f_1}{\partial r} & \frac{\partial f_1}{\partial \varphi} \\ \frac{\partial f_2}{\partial r} & \frac{\partial f_2}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{bmatrix}$
- $J_{f_7} = \begin{bmatrix} \frac{\partial f_1}{\partial t} \\ \frac{\partial f_2}{\partial t} \end{bmatrix} = \begin{bmatrix} -r\sin t \\ r\cos t \end{bmatrix}$

d) The divergences are:

- $\operatorname{div} f_8 = \frac{\partial(-y)}{\partial x} + \frac{\partial x}{\partial y} = 0$
- $\operatorname{div} f_9 = \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} = 2$

4 Sigmoid derivative

a)

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{d}{dx}(1+e^{-x})^{-\frac{1}{2}} = \frac{-(1+e^{-x})^{-2}(-e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

b)

$$\begin{aligned}
&\frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{e^{-x}+1-1}{(1+e^{-x})^2} \\
&= \frac{1+e^{x}}{(1+e^{x})^2} - \frac{1}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}$$

5 Softmax derivative

5.1 1st approach - two cases

When deriving $\sigma(z)$ with respect to z, there are $n \times n$ partial derivates but we notice that they reduce to only two distinct kinds:

- $\hat{y}_i = \sigma(z)_i$ w.r.t z_i. For example, deriving $\frac{e^{z_1}}{\sum_{k=1}^{n} e^{z_k}}$ w.r.t z_1. (z_1 appears both in the nominator and in the denominator)
- $\hat{y}_i = \sigma(z)_i$ w.r.t $z_j, i \neq j$. For example, deriving $\frac{e^{z_1}}{\sum_{k=1}^{n} e^{z_k}}$ w.r.t z_2 (z_2 appears only in the denominator).

We first derive the first kind:

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

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

In the last and second to last equality, we used a trick, or the observation, that we can express these terms in means of $\hat{y}$. In a similar fashion, we derive the second kind:

$$\frac{\partial \hat{y}_1}{\partial z_2} = \partial \left(\frac{e^{z_1}}{\sum_{k=1}^{n} e^{z_k}} \right) / \partial z_2 = \frac{\cancelto{0}{0 \cdot \sum_{k=1}^{n} e^{z_k}} - e^{z_2} \cdot e^{z_1}}{\left(\sum_{k=1}^{n} e^{z_k}\right)\left(\sum_{k=1}^{n} e^{z_k}\right)} = -\frac{e^{z_2}}{\left(\sum_{k=1}^{n} e^{z_k}\right)} \cdot \frac{e^{z_1}}{\left(\sum_{k=1}^{n} e^{z_k}\right)} = -\hat{y}_1 \hat{y}_2.$$

In conclusion, the partial derivatives of the softmax layer $\hat{y} = \sigma(z)$ with respect to its input z are given by:

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

5.2 2nd approach - solve all in one!

A nice trick to solve both cases in one. First, we derive:

$$\frac{\partial \log(s_i)}{\partial z_j} = \frac{1}{s_i} \frac{\partial s_i}{\partial z_j}$$

Therefore:

$$\frac{\partial s_i}{\partial z_j} = s_i \cdot \frac{1}{s_i} \frac{\partial s_i}{\partial z_j} = s_i \cdot \frac{\partial \log(s_i)}{\partial z_j} = s_i \frac{\partial}{\partial z_j} \log\left(\frac{e^{z_i}}{\sum_{k=1}^{C} e^{z_k}}\right) = s_i \frac{\partial}{\partial z_j}[z_i - \log(\sum_{k=1}^{C} e^{z_k})]$$

$$= s_i(\delta_{ij} - \frac{1}{\sum_{k=1}^{C} e^{z_k}} e^{z_j}) = s_i(\delta_{ij} - s_j)$$

With

$$\begin{cases} \delta_{ij} = 1 & i = j \\ \delta_{ij} = 0 & i \neq j \end{cases}$$

6 Probability

a) We use the definition of the variance, namely

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{1}$$

and equivalently,

$$\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2. \tag{2}$$

Since $X, Y \sim \mathcal{N}(0, \sigma^2)$, we are given that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. With these observations, we obtain

$$\begin{aligned}
\mathrm{Var}(XY) &\overset{(1)}{=} \mathbb{E}[X^2Y^2] - \mathbb{E}[XY]^2 \\
&\overset{(*)}{=} \mathbb{E}[X^2]\mathbb{E}[Y^2] - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&\overset{(2)}{=} (\mathrm{Var}(X) + \mathbb{E}[X]^2)(\mathrm{Var}(Y) + \mathbb{E}[Y]^2) - \mathbb{E}[X]^2\mathbb{E}[Y]^2 \\
&= \mathrm{Var}(X)\mathrm{Var}(Y) + \mathrm{Var}(X)\underbrace{\mathbb{E}[Y]^2}_{=0} + \mathrm{Var}(Y)\underbrace{\mathbb{E}[X]^2}_{=0} \\
&= \mathrm{Var}(X)\mathrm{Var}(Y)
\end{aligned}$$

$(*) X, Y$ are independent.

b) We use the properties of the expectation and the variance of a random variable. For the mean of Z, we observe:

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\frac{X - \mu}{\sigma}\right] \\
&= \frac{1}{\sigma} \cdot \mathbb{E}[X - \mu] \\
&= \frac{1}{\sigma} \cdot (\mathbb{E}[X] - \mathbb{E}[\mu]) \\
&= \frac{1}{\sigma} \cdot (\mu - \mu) \\
&= 0
\end{aligned}$$

For the variance, remember that:

$$\begin{aligned}
&\mathrm{Var}\left[\frac{X - \mu}{\sigma}\right] \\
&= \mathbb{E}\left[\left(\frac{X - \mu}{\sigma} - \mathbb{E}\left[\frac{X - \mu}{\sigma}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{X - \mu}{\sigma} - \frac{\mathbb{E}[X] - \mu}{\sigma}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^2\right] \\
&= \frac{1}{\sigma^2}\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \frac{1}{\sigma^2} \cdot \mathrm{Var}[X].
\end{aligned}$$

Therefore, we observe that:

$$
\begin{aligned}
&\mathrm{Var}[Z] \\
&= \mathrm{Var}\left[\frac{X-\mu}{\sigma}\right] \\
&= \frac{1}{\sigma^2}\mathrm{Var}[X] \\
&= \frac{1}{\sigma^2}\sigma^2 \\
&= 1.
\end{aligned}
$$

In summary, we conclude that $Z \sim \mathcal{N}(0,1)$.