Queen Mary
University of London

Main Examination period 2021 – 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 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 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 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

 Continue to next page

The notation log refers to the natural logarithm. 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 $f(x) = \frac{dx-2x^2}{1+4x^2}$ 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). **[8]**

(c) Rewrite the function $f(x) = \frac{1}{2}\langle Qx, x\rangle$, acting on a vector $x = a + \varepsilon b$ of dual numbers, to $g(a) + \varepsilon h(a,b)$. Here, $Q \in \mathbb{R}^{n\times n}$ is a square matrix. Specify both g and h, and compute the partial derivative $\partial/\partial a_l f(a_1, \ldots, a_n)$ for some $l \in \{1, \ldots, n\}$. **[6]**

(d) Characterise all cases of the subdifferential $\partial\chi_{\geq 0}$ of the characteristic function

$$\chi_{\geq 0}(x) := \begin{cases} 0 & x \geq 0 \\ \infty & \text{otherwise} \end{cases}.$$

[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 four different cases. **[8]**

 Continue to next page

Solution:

(a) We compute

$$\begin{aligned} f(a+\varepsilon b) &= \frac{d(a+\varepsilon b) - 2(a+\varepsilon b)^2}{1+4(a+\varepsilon b)^2} \\ &= \frac{da+\varepsilon db - 2a^2 - \varepsilon 4ab}{1+4a^2+\varepsilon 8ab} = \frac{da-2a^2+\varepsilon\,(db-4ab)}{1+4a^2+\varepsilon 8ab} \\ &= \frac{u+\varepsilon v}{w+\varepsilon z}\,, \end{aligned}$$

for $u := da - 2a^2$, $v := db - 4ab$, $w := 1 + 4a^2$ and $z := 8ab$. We further compute

$$\begin{aligned} f(a+\varepsilon b) &= \frac{u+\varepsilon v}{w+\varepsilon z} \\ &= \frac{(u+\varepsilon v)(w-\varepsilon z)}{(w+\varepsilon z)(w-\varepsilon z)} = \frac{uw+\varepsilon(vw-uz)}{w^2} \\ &= \frac{u}{w} + \varepsilon\frac{vw-uz}{w^2} \\ &= \frac{da-2a^2}{1+4a^2} + \varepsilon\frac{(db-4ab)(1+4a^2)-(da-2a^2)8ab}{(1+4a^2)^2} \\ &= \frac{da-2a^2}{1+4a^2} + \varepsilon\, b\,\frac{d-4a(1+ad)}{(1+4a^2)^2} \\ &= g(a) + \varepsilon h(a,b)\,, \end{aligned}$$

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

This exercise is similar to a future coursework exercise.

(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

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

This exercise is similar to a future coursework exercise.

(c) Similar to the previous exercise we conclude

$$\begin{aligned} f(x) = f(a+\varepsilon b) &= \frac{1}{2}\langle Q(a+\varepsilon b), a+\varepsilon b\rangle \\ &= \frac{1}{2}\langle Q(a+\varepsilon b), a\rangle + \frac{\varepsilon}{2}\langle Q(a+\varepsilon b), b\rangle \\ &= \frac{1}{2}\langle Qa, a\rangle + \frac{\varepsilon}{2}\langle Qb, a\rangle + \frac{\varepsilon}{2}\langle Qa, b\rangle + \underbrace{\frac{\varepsilon^2}{2}}_{=0}\langle Qb, b\rangle \\ &= \frac{1}{2}\langle Qa, a\rangle + \varepsilon\left(\frac{1}{2}\langle Qa, b\rangle + \frac{1}{2}\langle Qb, a\rangle\right) \\ &= g(a) + \varepsilon h(a,b)\,, \end{aligned}$$

for $g(a) = \frac{1}{2}\langle Qa, a\rangle$ and $h(a,b) = \frac{1}{2}\left(\langle Qa, b\rangle + \langle Qb, a\rangle\right)$. If we specify $e_l = (0 \ \dots \ 0 \underbrace{1}_{l-\text{th position}} 0 \ \dots \ 0)$, we can compute the l-th partial derivative by evaluating $h(a, e_l)$, which reads

$$\begin{aligned} h(a,e_l) = \frac{1}{2}\left(\langle Qa, e_l\rangle + \langle Qe_l, a\rangle\right) &= \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}a_j(e_l)_i + \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}(e_l)_j a_i\right) \\ &= \frac{1}{2}\left(\sum_{j=1}^{n} q_{lj}a_j + \sum_{i=1}^{n} q_{il}a_i\right) \\ &= \left(\frac{1}{2}\left(Q + Q^\top\right)a\right)_l . \end{aligned}$$

If Q is symmetric, we have $Q = Q^\top$ and the l-th partial derivative simplifies to $(Qa)_l$.

