Queen Mary
University of London

Late-Summer Examination period 2021

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 and notes.**
- **You may use calculators and computers, but you must show your working for any calculations you do.**
- **You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.**
- **You must not seek or obtain help from anyone else.**

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

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

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- 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 expected to spend about 2 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 submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

Examiners: M. Benning

The notation log refers to the natural logarithm. The determinant of a matrix is denoted by det. The set of real numbers is denoted by $\mathbb{R}$. All computations should be done by hand where possible, with marks being awarded for intermediate steps in order to discourage computational aids.

Question 1 [38 marks].

(a) Rewrite the function $f(x) = \frac{3x^2-2x+4}{1+dx+2x^2}$ to $g(a) + \varepsilon h(a, b)$, where the argument x is a dual number of the form $x = a + \varepsilon b$ with $\varepsilon^2 = 0$, and specify both g and h. The number d is one added to the last digit of your student ID number. **[8]**

(b) Compute the derivative $f'(a)$ of f as defined in Question 1(a) at argument a by making use of your result of Question 1(a). **[6]**

(c) Consider the function $f(X) = \log(\det(X))$, acting on an invertible matrix $X \in \mathbb{R}^{2\times 2}$. Compute the partial derivatives with the help of dual number calculus, assuming each entry x_{ij} of X is a dual number of the form $a_{ij} + \varepsilon b_{ij}$ with $\varepsilon^2 = 1$, for $i, j \in \{1, 2\}$. Subsequently, show that the entire gradient equals $(X^\top)^{-1}$. **[8]**

(d) Verify that the subdifferential $\partial\|\cdot\|$ of the (non-squared!) Euclidean norm $f(x) := \|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ for any $x \in \mathbb{R}^n$ is characterised via

$$\partial\|x\| = \begin{cases} \left\{\frac{x}{\|x\|}\right\} & x \neq 0 \\ \{p \in \mathbb{R}^n \mid \|p\| \leq 1\} & x = 0 \end{cases} .$$

[8]

(e) Compute the generalised Bregman distance $D_f^p(x, y)$ with respect to the function $f(x) = \|x\|$ for a specific subgradient $p \in \partial f(y)$. Make sure to characterise all different cases. **[8]**

Solution:

(a) We compute

$$\begin{aligned} f(a+\varepsilon b) &= \frac{3(a+\varepsilon b)^2 - 2(a+\varepsilon b) + 4}{1 + d(a+\varepsilon b) + 2(a+\varepsilon b)^2} \\ &= \frac{3a^2 - 2a + 4 + \varepsilon b(6a-2)}{1 + da + 2a^2 + \varepsilon b(d+4a)} \\ &= \frac{u + \varepsilon v}{w + \varepsilon z}, \end{aligned}$$

for $u = 3a^2 - 2a + 4$, $v = b(6a-2)$, $w = 1 + da + 2a^2$ and $z = b(d+4a)$. We further compute

$$\begin{aligned} f(a+\varepsilon b) &= \frac{u + \varepsilon v}{w + \varepsilon z} \\ &= \frac{(u+\varepsilon v)(w-\varepsilon z)}{(w+\varepsilon z)(w-\varepsilon z)} = \frac{uw + \varepsilon(vw - uz)}{w^2} \\ &= \frac{u}{w} + \varepsilon\frac{vw - uz}{w^2} \\ &= \frac{3a^2 - 2a + 4}{1 + da + 2a^2} + \varepsilon\frac{b(6a-2)(1+da+2a^2) - (3a^2 - 2a + 4)b(d+4a)}{(1+da+2a^2)^2} \\ &= \frac{3a^2 - 2a + 4}{1 + da + 2a^2} + \varepsilon\, b\, \frac{(3a^2-4)d + 4a^2 - 10a - 2}{(1+da+2a^2)^2} \\ &= g(a) + \varepsilon h(a,b), \end{aligned}$$

for $g(a) = (3a^2 - 2a + 4)/(1 + da + 2a^2)$ and
$h(a,b) = b((3a^2-4)d + 4a^2 - 10a - 2)/(1 + da + 2a^2)^2$.

This exercise is similar to Exercise 1 on Coursework 10.

