Queen Mary
University of London

Main Examination period 2020 – May/June – Semester B
Online Alternative Assessments

MTH793P: Advanced machine learning

You should attempt ALL questions. Marks available are shown next to the questions.

In completing this assessment, you may use books, notes, and the internet. You may use calculators and computers, but you should show your working for any calculations you do. You must not seek or obtain help from anyone else.

At the start of your work, please **copy out and sign** the following declaration:

> I declare that my submission is entirely my own, and I have not sought or obtained help from anyone else.

All work should be **handwritten**, and should **include your student number**.

You have **24 hours** in which to complete and submit this assessment. When you have finished your work:

- scan your work, convert it to a **single PDF file** and upload this using the upload tool on the QMplus page for the module;
- e-mail a copy to maths@qmul.ac.uk with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

You are not expected to spend a long time working on this assessment. We expect you to spend about **3 hours** to complete the assessment, plus the time taken to scan and upload your work. Please try to upload your work well before the end of the assessment period, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

Examiners: M. Benning

 Continue to next page

The notation log refers to the natural logarithm. The set of all natural numbers (starting from one) is denoted by $\mathbb{N}$. The function $\text{rank}(L)$ returns the rank of a matrix L. All computations should be done by hand where possible, with marks being awarded for intermediate steps in order to discourage computational aids.

Question 1 [36 marks].

(a) Compute the expected value $\mathbb{E}_x$ of a (discrete) Poisson-distributed random variable X with probability

$$p_x := \exp(-\lambda)\frac{\lambda^x}{x!}, \quad x = 1, 2, \ldots, s$$

for a constant $\lambda > 0$. What is the solution for $s \to \infty$? **Hint**: Make use of the identity $\exp(\lambda) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$. **[6]**

(b) For a uniform (and absolutely continuous) random variable X in $[0, 1]$ compute the expectation of $f(X)$ for

$$f(x) := \begin{cases} -\log(x) & x \in [0, 1/d] \\ 0 & \text{otherwise} \end{cases},$$

where d is the maximum of the last digit of your student ID and 1. Make use of the convention $0\log(0) = 0$. **[6]**

(c) Let X be a random variable with expectation μ and variance σ^2. Show that the variance of $aX + b$, where $a, b \in \mathbb{R}$, is

$$\text{Var}_x[ax + b] = a^2\sigma^2 .$$

[6]

(d) Verify that the gradient of the function $J(x) := \frac{1}{2}\langle Qx, x\rangle$, where $Q \in \mathbb{R}^{n\times n}$ is a (square) matrix, is $\nabla J(x) = \frac{1}{2}(Q + Q^\top)x$. What does the gradient simplify to if Q is also symmetric? **[6]**

(e) Compute the Bregman distance with respect to the function $J(x) = \frac{1}{2}\langle Qx, x\rangle$, where $Q \in \mathbb{R}^{n\times n}$ is a (square) matrix. **[6]**

(f) If Q in Question 1(e) is a symmetric, positive semi-definite matrix, the function J is guaranteed to be convex for all arguments. What does this imply for the corresponding Bregman distance? **[6]**

 Continue to next page

Solution:

(a) The expectation for a discrete Poisson-distributed random variable X reads

$$\begin{aligned}\mathbb{E}_x[x] &= \sum_{x=1}^{s} x p_x = \sum_{x=1}^{s} x \exp(-\lambda)\frac{\lambda^x}{x\,!} \\ &= \lambda \exp(-\lambda) \sum_{x=1}^{s} \frac{\lambda^{x-1}}{(x-1)\,!} = \lambda \exp(-\lambda) \sum_{x=0}^{s-1} \frac{\lambda^x}{x\,!}\,.\end{aligned}$$

Taking the limit $s \to \infty$ therefore yields

$$\begin{aligned}\mathbb{E}_x[x] &= \lambda \exp(-\lambda) \sum_{x=0}^{\infty} \frac{\lambda^x}{x\,!} \\ &= \lambda \exp(-\lambda) \exp(\lambda) = \lambda\,.\end{aligned}$$

This is a new exercise, which can easily be computed with help of the lecture notes.

