Prof. Dr. Frauke Liers, Christian Biefel Summer term 2022, 04.08.2022

Exam
Mathematics of Learning
Solution sketches

Name ....

Student registration no. ....
(Matrikelnummer)

Signature ....

- **Do not turn over this page until instructed to do so by the examiner!**

- The time for completing the exam is **60 min**.
- The exam consists of 4 questions with a total of 40 points.
- Answers have to be readable and justified.
- Write in black or blue.
- As auxiliary tool you may use one handwritten sheet of paper (DinA4, both sides). There are no other tools allowed (no books, no calculator or phone).

Good luck!

Q1 (8P)	Q2 (12P)	Q3 (13P)	Q4 (7P)	$\sum = 40$P

Grade:

Question 1. (8 points)
Let input data $x^1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$, $x^2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, $x^3 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$, $x^4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ be given. Compute for all data points the first principal component (i.e. dimension $k = 1$).

Solution Question 1:

1. (1P) Compute mean value
$$\overline{X} = \frac{1}{4}\sum_{i=1}^{4} x^i = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

2. (1P) Center data $y^i = x^i - \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

3. (2P) Compute covariance matrix
$$C = \frac{1}{4}\sum_{i=1}^{4} y^i (y^i)^T = \frac{1}{4}\left(\begin{pmatrix} 4 & 4 \\ 4 & 4 \end{pmatrix} + \begin{pmatrix} 4 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 4 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$
(Alternative: $Y = (y^1\ y^2\ y^3\ y^4)$, $C = \frac{1}{4}YY^T$)

4. (2P) Compute eigenvalues and eigenvectors. First, compute the roots of the characteristic polynomial of C:
$$\chi_C(\lambda) = (\lambda - 2)(\lambda - 2) - 1 = \lambda^2 - 4\lambda + 3.$$
Using the quadratic formula we can calculate the eigenvalues $\lambda = 3$ and $\lambda = 1$. Therefore, the largest eigenvalue is $\lambda = 3$ and since we want to compute one principal component it is enough to calculate the eigenvector for $\lambda = 3$.

5. (1P) This can be done with Gaussian elimination:
$$C - 3 \cdot \mathbb{1} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \Leftrightarrow \begin{pmatrix} -1 & 1 \\ 0 & 0 \end{pmatrix}.$$
This implies that the eigenvector is given by $v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

6. (0P) $T = (v) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

7. (1P) Compute the first principal components for each data point:
$$z^1 = T^T y^1 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = 4,$$
$$z^2 = T^T y^2 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} -2 \\ 0 \end{pmatrix} = -2,$$
$$z^1 = T^T y^1 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ -2 \end{pmatrix} = -2,$$
$$z^1 = T^T y^1 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0.$$

alternative solution without centralizing (2 points less)

3. (2P) Compute covariance matrix

$$C = \frac{1}{4}\sum_{i=1}^{4} x^i (x^i)^T = \frac{1}{4}\left(\begin{pmatrix} 9 & 6 \\ 6 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right) = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

4. (2P) Compute eigenvalues and eigenvectors. First, compute the roots of the characteristic polynomial of C:

$$\chi_C(\lambda) = (\lambda - 3)(\lambda - 2) - 1 = \lambda^2 - 4\lambda + 3.$$

Using the quadratic formula we can calculate the eigenvalues $\lambda = \frac{5+\sqrt{5}}{2}$ and $\lambda = \frac{5-\sqrt{5}}{2}$. Therefore, the largest eigenvalue is $\lambda = \frac{5+\sqrt{5}}{2}$ and since we want to compute one principal component it is enough to calculate the eigenvector for $\lambda = \frac{5+\sqrt{5}}{2}$.

5. (1P) This can be done with Gaussian elimination:

$$C - \frac{5+\sqrt{5}}{2} \cdot \mathbb{1} = \begin{pmatrix} \frac{1-\sqrt{5}}{2} & 1 \\ 1 & \frac{-1-\sqrt{5}}{2} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \frac{1-\sqrt{5}}{2} & 1 \\ 0 & 0 \end{pmatrix}.$$

This implies that the eigenvector is given by $v = \begin{pmatrix} -1 \\ \frac{1-\sqrt{5}}{2} \end{pmatrix}$.