(b) From the lecture notes we know that $h(a,1) = f'(a)$. Hence, the derivative of f w.r.t. the argument a is

$$f'(a) = h(a,1) = \frac{(3a^2-4)d + 4a^2 - 10a - 2}{(1+da+2a^2)^2}.$$

This exercise is similar to a Exercise 1 on Coursework 10.

(c) Similar to the previous exercise we conclude

$$\begin{aligned} f(X) = f(A + \varepsilon B) &= \log(\det(A + \varepsilon B)) \\ &= \log\left((a_{11} + \varepsilon b_{11})(a_{22} + \varepsilon b_{22}) - (a_{12} + \varepsilon b_{12})(a_{21} + \varepsilon b_{21})\right) \\ &= \log\left(a_{11}a_{22} - a_{12}a_{21} + \varepsilon\left(a_{11}b_{22} + b_{11}a_{22} - a_{12}b_{21} - b_{12}a_{21}\right)\right) \\ &= \log\left(\det(A) + \varepsilon\left\langle\begin{pmatrix} a_{22} \\ -a_{12} \\ -a_{21} \\ a_{11}\end{pmatrix}, \begin{pmatrix} b_{11} \\ b_{21} \\ b_{12} \\ b_{22}\end{pmatrix}\right\rangle\right) \\ &= \log(\det(A)) + \varepsilon\frac{\left\langle\begin{pmatrix} a_{22} \\ -a_{12} \\ -a_{21} \\ a_{11}\end{pmatrix}, \begin{pmatrix} b_{11} \\ b_{21} \\ b_{12} \\ b_{22}\end{pmatrix}\right\rangle}{\det(A)} \, . \end{aligned}$$

If we specify $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ we can compute the first partial derivative by evaluating f, which reads $a_{22}/\det(A)$. Computing all partial derivatives and combining them in compact matrix-notation yields

$$\frac{1}{a_{11}a_{22} - a_{12}a_{21}}\begin{pmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{22} \end{pmatrix} = (A^\top)^{-1} \, .$$

This exercise is similar to a Exercise 2 in Coursework 10.

(d) The definition of the subdifferential for the Euclidean norm $\|\cdots\|$ reads

$$\partial\|x\| = \{p \in \mathbb{R}^n \mid \|y\| \geq \|x\| + \langle p, y - x\rangle \, , \forall y \in \mathbb{R}^n\} \, .$$

We characterise this subdifferential by case analysis.
Case 1: suppose $x \neq 0$, then $\|x\|$ is differentiable, with partial derivative $x_l/\|x\|$ for $l \in \{1, \ldots, n\}$. We know from the lecture notes that the subdifferential of a differentiable function only contains the gradient. Hence, we have $\partial\|x\| = \{x/\|x\|\}$. **Case 2**: suppose $x = 0$, then we have

$$\partial\|0\| = \{p \in \mathbb{R}^n \mid \|y\| \geq \langle p, y\rangle \, , \forall y \in \mathbb{R}^n\} \, .$$

The inequality $\langle p, y\rangle \leq \|y\|$ is satisfied for every p with $\|p\| \leq 1$ and all y, which can be seen from the Cauchy-Schwartz inequality $\langle p, y\rangle \leq \|p\|\|y\| = \|y\|$. For p with $\|p\| > 1$ we can always choose a $y = p/\|p\|$ for which we observe $\langle p, y\rangle = \|p\| > 1 = \|y\|$, so that the inequality is not satisfied for all y. Hence, the subdifferential reads

$$\partial\|0\| = \{p \in \mathbb{R}^n \mid \|p\| \leq 1\} \, .$$

Combining both cases yields

$$\partial\|x\| = \begin{cases} \left\{\frac{x}{\|x\|}\right\} & x \neq 0 \\ \{p \in \mathbb{R}^n \mid \|p\| \leq 1\} & x = 0 \end{cases} \, .$$

This exercise is similar to Exercise 1.3 on Coursework 5.