(b) For $d = 5$ we compute

$$\begin{aligned}\mathbb{E}_x[f(x)] &= \int_0^1 f(x)\,dx = -\int_0^{\frac{1}{5}} \log(x)\,dx = -\left[x\log x - x\right]_0^{\frac{1}{5}} \\ &= \frac{1}{5} - \frac{1}{5}(\log(1) - \log(5)) = \frac{1}{5}\,(1 + \log(5)) \approx 0.5218875825\,.\end{aligned}$$

This exercise is similar to Exercise 1, Coursework 3

(c) With the definition of the variance we compute

$$\begin{aligned}\mathrm{Var}_x[ax+b] &= \mathbb{E}_x\left[(ax + b - \mathbb{E}_x[ax+b])^2\right] \\ &= \mathbb{E}_x\left[(ax + b - a\mathbb{E}_x[x] + b)^2\right] \\ &= \mathbb{E}_x\left[(ax - a\mathbb{E}_x[x])^2\right] \\ &= \mathbb{E}_x\left[a^2\,(x - \mathbb{E}_x[x])^2\right] \\ &= a^2\mathbb{E}_x\left[(x - \mathbb{E}_x[x])^2\right] \\ &= a^2\,\mathrm{Var}_x[x] \\ &= a^2\sigma^2\,.\end{aligned}$$

This question is similar to a question in the (yet unpublished) mock exam.

(d) For the function $J(x) := \frac{1}{2}\langle Qx, x\rangle$ we observe

$$J(x) = \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} q_{ij}x_j\right)x_i\,,$$

where q_{ij} are the entries of Q. Computing the partial derivatives w.r.t. x_l therefore yields

$$\frac{\partial}{\partial x_l} J = \frac{1}{2}\sum_{j=1}^{n} q_{lj}x_j + \frac{1}{2}\sum_{i=1}^{n} q_{il}x_i\,.$$

In matrix-form we therefore have

$$\nabla J(x) = \frac{1}{2}Qx + \frac{1}{2}Q^\top x\,.$$

If Q is symmetric, the gradient simplifies to

$$\nabla J(x) = Qx = Q^\top x\,.$$

This question can be answered based on basic calculus.

(e) For the stated function we compute $\nabla J(y) = \frac{1}{2}(Q + Q^\top)y$ and therefore

$$\begin{aligned}
D_J(x,y) &= \frac{1}{2}\langle Qx, x\rangle - \frac{1}{2}\langle Qy, y\rangle - \left\langle \frac{1}{2}(Q+Q^\top)y, x-y\right\rangle \\
&= \frac{1}{2}\Big(\langle Qx, x\rangle - \langle Qy, y\rangle - \langle Qy, x-y\rangle + \langle Q^\top y, x-y\rangle\Big) \\
&= \frac{1}{2}\left(\langle Qx, x\rangle - \langle Qy, y\rangle - \langle Qy, x-y\rangle - \langle y, Q(x-y)\rangle\right) \\
&= \frac{1}{2}\left(\langle Qx, x\rangle - \langle Qy, y\rangle - \langle Qy, x-y\rangle - \langle y, Qx\rangle + \langle y, Qy\rangle\right) \\
&= \frac{1}{2}\left(\langle Qx, x\rangle - \langle Qy, x-y\rangle - \langle y, Qx\rangle\right) \\
&= \frac{1}{2}\left(\langle Qx, x-y\rangle - \langle Qy, x-y\rangle\right) \\
&= \frac{1}{2}\langle Q(x-y), x-y\rangle
\end{aligned}$$

This exercise is similar to Exercise 2, Coursework 3

(f) It implies that the Bregman distance $D_J(x,y)$ is non-negative for all $x, y \in \mathbb{R}^n$, i.e. $D_J(x,y) \geq 0$ for all $x, y \in \mathbb{R}^n$.

This question can be answered based on the lecture notes content.

 Continue to next page

Question 2 [34 marks].