6. (0P) $T = \begin{pmatrix} v \end{pmatrix} = \begin{pmatrix} -1 \\ \frac{1-\sqrt{5}}{2} \end{pmatrix}$

7. (1P) Compute the first principal components for each data point:

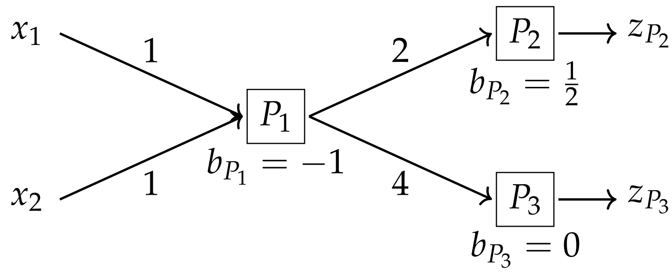
$$z^1 = T^T y^1 = \begin{pmatrix} -1 & \frac{1-\sqrt{5}}{2} \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = -2 - \sqrt{5},$$

$$z^2 = T^T y^2 = \begin{pmatrix} -1 & \frac{1-\sqrt{5}}{2} \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} = 1,$$

$$z^1 = T^T y^1 = \begin{pmatrix} -1 & \frac{1-\sqrt{5}}{2} \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = -2 + \sqrt{5},$$

$$z^1 = T^T y^1 = \begin{pmatrix} -1 & \frac{1-\sqrt{5}}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -1.$$

Question 2. (10+2=12 points)
Consider the following network with 3 neurons P_1, P_2, P_3



with initial weights (as denoted in the graph)

$$w_{P_1x_1} = 1,\ w_{P_1x_2} = 1,\ w_{P_2P_1} = 2,\ w_{P_3P_1} = 4$$

initial biases (as denoted in the graph) $b_{P_1} = -1$, $b_{P_2} = \frac{1}{2}$, $b_{P_3} = 0$, and activation functions

$$\psi_{P_1}(t) = \frac{1}{1+3^{-t}},\ \ \psi_{P_2}(t) = t^2,\ \ \psi_{P_3}(t) = t^2.$$

You may use without proof that the derivative of ψ_{P_1} is given by

$$\psi'_{P_1}(t) \approx \psi_{P_1}(t)(1-\psi_{P_1}(t)).$$

Let

$$\theta = (w_{P_1x_1}, w_{P_1x_2}, w_{P_2P_1}, w_{P_3P_1}, b_{P_1}, b_{P_2}, b_{P_3})^T$$

and let $f_\theta(x) = (z_{P_2}, z_{P_3})^T \in \mathbb{R}^2$ denote the output of the network using parameters θ and input $x = (x_1, x_2)^T \in \mathbb{R}^2$. Consider the loss function $C(\theta; x, y) = \frac{1}{2}||f_\theta(x) - y||^2$ for a given training pair (x, y).

a) Perform one training iteration using the input data $x^1 = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$, $y^1 = \begin{pmatrix} 3 \\ 8 \end{pmatrix}$, and step size $\eta = 0.1$. State the updated weights and biases.

b) Assume that you are given a second point (x^2, y^2) with

$$\nabla C(\theta; x^2, y^2) = (6, 2, 1, 5.5, 10, -4, 2)^T.$$

What are the updated weights and biases if you use both points and the mean squared loss function in the first training iteration instead of only using x^1 as in a) (again with stepsize $\eta = 0.1$)?

Solution Question 2:

a) We start computing the layers' outputs using a forward pass (3P in total):

$$\begin{aligned}
a_{P_1} &= w_{P_1x_1} \cdot x_1 + w_{P_1x_2} \cdot x_2 + b_{P_1} = 1 \cdot (-1) + 1 \cdot 3 - 1 = 1 \ (0.5P) \\
z_{P_1} &= \psi_{P_1}(a_{P_1}) = \frac{1}{1+3^{-1}} = \frac{1}{1+\frac{1}{3}} = \frac{3}{4} \ (1P) \\
a_{P_2} &= w_{P_2P_1} \cdot z_{P_1} + b_{P_2} = 2 \cdot \frac{3}{4} + \frac{1}{2} = 2 \ (0.5P) \\
z_{P_2} &= \psi_{P_2}(a_{P_2}) = 2^2 = 4 \\
a_{P_3} &= w_{P_3P_1} \cdot z_{P_1} + b_{P_3} = 4 \cdot \frac{3}{4} + 0 = 3 \ (0.5P) \\
z_{P_3} &= \psi_{P_3}(a_{P_3}) = 3^2 = 9. \ (z_{P_2} + z_{P_3} : 0.5P)
\end{aligned}$$

We need the derivatives of the activation functions:

$$\begin{aligned}
\psi'_{P_1}(t) &\approx \psi_{P_1}(t)(1 - \psi_{P_1}(t)) \\
\psi'_{P_2}(t) &= \psi'_{P_3}(t) = 2t.
\end{aligned}$$