(e) The generalised Bregman distance with respect to the function $f(x) = \|x\|$ reads

$$D_f^p(x, y) = \|x\| - \|y\| - \langle p, x - y \rangle \,,$$

for $p \in \partial\|y\|$. From the previous exercise we know

$$\partial\|y\| = \begin{cases} \left\{ \frac{y}{\|y\|} \right\} & y \neq 0 \\ \{p \mid \|p\| \leq 1\} & y = 0 \end{cases} .$$

Hence, for $y \neq 0$ we have $p = y/\|y\|$ and therefore observe

$$D_f^{\frac{y}{\|y\|}}(x, y) = \|x\| - \|y\| - \left\langle \frac{y}{\|y\|}, x - y \right\rangle = \|x\| - \frac{\langle x, y \rangle}{\|y\|} = \frac{\|x\|\|y\| - \langle x, y \rangle}{\|y\|} .$$

For $y = 0$ we have $\partial\|0\| = \{p \mid \|p\| \leq 1\}$ and therefore conclude

$$D_f^p(x, 0) = \|x\| - \langle p, x \rangle \,.$$

Combining both cases yields

$$D_f^p(x, y) = \begin{cases} \frac{\|x\|\|y\| - \langle x, y \rangle}{\|y\|} & y \neq 0 \\ \|x\| - \langle p, x \rangle & y = 0 \end{cases} .$$

This exercise is similar to Exercise 2 on Coursework 10.

Question 2 [31 marks].

(a) Determine parameters $W \in \mathbb{R}^{2\times 2}$, $w \in \mathbb{R}^2$, $b \in \mathbb{R}^2$ and $c \in \mathbb{R}$ of a neural network of the form

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

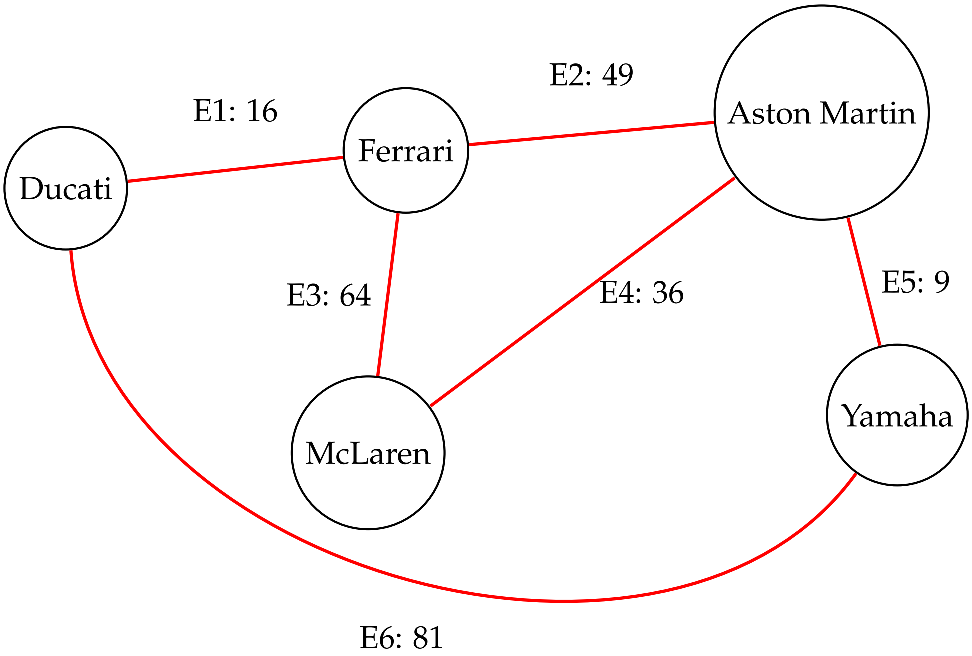
that is the logical XNOR-function, i.e.

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

.

[8]

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



Use the definition of an incidence matrix from the lecture notes and 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). **[8]**

(d) We want to use the graph from Question 2(b) to determine whether a node in the graph belongs to the class "cars" or the class "motorbikes". Suppose we are in a semi-supervised setting, where the node "McLaren" is already labelled $v_{\text{McLaren}} = 1$ (class "cars") and the node "Yamaha" is labelled as $v_{\text{Yamaha}} = 0$ (class "motorbikes"). Determine the remaining labels with the same procedure as described in the lecture notes and classify each node. **[8]**

Solution:

(a) A possible choice of weights W, w and biases b and c is

$$W = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad w = \begin{pmatrix} -1 \\ 2 \end{pmatrix} \quad b = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \quad \text{and } c = 1\,.$$

This way we obtain $f(0,0) = 1$, $f(1,0) = 0$, $f(0,1) = 0$ and $f(1,1) = 1$. 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 3 on Coursework 10.