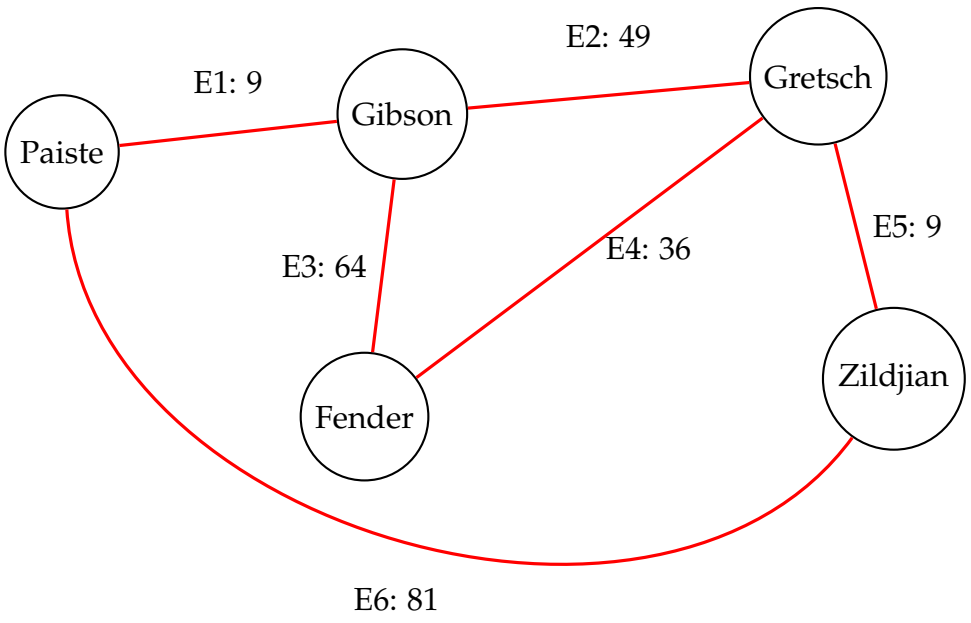
(a) You want to decide whether or not to become a data scientist. You base your decision on three factors:

- Am I excited about machine learning?
- Does becoming a data scientist involve understanding complicated mathematics?
- Do companies likely hire data scientists?

Suppose you do not mind if becoming a data scientist involves understanding complicated mathematics as long as many companies likely hire data scientists. However, you really would not want to become a data scientist if machine learning does not excite you.

Model this binary decision process with a perceptron and choose some appropriate weights to mimic the decision process accurately. **[7]**

(b) Write down the incidence matrix for the following weighted, undirected graph:



Order the columns of the incidence matrix alphabetically according to the vertex name and the rows according to the edge numbering (E1, E2, E3, ...). **[7]**

(c) Compute the corresponding graph Laplacian for the incidence matrix in Question 2(b). **[6]**

(d) We want to use the graph from Question 2(b) to determine whether a node in the graph belongs to the class "guitars" or the class "cymbals". Suppose we are in a semi-supervised setting, where the node "Fender" is already labelled $v_{\text{Fender}} = 1$ (class "guitars") and the node "Zildjian" is labelled as $v_{\text{Zildjian}} = 0$ (class "cymbals"). Determine the labels for all remaining nodes, and classify each node. **[8]**

(e) Determine manually some parameters $w \in \mathbb{R}^2$ and $b \in \mathbb{R}$ of a neural network of the form

$$f(x_1, x_2) = \max\left(0, w^\top \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + b\right)$$

that is supposed to mimic the logical AND function, i.e.

x_1	x_2	$f(x_1, x_2)$
0	0	0
1	0	0
0	1	0
1	1	1

.

[6]

Solution:

(a) This binary decision process can for example be modelled with the following perceptron:

$$f(x_1, x_2, x_3) = \begin{cases} 0 & 5x_1 - x_2 + 3x_3 \leq 4 \\ 1 & 5x_1 - x_2 + 3x_3 > 4 \end{cases},$$

where $x_1, x_2, x_3 \in \{0, 1\}$ represent variables associated to the three factors mentioned in the problem description. Different states are represented as follows:

$f(x_1, x_2, x_3)$	x_1	x_2	x_3
1	1	1	1
0	0	1	1
1	1	0	1
0	1	1	0
1	1	0	0
0	0	1	0
0	0	0	1
0	0	0	0

Hence, one would never become a data scientist if one finds machine learning not exciting. If one finds machine learning exciting but becoming a data scientist also involves complicated mathematics one will only become a data scientist if companies likely hire data scientists.

This exercise is similar to Exercise 3 on Coursework 3

(b) The incidence matrix for the displayed graph is

$$M_w = \begin{pmatrix} \text{E1} & 0 & -3 & 0 & 3 & 0 \\ \text{E2} & 0 & -7 & 7 & 0 & 0 \\ \text{E3} & -8 & 8 & 0 & 0 & 0 \\ \text{E4} & -6 & 0 & 6 & 0 & 0 \\ \text{E5} & 0 & 0 & -3 & 0 & 3 \\ \text{E6} & 0 & 0 & 0 & -9 & 9 \\ \hline & \text{Fender} & \text{Gibson} & \text{Gretsch} & \text{Paiste} & \text{Zildjian} \end{pmatrix}$$

This question is similar to Exercise 2 on Coursework 4.