Now, we compute the partial derivatives of the loss function C with respect to all elements in θ using backpropagation (5P in total):

$$\begin{aligned}
\frac{\partial C}{\partial b_{P_2}} &= (z_{P_2} - y_1) \cdot \psi'_{P_2}(a_{P_2}) = (4-3) \cdot 2 \cdot 2 = 4 \ (1P) \\
\frac{\partial C}{\partial b_{P_3}} &= (z_{P_3} - y_2) \cdot \psi'_{P_3}(a_{P_3}) = (9-8) \cdot 2 \cdot 3 = 6 \ (1P) \\
\frac{\partial C}{\partial b_{P_1}} &= \left(\frac{\partial C}{\partial b_{P_2}} \cdot w_{P_2P_1} + \frac{\partial C}{\partial b_{P_3}} \cdot w_{P_3P_1} \right) \cdot \psi'_{P_1}(a_{P_1}) \\
&= (4 \cdot 2 + 6 \cdot 4) \cdot z_{P_1} \cdot (1 - z_{P_1}) \\
&= (8+24) \cdot \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) = 32 \cdot \frac{3}{4} \cdot \frac{1}{4} = 6 \ (1P) \\
\frac{\partial C}{\partial w_{P_2P_1}} &= \frac{\partial C}{\partial b_{P_2}} \cdot z_{P_1} = 4 \cdot \frac{3}{4} = 3 \ (0.5P) \\
\frac{\partial C}{\partial w_{P_3P_1}} &= \frac{\partial C}{\partial b_{P_3}} \cdot z_{P_1} = 6 \cdot \frac{3}{4} = 4,5 \ (0.5P) \\
\frac{\partial C}{\partial w_{P_1x_1}} &= \frac{\partial C}{\partial b_{P_1}} \cdot x_1 = 6 \cdot (-1) = -6 \ (0.5P) \\
\frac{\partial C}{\partial w_{P_1x_2}} &= \frac{\partial C}{\partial b_{P_1}} \cdot x_2 = 6 \cdot 3 = 18. \ (0.5P)
\end{aligned}$$

Therefore, the gradient reads:

$$\nabla C(\theta) = \begin{pmatrix} -6 \\ 18 \\ 3 \\ 4,5 \\ 6 \\ 4 \\ 6 \end{pmatrix}.$$

We update the parameters with a gradient step (1P):

$$\theta^{new} = \theta - \eta \nabla C(\theta) = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 4 \\ -1 \\ 0.5 \\ 0 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} -6 \\ 18 \\ 3 \\ 4,5 \\ 6 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 1,6 \\ -0,8 \\ 1,7 \\ 3,55 \\ -1,6 \\ 0,1 \\ -0,6 \end{pmatrix}.$$

The updated parameters are therefore given by (1P, aber wenn Zahlen oben stimmen auch)

$$w_{P_1x_1}^{new} = 1,6 \quad w_{P_1x_2}^{new} = -0,8 \quad w_{P_2P_1}^{new} = 1,7 \quad w_{P_3P_1}^{new} = 3,55$$
$$b_{P_1}^{new} = -1,6 \quad b_{P_2}^{new} = 0,1 \quad b_{P_3}^{new} = -0,6.$$

b) Using a second data point we have to compute the averaged gradient $\overline{\nabla} C(\theta)$ and use it for the update step. The averaged gradient is given by (1P)

$$\overline{\nabla} C(\theta) = \frac{1}{2}(\nabla(\theta, x^1, y^1) + \nabla(\theta, x^2, y^2)) = \frac{1}{2} \begin{pmatrix} 0 \\ 20 \\ 4 \\ 10 \\ 16 \\ 0 \\ 8 \end{pmatrix} = \begin{pmatrix} 0 \\ 10 \\ 2 \\ 5 \\ 8 \\ 0 \\ 4 \end{pmatrix}.$$

The update is given by (1P):

$$\theta^{new} = \theta - \eta \overline{\nabla} C(\theta) = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 4 \\ -1 \\ 0.5 \\ 0 \end{pmatrix} - 0.1 \cdot \begin{pmatrix} 0 \\ 10 \\ 2 \\ 5 \\ 8 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1,8 \\ 3,5 \\ -1,8 \\ 0,5 \\ -0,4 \end{pmatrix}.$$

The updated parameters are therefore given by

$$w_{P_1x_1}^{new} = 1 \quad w_{P_1x_2}^{new} = 0 \quad w_{P_2P_1}^{new} = 1,8 \quad w_{P_3P_1}^{new} = 3,5$$
$$b_{P_1}^{new} = -1,8 \quad b_{P_2}^{new} = 0,5 \quad b_{P_3}^{new} = -0,4.$$

Question 3. (3+3+4+3=13 points)

a) Show or give a counterexample: the output of the k-means algorithm for $k = 2$ is the same clustering for any initial cluster centers.

b) Write a short paragraph on the analysis of the convergence of stochastic gradient descent. Cover at least three main ingredients of the analysis.