(b) The incidence matrix for the displayed graph is

$$M_w = \left(\begin{array}{c|ccccc} \text{E1} & 0 & -4 & 4 & 0 & 0 \\ \text{E2} & -7 & 0 & 7 & 0 & 0 \\ \text{E3} & 0 & 0 & -8 & 8 & 0 \\ \text{E4} & -6 & 0 & 0 & 6 & 0 \\ \text{E5} & -3 & 0 & 0 & 0 & 3 \\ \text{E6} & 0 & -9 & 0 & 0 & 9 \\ \hline & \text{Aston Martin} & \text{Ducati} & \text{Ferrari} & \text{McLaren} & \text{Yamaha} \end{array}\right)$$

This question is similar to Exercise 1.1 of Coursework 1.

(c) The corresponding graph Laplacian then reads

$$L_w = M_w^\top M_w = \left(\begin{array}{c|ccccc} \text{A. Martin} & 94 & 0 & -49 & -36 & -9 \\ \text{Ducati} & 0 & 97 & -16 & 0 & -81 \\ \text{Ferrari} & -49 & -16 & 129 & -64 & 0 \\ \text{McLaren} & -36 & 0 & -64 & 100 & 0 \\ \text{Yamaha} & -9 & -81 & 0 & 0 & 90 \\ \hline & \text{A. Martin} & \text{Ducati} & \text{Ferrari} & \text{McLaren} & \text{Yamaha} \end{array}\right)$$

This question is similar to Exercise 1.2 of Coursework 1.

(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

$$v = \begin{pmatrix} v_{\text{Aston Martin}} \\ v_{\text{Ducati}} \\ v_{\text{Ferrari}} \\ v_{\text{McLaren}} \\ v_{\text{Yamaha}} \end{pmatrix}$$

we know the fourth and fifth entry; the fourth belongs to the class "cars" and therefore takes on the value $v_{\text{McLaren}} = 1$, whereas the fifth entry belongs to the class "motorbikes", hence $v_{\text{Yamaha}} = 0$. Thus, for $y = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top$ we have

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

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

$$\begin{pmatrix} -94 & 0 & -49 \\ 0 & -97 & 16 \\ 49 & 16 & -129 \end{pmatrix} \tilde{v} = \begin{pmatrix} -36 \\ 0 \\ -64 \end{pmatrix}\,,$$

Solving this linear system leads to the (approximate) solution

$$\tilde{v} \approx \begin{pmatrix} 0.8109 \\ 0.1354 \\ 0.8209 \end{pmatrix}\,.$$

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

$$v = \begin{pmatrix} v_{\text{Aston Martin}} \\ v_{\text{Ducati}} \\ v_{\text{Ferrari}} \\ v_{\text{McLaren}} \\ v_{\text{Yamaha}} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}\,.$$

This question is similar to Question 1.3 of Coursework 1.

Question 3 [31 marks].

(a) Perform k-means clustering by hand for the five data points $x_1 = \begin{pmatrix} -3 \\ 7 \end{pmatrix}$, $x_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $x_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $x_4 = \begin{pmatrix} -5 \\ 6 \end{pmatrix}$, and $x_5 = \begin{pmatrix} -4 \\ 5 \end{pmatrix}$. Assume $k = 2$ clusters and initialise your centroids as

$$\mu_1^0 := \begin{pmatrix} 0 \\ d \end{pmatrix} \qquad \text{and} \qquad \mu_2^0 := \begin{pmatrix} d \\ 0 \end{pmatrix} ,$$

where d is one added to the eighth digit of your student ID number. For each iteration, update the assignments first, and then the centroids. Perform as many iterations as are required for the k-means clustering algorithm to converge. **[8]**