(c) The corresponding graph Laplacian then reads

$$L_w = M_w^\top M_w = \begin{pmatrix} \text{Fender} & 100 & -64 & -36 & 0 & 0 \\ \text{Gibson} & -64 & 122 & -49 & -9 & 0 \\ \text{Gretsch} & -36 & -49 & 94 & 0 & -9 \\ \text{Paiste} & 0 & -9 & 0 & 90 & -81 \\ \text{Zildjian} & 0 & 0 & -9 & -81 & 90 \\ \hline & \text{Fender} & \text{Gibson} & \text{Gretsch} & \text{Paiste} & \text{Zildjian} \end{pmatrix}$$

This question is similar to Exercise 2 on Coursework 4.

 Continue to next page

(d) From the lecture notes we know that the label vector $v \in \mathbb{R}^5$ can be decomposed as

$$v = P_{R^\perp}^\top \tilde{v} + P_R^\top y\,,$$

where P_R denotes the projection of v onto the known indices, and $P_{R^\perp}$ onto the unknown indices. The known indices are denoted by y, the unknown by $\tilde{v}$. For

$$v = \begin{pmatrix} v_{\text{Fender}} \\ v_{\text{Gibson}} \\ v_{\text{Gretsch}} \\ v_{\text{Paiste}} \\ v_{\text{Zildjian}} \end{pmatrix}$$

we know the first and the last entry; the first belongs to the class "guitars" and therefore takes on the value $v_{\text{Fender}} = 1$, whereas the last entry belongs to the class "cymbals", hence $v_{\text{Zildjian}} = 0$. Thus, for $y = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top$ we have

$$v = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \tilde{v} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

From the lecture notes we also know that we can estimate $\tilde{v}$ via

$$\begin{aligned} \tilde{v} &= \arg\min_{\overline{v}} \left\| M_w \left(P_{R^\perp}^\top \overline{v} + P_R^\top y \right) \right\|^2, \\ &= -\left(P_{R^\perp} L_w P_{R^\perp}^\top \right)^{-1} \left(P_{R^\perp} L_w P_R^\top y \right), \end{aligned}$$

which for our matrices reads

$$\begin{pmatrix} 122 & -49 & -9 \\ -49 & 94 & 0 \\ -9 & 0 & 90 \end{pmatrix} \tilde{v} = \begin{pmatrix} 64 \\ 36 \\ 0 \end{pmatrix},$$

Solving this linear system leads to the (approximate) solution

$$\tilde{v} \approx \begin{pmatrix} 0.8661 \\ 0.8345 \\ 0.0666 \end{pmatrix}.$$

Rounding all values above $1/2$ to one and below $1/2$ to zero then yields the classification

$$v = \begin{pmatrix} v_{\text{Fender}} \\ v_{\text{Gibson}} \\ v_{\text{Gretsch}} \\ v_{\text{Paiste}} \\ v_{\text{Zildjian}} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

This question is similar to Question 3 of the mock exam.

(e) A possible choice of weights w and bias b is

$$w = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad b = -1\,.$$

This way we obtain $f(0,0) = \max(0,-1) = 0$, $f(1,0) = \max(0,0) = 0$, $f(0,1) = \max(0,0) = 0$ and $f(1,1) = \max(0,1) = 1$. I will accept any weights and biases as correct answers that yield $f(0,0) = 0$, $f(1,0) = 0$, $f(0,1) = 0$ and $f(1,1) = 1$.

This question is similar to Exercise 1 of Coursework 4.

Question 3 [30 marks].

(a) Perform two steps of k-means clustering by hand for the six data points $x_1 = 1$, $x_2 = 3$, $x_3 = 0$, $x_4 = 15$, and $x_5 = 17$. Assume $k = 2$ clusters and initialise your variables as

$$z_0 := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}^\top ,$$

and

$$\mu_0 := \begin{pmatrix} d \\ 0 \end{pmatrix} ,$$

where d is the eighth digit of your student ID number. For each iteration, update the variable z^l first, and then μ^l. Here $l \in \{1, 2\}$ denotes the iteration index. Did the iteration converge? **[8]**

(b) Complete the following matrix such that it has minimal rank:

$$\begin{pmatrix} 1 & -2 & d \\ -3 & 6 & ? \end{pmatrix} .$$

Here d is the maximum of the seventh digit of your student ID number and 1. Justify your choice. **[6]**

