Technical University Munich Informatics



Introduction to Deep Learning (IN 2346)

Prof. Dr. Matthias Nießner

Exercise 2: Math Background (Solution)

Exercise 1.1

- 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) $f(\mathbf{x}) = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j M_{ij} = \sum_{i=1}^{N} x_i \sum_{j=1}^{N} x_j M_{ij} = \sum_{i=1}^{N} x_i (\mathbf{M} \cdot \mathbf{x})_i = \mathbf{x}^{\top} \mathbf{M} \mathbf{x}.$
- c) Proof: Consider $||\boldsymbol{u} \boldsymbol{v}||^2$, we have:

$$||\boldsymbol{u} - \boldsymbol{v}||^2 = \langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v} \rangle$$

$$= \langle \boldsymbol{u}, \boldsymbol{u} \rangle - \langle \boldsymbol{u}, \boldsymbol{v} \rangle - \langle \boldsymbol{v}, \boldsymbol{u} \rangle + \langle \boldsymbol{v}, \boldsymbol{v} \rangle$$

$$= ||\boldsymbol{u}||^2 - 2\langle \boldsymbol{u}, \boldsymbol{v} \rangle + ||\boldsymbol{v}||^2$$

$$= 0$$

Hence, u = v.

Exercise 1.2

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

have

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

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
$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$
.

b) Similar to the first part, we obtain $f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$ and the partial derivative of the

variable x_k with $1 \le k \le n$ is

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \left(\sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j \right)
= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial}{\partial x_k} \left(A_{ij} x_i x_j \right)
= \sum_{j=1, j \neq k}^n \frac{\partial}{\partial x_k} \left(x_k A_{kj} x_j \right) + \sum_{i=1, i \neq k}^n \frac{\partial}{\partial x_k} \left(A_{ik} x_i x_k \right) + \frac{\partial}{\partial x_k} \left(A_{kk} x_k^2 \right)
= \sum_{j=1, j \neq k}^n A_{kj} x_j + \sum_{i=1, i \neq k}^n A_{ik} x_i + 2A_{kk} x_k
= \sum_{j=1}^n A_{kj} x_j + \sum_{i=1}^n A_{ik} x_i
A \in \mathbb{S}_n \ 2 \cdot \sum_{j=1}^n A_{kj} x_j
= 2 \cdot (Ax)_k.$$

Hence, we obtain
$$\nabla_x f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} 2 \cdot (Ax)_1 \\ 2 \cdot (Ax)_2 \\ \vdots \\ 2 \cdot (Ax)_n \end{pmatrix} = 2Ax.$$

c) Let us first rewrite the expression:

$$f(x) = ||Ax - b||_{2}^{2}$$

$$= (Ax - b)^{\top} (Ax - b)$$

$$= ((Ax)^{\top} - b^{\top}) (Ax - b)$$

$$= (x^{\top} A^{\top} - b^{\top}) (Ax - b)$$

$$= x^{\top} A^{\top} Ax - x^{\top} A^{\top} b - b^{\top} Ax + b^{\top} b$$

$$= x^{\top} A^{\top} Ax - 2x^{\top} A^{\top} b + b^{\top} b.$$

Using part (a) and (b), we obtain

$$\nabla_x f(x) = \nabla_x (x^\top A^\top A x - 2x^\top A^\top b + b^\top b) = \nabla_x x^\top A^\top A x - \nabla_x 2x^\top A^\top b + 0$$
$$= 2A^\top A x - 2A^\top b$$

Exercise 1.3

a) The derivatives are:

•
$$f_1'(x) = \left[(x^3 + x + 1)^2 \right]' = 2(x^3 + x + 1)(x^3 + x + 1)' = 2(x^3 + x + 1)(3x^2 + 1)$$

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

•
$$f_3'(x) = [(1-x)\log(1-x)]' = -\log(1-x) - 1$$

b) The gradients are:

$$\bullet \quad \nabla f_4 = (x_1, x_2)^\top = \mathbf{x}$$

•
$$\nabla f_5 = \frac{1}{2}(x_1^2 + x_2^2)^{-\frac{1}{2}}(x_1, x_2)^{\top} = \frac{1}{2} \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$$

c) The Jacobians are:

•
$$J_{f_6} = \begin{bmatrix} \cos(\varphi) & -r\sin(\varphi) \\ \sin(\varphi) & r\cos(\varphi) \end{bmatrix}$$

•
$$J_{f_7} = \begin{bmatrix} -r\sin(t) \\ r\cos(t) \end{bmatrix}$$

d) The divergences are:

•
$$\operatorname{div} f_8 = 0$$

•
$$\operatorname{div} f_0 = 2$$

Exercise 1.4

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:

- $\sigma(z)_i$ 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)
- $\sigma(z)_i$ 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:

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

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:

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

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

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

Exercise 1.5

a) We use the definition of the variance, namely

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

and equivalently,

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

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

$$\operatorname{Var}(XY) \stackrel{(1)}{=} \mathbb{E}[X^{2}Y^{2}] - \mathbb{E}[XY]^{2}$$

$$\stackrel{(*)}{=} \mathbb{E}[X^{2}]\mathbb{E}[Y^{2}] - \mathbb{E}[X]^{2}\mathbb{E}[Y]^{2}$$

$$\stackrel{(2)}{=} (\operatorname{Var}(X) + \mathbb{E}[X]^{2})(\operatorname{Var}(Y) + \mathbb{E}[Y]^{2}) - \mathbb{E}[X]^{2}\mathbb{E}[Y]^{2}$$

$$= \operatorname{Var}(X)\operatorname{Var}(Y) + \operatorname{Var}(X)\underbrace{\mathbb{E}[Y]^{2}}_{=0} + \operatorname{Var}(Y)\underbrace{\mathbb{E}[X]^{2}}_{=0}$$

$$= \operatorname{Var}(X)\operatorname{Var}(Y)$$

- (*)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:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{X - \mu}{\sigma}\right]$$

$$= \frac{1}{\sigma} \cdot \mathbb{E}[X - \mu]$$

$$= \frac{1}{\sigma} \cdot (\mathbb{E}[X] - \mathbb{E}[\mu])$$

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

$$= 0$$

For the variance, we observe:

$$Var(Z) = Var\left[\frac{X - \mu}{\sigma}\right]$$

$$= \frac{1}{\sigma^2} \cdot Var[X - \mu]$$

$$= \frac{1}{\sigma^2} \cdot Var[X]$$

$$= \frac{1}{\sigma^2} \cdot \sigma^2$$

$$= 1$$

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