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

$$X = \begin{pmatrix} d & ? & 1 & 7 \\ ? & 6 & -3 & ? \\ ? & -12 & ? & 42 \end{pmatrix} .$$

Here d is the maximum of the seventh digit of your student ID number and 1. Depending on the rank, find a representation $UV^\top = X$ with suitable matrices U and V. **[6]**

(c) Show that for vectors $x, y \in \mathbb{R}^k$ and a function $g : \mathbb{R}^k \to \mathbb{R}^k$ the vector $z \in \mathbb{R}^k$ defined as

$$z_i = \frac{x_i \exp\left(-g(y_i)\right)}{\sum_{j=1}^k x_j \exp\left(-g(y_j)\right)} = \operatorname{softmax}\left(-\log(x)g(y)\right)_i ,$$

for all $i \in \{1, \ldots, k\}$, is the solution of the optimisation problem

$$z = \arg\min_{\tilde{z} \in \mathbb{R}^k} \left\{ D_f(\tilde{z}, x) + \langle \tilde{z}, g(y) \rangle \quad \text{subject to} \quad \tilde{z} \in [0,1]^k , \text{ and } \sum_{j=1}^k \tilde{z}_j = 1 \right\} ,$$

where D_f denotes the Bregman distance with respect to the convex function $f(z) := \sum_{j=1}^k z_j \log(z_j)$.
Hint: reformulate the unconstrained objective $D_f(\tilde{z}, x) + \langle \tilde{z}, g(y) \rangle$ to $D_f(\tilde{z}, z) + c\left(1 - \sum_{j=1}^k \tilde{z}_j\right) + d$ for constants c and d independent of $\tilde{z}$. **[8]**

(d) Design a (Bregman) proximal gradient descent algorithm for the solution of the convex optimisation problem

$$\hat{x} = \arg\min_{x \in \mathbb{R}^k} \left\{ h(x) \quad \text{subject to} \quad x \in [0,1]^k , \text{ and } \sum_{j=1}^k x_j = 1 \right\} .$$

Here $h : \mathbb{R}^k \to \mathbb{R}$ is a convex and continuously differentiable function. **Hint**: make use of Question 3(c). **[9]**

Solution:

(a) We show the solution for $d = 3$; other solutions, however, lead to the same results in terms of assignments and centroids. We compute the Euclidean distances of the data points with respect to the initial centroids:

$$\begin{pmatrix} 5 & 2 & \sqrt{17} & \sqrt{34} & \sqrt{20} \\ \sqrt{85} & \sqrt{10} & \sqrt{5} & 10 & \sqrt{74} \end{pmatrix},$$

leading to the following assignment:

$$z_1 = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Hence, we update the centroids to

$$\mu_1^1 = \frac{x_1 + x_2 + x_4 + x_5}{4} = \begin{pmatrix} -3 \\ \frac{19}{4} \end{pmatrix}$$

$$\mu_2^1 = x_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Second iteration: the (Euclidean) distances between the data points and the centroids from the first iteration are

$$\begin{pmatrix} \frac{9}{4} & \frac{\sqrt{369}}{4} & \frac{\sqrt{785}}{4} & \frac{\sqrt{89}}{4} & \frac{\sqrt{17}}{4} \\ \sqrt{80} & \sqrt{5} & 0 & \sqrt{85} & \sqrt{61} \end{pmatrix},$$

leading to the following assignment:

$$z_2 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

We update the centroids to

$$\mu_1^1 = \frac{x_1 + +x_4 + x_5}{3} = \begin{pmatrix} -4 \\ 6 \end{pmatrix}$$

$$\mu_2^1 = \frac{x_2 + x_3}{2} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$$

Third iteration: the (Euclidean) distances between the data points and the centroids from the first iteration are

$$\begin{pmatrix} \sqrt{2} & \sqrt{41} & \sqrt{74} & 1 & 1 \\ \frac{\sqrt{245}}{2} & \frac{\sqrt{5}}{2} & \frac{\sqrt{5}}{2} & \frac{\sqrt{265}}{2} & \frac{\sqrt{181}}{2} \end{pmatrix},$$

leading to the assignment:

$$z_3 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

We see that $z^3 = z^2$, hence $\mu^3 = \mu^2$ and we have converged.