(c) Compute by hand an approximation $\hat{L} \in \mathbb{R}^{2\times 3}$ with $\operatorname{rank}(\hat{L}) = 1$ of the matrix

$$X := d \begin{pmatrix} 6 & 4 & 4 \\ 4 & 6 & -4 \end{pmatrix} ,$$

where d is the maximum of the last digit of your student ID and 1, that satisfies $\|\hat{L} - X\|_{\text{Fro}} \leq \|L - X\|_{\text{Fro}}$, for all $L \in \mathbb{R}^{2\times 3}$ with $\operatorname{rank}(L) = 1$. **[8]**

(d) Formulate a projected (or proximal) gradient descent algorithm for a matrix completion problem of the form

$$\hat{L} = \arg\min_{L\in\mathbb{R}^{s\times n}} \left\{ \frac{1}{2}\|P_\Omega L - y\|_{\text{Fro}}^2 \quad \text{subject to} \quad \operatorname{rank}(L) \leq k \right\} .$$

Here $P_\Omega \in \mathbb{R}^{s\times n} \to \mathbb{R}^r$ is a (known) projection operator that projects the known entries, specified by the index set Ω, of its argument to a vector of length r. The vector of known entries is denoted by $y \in \mathbb{R}^r$ and $k \in \mathbb{N}$ is a fixed constant that determines the rank of $\hat{L}$. What does the proximal mapping (in the proximal gradient descent) look like? What is its closed-form solution? **[8]**

Solution:

(a) The update formulae for k-means clustering are

$$z_{ik}^{l+1} = \begin{cases} 1 & k = \arg\min_{j\in\{1,2\}} |x_i - \mu_j^l|^2 \\ 0 & \text{otherwise} \end{cases},$$

and

$$\mu^{l+1} = \frac{\sum_{i=1}^5 z_{ik}^{l+1} x_i}{\sum_{i=1}^5 z_{ik}^{l+1}}.$$

For $d = 9$ we compute

$$\begin{pmatrix} |x_1 - \mu_1^0|^2 & |x_2 - \mu_1^0|^2 & |x_3 - \mu_1^0|^2 & |x_4 - \mu_1^0|^2 & |x_5 - \mu_1^0|^2 \\ |x_1 - \mu_2^0|^2 & |x_2 - \mu_2^0|^2 & |x_3 - \mu_2^0|^2 & |x_4 - \mu_2^0|^2 & |x_5 - \mu_2^0|^2 \end{pmatrix}$$
$$= \begin{pmatrix} 64 & 36 & 81 & 36 & 64 \\ 1 & 9 & 0 & 225 & 289 \end{pmatrix}$$

for the squared differences. Hence, we compute

$$z^1 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix},$$

and, consequently, also

$$\mu^1 = \begin{pmatrix} \frac{\sum_{i=1}^5 z_{i1}^1 x_i}{\sum_{i=1}^5 z_{i1}^1} \\ \frac{\sum_{i=1}^5 z_{i2}^1 x_i}{\sum_{i=1}^5 z_{i2}^1} \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=4}^5 z_{i1}^1 x_i}{\sum_{i=4}^5 z_{i1}^1} \\ \frac{\sum_{i=1}^3 z_{i2}^1 x_i}{\sum_{i=1}^3 z_{i2}^1} \end{pmatrix} = \begin{pmatrix} \frac{15+17}{2} \\ \frac{1+3+0}{3} \end{pmatrix} = \begin{pmatrix} 16 \\ \frac{4}{3} \end{pmatrix}.$$

This completes the first iteration. Computing the squared differences for the second iteration then yields

$$\begin{pmatrix} |x_1 - \mu_1^1|^2 & |x_2 - \mu_1^1|^2 & |x_3 - \mu_1^1|^2 & |x_4 - \mu_1^1|^2 & |x_5 - \mu_1^1|^2 \\ |x_1 - \mu_2^1|^2 & |x_2 - \mu_2^1|^2 & |x_3 - \mu_2^1|^2 & |x_4 - \mu_2^1|^2 & |x_5 - \mu_2^1|^2 \end{pmatrix}$$
$$= \begin{pmatrix} 255 & 169 & 256 & 1 & 1 \\ \frac{1}{9} & \frac{25}{9} & \frac{16}{9} & \frac{1681}{9} & \frac{2209}{9} \end{pmatrix},$$

for which we obtain

$$z^2 = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

Since $z^2 = z^1$, μ^2 will also equal μ^1 and the algorithm has converged.

This question is similar to Exercise 1 on Coursework 5.