Let $N, p \in \mathbb{N}$, $X \in \mathbb{R}^{N\times p}$, and $0 \neq Y \in \mathbb{R}^N$. Consider the corresponding linear regression problem minimizing the squared error

$$\min_{\beta\in\mathbb{R}^p} ||X\beta - Y||_2^2. \tag{1}$$

c) Prove or disprove: Assume that we add the column $X_{p+1} = \sum_{i=1}^{p} X_i$ (where X_i are the columns of X) to X to obtain $\tilde{X} = (X, X_{p+1})$. Then, the minimal squared error does not change, i.e.,

$$\min_{\beta\in\mathbb{R}^p} ||X\beta - Y||_2^2 = \min_{\tilde{\beta}\in\mathbb{R}^{p+1}} ||\tilde{X}\tilde{\beta} - Y||_2^2.$$

d) Prove or disprove: If the optimal solution β to (1) is unique, then $N \geq p$.

Solution Question 3:

a) Wrong: any counterexample where different centers give different clusterings, for example in a triangle. (3P if counterexample, 1-2P if explanation but no counterexample, 0P if only "wrong")

b) 1P for any of those: $E(g(x)| = \nabla\mathcal{L}$, $\mathcal{L}$ L-cont, $\sum\eta = \infty$ and $\sum\eta^2 < \infty$, subsequence converges to local minimum ...

c) True: Let β^* be optimal for the left. Then, $\tilde{beta} = (\beta^*, 0)^T$ (1P) is feasible for the right. $\tilde{\beta}$ yields same objective value for the right and thus $\geq$ (1P).

   Let $\tilde{\beta}^*$ be optimal for the right. Then, β given by $\beta_i = \tilde{\beta}_i^* + \tilde{\beta}_{p+1}^*$ (1P) is feasible for the left. It has same objective value and thus $\leq$ (1P).

d) True: The optimal solution satisfies $\nabla = 0$ and thus $X^TY - X^TX\beta = 0$. If X^TX has full rank, the solution is uniquely given by $\beta = (X^TX)^{-1}X^TY$ (1P). X^TX has full rank if X has full column rank (1P). If X has full column rank, then $N \geq p$ (1P).

Question 4. (2+5=7 points)
Consider a list of samples $(x^i, y^i)_{i=1}^N$ with $x^i \in \mathbb{R}^p$ and $y^i \in \{-1, 1\}$ and the soft margin SVM problem

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}, z \in \mathbb{R}^N} \quad & F(\beta, \beta_0, z) = \frac{1}{N} \sum_{i=1}^N z_i + \frac{1}{2} ||\beta||_2^2 \\ \text{s.t.} \quad & 1 - y^i(\beta^T x^i + \beta_0) \le z_i, \qquad i = 1, ..., N, \\ & z_i \ge 0, \qquad i = 1, ..., N. \end{aligned} \tag{2}$$

Note that $(0, 0, \mathbb{1}_N)$, where $\mathbb{1}_N \in \mathbb{R}^N$ denotes the vector containing only ones, is always feasible for (2).

a) Show that the objective function F is bounded from below on the feasible set of (2).

b) Assume that samples from both classes exist. Show that we can bound the feasible set of (2) (i.e., all variables) without changing the optimal value.

Solution Question 4:

a) $z_i \ge 0$, therefore $F(\beta, \beta_0, z) \ge 0$ for all (β, β_0, z). (2P)

b) $(0, 0, \mathbb{1}_N)$ is feasible and thus, for an optimal solution (β, β_0, z) we know that $F(\beta, \beta_0, z) \le 1$ (1P). This implies $0 \le z \le N \cdot \mathbb{1}_N$ (1P) and $\frac{1}{2}||\beta||_2^2 \le 1$ (1P).

   There is a point with $y^i = 1$ and thus

$$\beta_0 \ge -\beta^T x^i + 1 - z_i \ge -2||x^i||_2 + 1 - N$$

   (0.5P for $y^i = 1$ and the constraint, 0.5P for inserting the other bounds). There is a point with $y^i = -1$ and thus

$$\beta_0 \le -\beta^T x^i - 1 + z_i \le 2||x^i||_2 - 1 + N$$

   (0.5P for $y^i = -1$ and the constraint, 0.5P for inserting the other bounds).