This question is similar to Exercise 1.3 of Coursework 3.

(b) The minimal rank for the matrix is one, which is why the question marks have to be replaced with numbers such that the matrix has rank one. This can be achieved by ensuring that each row is a multiple of the first row. The second row is the first row multiplied by -3, and the second row is the first row multiplied by 6. Hence, we set

$$X = \begin{pmatrix} d & -2 & 1 & 7 \\ -3d & 6 & -3 & -21 \\ 6d & -12 & 6 & 42 \end{pmatrix} .$$

to complete the matrix such that it has rank one. Since the matrix has rank one, it can be decomposed into two vectors U and V via $X = UV^\top$, where U and V are of the form

$$U = \begin{pmatrix} 1 \\ -3 \\ 6 \end{pmatrix} \qquad \text{and} \qquad V = \begin{pmatrix} d \\ -2 \\ 1 \\ 7 \end{pmatrix} .$$

This question is similar to Exercise 2.2 of Coursework 5.

(c) First, we verify that the Bregman distance w.r.t. $f(z) := \sum_{j=1}^k z_j \log(z_j)$ reads

$$\begin{aligned} D_f(z,x) &= \sum_{j=1}^k z_j \log(z_j) - \sum_{j=1}^k x_j \log(x_j) - \sum_{j=1}^k (1+\log(x_j))(z_j - x_j) \\ &= \sum_{j=1}^k \left[z_j \log\left(\frac{z_j}{x_j}\right) + x_j - z_j \right] , \end{aligned}$$

which is also known as the Kullback-Leibler divergence. Next, we show that $z = \text{softmax}\,(-\log(x) f(y))$ is a global minimiser of our objective. We do this by reformulating

$$\begin{aligned} D_f(\tilde{z}, x) + \langle \tilde{z}, g(y) \rangle &= \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{\tilde{z}_j}{x_j}\right) + x_j - \tilde{z}_j \right] + \sum_{j=1}^k \tilde{z}_j g(y)_j \\ &= \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{\tilde{z}_j}{x_j}\right) + x_j - \tilde{z}_j + g(y)_j \tilde{z}_j \right] \\ &= \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{\tilde{z}_j}{x_j}\right) + x_j - \tilde{z}_j + g(y)_j \tilde{z}_j + \tilde{z}_j \log(z_j) - \tilde{z}_j \log(z_j) \right] \\ &= \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{\tilde{z}_j}{z_j}\right) + x_j - \tilde{z}_j + g(y)_j \tilde{z}_j + \tilde{z}_j \log(z_j) - \tilde{z}_j \log(x_j) \right] \end{aligned}$$

We replace $\tilde{z}_j \log(z_j)$ with

$$\begin{aligned}\tilde{z}_j \log(z_j) &= \tilde{z}_j \left(\log(x_j \exp(-g(y)_j)) - \log\left(\sum_{i=1}^k x_i \exp(-g(y)_i) \right) \right) \\ &= \tilde{z}_j \log(x_j) - \tilde{z}_j g(y)_j - \tilde{z}_j \log\left(\sum_{i=1}^k x_i \exp(-g(y)_i) \right) ,\end{aligned}$$

and, thus, obtain

$$\begin{aligned}&D_f(\tilde{z}, x) + \langle \tilde{z}, g(y) \rangle \\ &= \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{\tilde{z}_j}{z_j} \right) + x_j - \tilde{z}_j + \tilde{z}_j \log\left(\frac{1}{\sum_{i=1}^k x_i \exp(-g(y)_i)} \right) \right] \\ &= \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{\tilde{z}_j}{z_j} \right) + z_j - \tilde{z}_j + \tilde{z}_j \log\left(\frac{1}{\sum_{i=1}^k x_i \exp(-g(y)_i)} \right) + x_j - z_j \right] \\ &= D_f(\tilde{z}, z) + \sum_{j=1}^k \left[\tilde{z}_j \log\left(\frac{1}{\sum_{i=1}^k x_i \exp(-g(y)_i)} \right) + x_j - z_j \right] \\ &= D_f(\tilde{z}, z) + c\left(1 - \sum_{j=1}^k \tilde{z}_j \right) - \underbrace{c + \sum_{j=1}^k [z_j - x_j]}_{\text{constant, independent of } \tilde{z}} ,\end{aligned}$$