(b) The minimal rank for non-zero 2×3-matrix is one. Completing the matrix for $d = 7$ to

$$\begin{pmatrix} 1 & -2 & 7 \\ -3 & 6 & -21 \end{pmatrix}$$

ensures that the entries of the second row are the entries of the first row multiplied by -3. This way both rows are linearly dependent, leading to a matrix of rank one.

This question can be answered based on the lecture notes content.

(c) From the lecture notes we know that the best possible rank-one approximation in terms of the Frobenius norm can be computed by computing the (incomplete) singular value decomposition of X. Like in a similar coursework exercise, we compute the eigenvalues of $XX^\top$ by solving the characteristic polynomial $\det(XX^\top - \lambda I) = 0$, i.e.

$$\begin{aligned}\det(XX^\top - \lambda I) &= \det\left(d\begin{pmatrix}68-\lambda & 32\\ 32 & 68-\lambda\end{pmatrix}\right) = d^2(68-\lambda)^2 - d^2 1024\\ &= d^2\left(\lambda^2 - 136\lambda + 3600\right) = 0\,,\end{aligned}$$

whose solutions are $\lambda_1 = 100d^2$ and $\lambda_2 = 36d^2$. Since the singular values are $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, 2$, we obtain $\sigma_1 = 10d$ and $\sigma_2 = 6d$. The best rank one approximation can be computed by computing $\tilde{X} = u_1 u_1^\top X$, where u_1 is the singular vector that corresponds to σ_1. We determine u_1 by computing the kernel of $XX^\top - \lambda_1 I$, i.e.

$$\ker(XX^\top - \lambda_1 I) = \ker\left(d\begin{pmatrix}-32 & 32\\ 32 & -32\end{pmatrix}\right) = \left\{ t\begin{pmatrix}1\\1\end{pmatrix} \,\middle|\, t \in \mathbb{R}\right\}.$$

Since $u_1 \in \ker(XX^\top - \lambda_1 I)$ has to have norm one, we easily compute

$$u_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}.$$

As a consequence, the best rank-one approximation of X in terms of the Frobenius norm is computed via

$$\begin{aligned}\tilde{X} = u_1 u_1^\top X &= \frac{d}{2}\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}6 & 4 & 4\\ 4 & 6 & -4\end{pmatrix}\\ &= d\begin{pmatrix}5 & 5 & 0\\ 5 & 5 & 0\end{pmatrix}.\end{aligned}$$

This exercise is similar to Exercise 2(1) on Coursework 5.

(d) A proximal or projected gradient descent method for this (non-convex) problem can be formulated by splitting the objective into two parts

$$F(L) := \frac{1}{2}\|P_\Omega L - y\|_{\text{Fro}}^2 \qquad \text{and} \qquad R(L) := \chi_{\text{rank}(\cdot)\le k}(L)\,,$$

where $\chi_{\text{rank}(\cdot)\le k}(L)$ denotes the characteristic function

$$\chi_{\text{rank}(\cdot)\le k}(L) := \begin{cases}0 & \text{rank}(L) \le k\\ \infty & \text{otherwise}\end{cases},$$

and formulating

$$\begin{aligned} L^{j+1} &= (I + \tau \partial R)^{-1} \left(L^j - \tau \nabla F(L^j) \right) \\ &= \left(I + \chi_{\text{rank}(\cdot) \leq k}(L) \right)^{-1} \left(L^j - \tau P_\Omega^\top \left(P_\Omega L^j - y \right) \right) \\ &= \arg \min_{L \in \mathbb{R}^{s \times n}} \left\{ \frac{1}{2} \left\| L - \left(L^j - \tau P_\Omega^\top \left(P_\Omega L^j - y \right) \right) \right\|_{\text{Fro}}^2 \text{ subject to } \text{rank}(L) \leq k \right\}, \end{aligned}$$

for a step-size parameter $0 < \tau < 1$, and some initial value L^0. Here j denotes the iteration index. The closed-form-solution of the projection / proximal map can be deduced from the lecture notes in form of Theorem 2.1. Then each iteration reads

$$\begin{aligned} Y^j &= L^j - \tau P_\Omega^\top \left(P_\Omega L^j - y \right) \\ \text{Compute SVD} \quad Y^j &= U_j \Sigma_j V_j^\top \\ L^{j+1} &= (U_j)_k (U_j)_k^\top Y^j, \end{aligned}$$

where $(U_j)_k$ denotes the first k-columns of U_j.

This is a new exercise.

End of Paper.