This exercise is similar to a future coursework exercise.

(d) The definition of the subdifferential for the characteristic function $\chi_{\geq 0}$ reads

$$\partial\chi_{\geq 0}(x) = \{p \in \mathbb{R} \,|\, \chi_{\geq 0}(y) \geq \chi_{\geq 0}(x) + p(y-x)\,,\ \forall y \in \mathbb{R}\}\ .$$

We characterise this subdifferential by case analysis.
Case 1: suppose $x < 0$, then $\partial\chi_{\geq 0}(x) = \emptyset$, since the right-hand-side of the inequality will be $+\infty$ regardless of the choice of p, while the left-hand-side is 0 for every $y > 0$.
Case 2: suppose $x > 0$, then $p = 0$ satisfies the inequality for all $y \in \mathbb{R}$. Hence, we observe $\partial\chi_{\geq 0}(x) = \{0\}$ in this case.
Case 3: suppose $x = 0$. In this case, the inequality reads $\chi_{\geq 0}(y) \geq py$ for all $y \in \mathbb{R}$. For this to be valid for all $y \in \mathbb{R}$, p can only take non-positive values, in order to guarantee that $yp \leq 0$ for $y > 0$. Hence, we observe $\partial\chi_{\geq 0}(x) = (-\infty, 0]$ for $x = 0$.
Combining all cases yields

$$\partial\chi_{\geq 0}(x) = \begin{cases} \{0\} & x > 0 \\ (-\infty, 0] & x = 0 \\ \emptyset & x < 0 \end{cases}\ .$$

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

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

for $p \in \partial|y|$. From the lecture notes we know

$$\partial|y| = \begin{cases} \{1\} & x > 0 \\ [-1, 1] & x = 0 \\ \{-1\} & x < 0 \end{cases}\ .$$

 Continue to next page

Hence, for $y > 0$ we have $p = 1$ and therefore observe

$$D_f^p(x,y) = |x| - y - (x-y) = |x| - x = \begin{cases} 0 & x \geq 0 \\ -2x & x < 0 \end{cases}.$$

In similar fashion we conclude for $y < 0$ that we have $p = -1$ and

$$D_f^p(x,y) = |x| + y + (x-y) = |x| + x = \begin{cases} 2x & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

For $y = 0$, the Bregman distance depends on the particular subgradient $p \in [-1,1]$. We observe

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

which always equals zero if $x = 0$ or if $p = x/|x|$ for $x \neq 0$. For $x > 0$ we have

$$D_f^p(x,y) = (1-p)x\,,$$

while for $x < 0$ we get

$$D_f^p(x,y) = -(1+p)x\,.$$

Combining all four cases yields

$$D_f^p(x,y) = \begin{cases} 0 & (\operatorname{sign}(x) = \operatorname{sign}(y)) \vee (x = 0) \\ 2|x| & ((x > 0) \wedge (0 > y)) \vee ((y > 0) \wedge (0 > x)) \\ (1-p)x & (x > 0) \wedge (y = 0) \\ -(1+p)x & (x < 0) \wedge (y = 0) \end{cases}.$$

This exercise is similar to a future coursework exercise.

 Continue to next page

Question 2 [24 marks].