for $c := \log\left(\sum_{i=1}^k x_i \exp(-g(y)_i) \right)$ independent of $\tilde{z}$. Hence, we have

$$\begin{aligned}&\arg\min_{\tilde{z} \in \mathbb{R}^k} \left\{ D_f(\tilde{z}, x) + \langle \tilde{z}, g(y) \rangle \quad \text{subject to} \quad \tilde{z} \in [0,1]^k, \text{ and } \sum_{j=1}^k \tilde{z}_j = 1 \right\} \\ = &\arg\min_{\tilde{z} \in \mathbb{R}^k} \left\{ D_f(\tilde{z}, z) + c\left(1 - \sum_{j=1}^k \tilde{z}_j \right) + c + \sum_{j=1}^k [z_j - x_j] \right. \\ &\qquad \left. \text{subject to} \quad \tilde{z} \in [0,1]^k, \text{ and } \sum_{j=1}^k \tilde{z}_j = 1 \right\} \\ = &\arg\min_{\tilde{z} \in \mathbb{R}^k} \left\{ D_f(\tilde{z}, z) + c\left(1 - \sum_{j=1}^k \tilde{z}_j \right) \quad \text{subject to} \right. \\ &\qquad \left. \tilde{z} \in [0,1]^k, \text{ and } \sum_{j=1}^k \tilde{z}_j = 1 \right\} \\ = &\arg\min_{\tilde{z} \in \mathbb{R}^k} \left\{ D_f(\tilde{z}, z) \quad \text{subject to} \quad \tilde{z} \in [0,1]^k, \text{ and } \sum_{j=1}^k \tilde{z}_j = 1 \right\} .\end{aligned}$$

Here, the final equality follows from the fact that the constraint $\sum_{j=1}^k \tilde{z}_j = 1$ already ensures $1 - \sum_{j=1}^k \tilde{z}_j = 0$. Since f is convex, the Bregman distance is non-negative, i.e. $D_f(x, y) \geq 0$ for all $x, y \in \mathbb{R}^k$. Hence, the smallest value that we can attain is $D_f(x, y) = 0$, which we do attain for $D_f(\tilde{z}, z)$. Since z also satisfies the constraints, we know that $\tilde{z} = z$ is a global minimiser of the original optimisation problem.

This question tests the understanding of multiple concepts in the lecture notes and builds on coursework related to Bregman distances.

(d) Similar to proximal gradient descent as described in the lecture notes, we can convert the optimisation problem into an iterative procedure that approximates the original problem via

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^k} \left\{ h(x) + D_g(x, x^k) \text{ subject to } x \in [0,1]^k \,, \text{ and } \sum_{j=1}^k x_j = 1 \right\} . \quad (1)$$

Usually, in the lecture notes we always chose $g(x) = \frac{1}{2\tau}\|x\|^2 - h(x)$. However, in order to make use of Question 3(c) we chose $g(x) = \frac{1}{\tau} f(x) - h(x)$ for $f(x) = \sum_{j=1}^k \left(x_j \log(x_j) \right)$ instead. This transforms (1) into

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^k} \left\{ D_f(x, x^k) + \langle x, \tau \nabla h(x^k) \rangle \text{ subject to } x \in [0,1]^k \,, \text{ and } \sum_{j=1}^k x_j = 1 \right\} ,$$

which is the same optimisation problem as in Question 3(c). Hence, its closed form solution reads

$$x_i^{k+1} = \frac{x_i^k \exp\left(-\tau (\nabla h(x^k))_i\right)}{\sum_{j=1}^k x_j^k \exp\left(-\tau (\nabla h(x^k))_j\right)} ,$$

for all $i \in \{1, \ldots, k\}$.

This part of the question requires understanding of proximal gradient descent as introduced in Section 2.7.3 in the lecture notes, as well as the ability to transfer knowledge about the solution of Question 3(c) into a new context.

End of Paper.