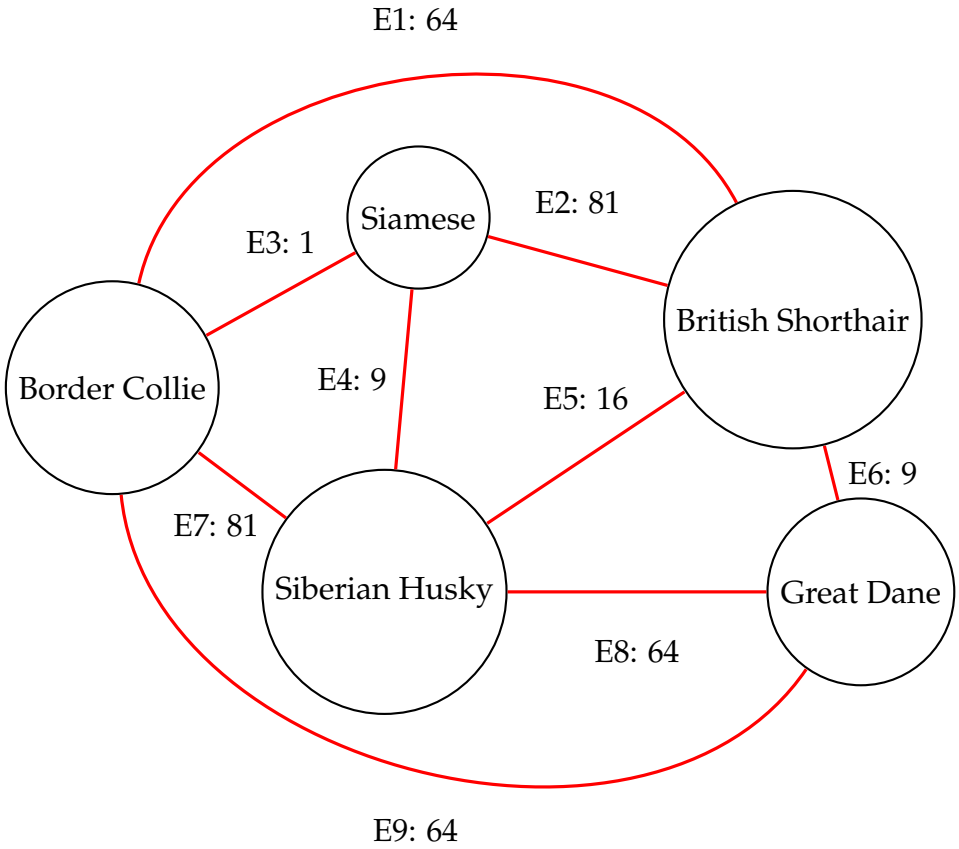
(a) It is the start of the Covid-19 vaccine role-out program and you want to decide whether or not to get vaccinated. You base your decision on three factors:

- Am I in any of the groups (age above threshold, living in care home, front-line worker etc.) eligible for the vaccine?
- Is the scientific advice in favour of getting vaccinated?
- Do people in my WhatsApp group chat think that I should get vaccinated?

Suppose you only get vaccinated if you are in any of the eligible vaccination groups and only if the scientific advice is in favour of you getting vaccinated. However, you really do not care about what people in your WhatsApp group chat think about you getting vaccinated or not.

Model this binary decision process with a perceptron and choose some appropriate weights to mimic the decision process accurately. **[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 "dogs" or the class "cats". Suppose we are in a semi-supervised setting, where the node "Great Dane" is already labelled $v_{\text{Great Dane}} = 1$ (class "dogs") and the node "Siamese" is labelled as $v_{\text{Siamese}} = 0$ (class "cats"). Determine the remaining labels with the same procedure as described in the lecture notes and classify each node. **[8]**

Solution:

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

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

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
1	1	1	0
0	1	0	1
0	1	0	0
0	0	1	1
0	0	1	0
0	0	0	1
0	0	0	0

Hence, one would never get vaccinated if either the scientific advice is not in favour or if one is not in any of the eligible groups. But the WhatApp group opinion has no influence on the decision.

This exercise is similar to a future coursework exercise.

(b) The incidence matrix for the displayed graph is

$$M_w = \left(\begin{array}{c|ccccc} \text{E1} & -8 & 8 & 0 & 0 & 0 \\ \text{E2} & 0 & -9 & 0 & 9 & 0 \\ \text{E3} & -1 & 0 & 0 & 1 & 0 \\ \text{E4} & 0 & 0 & 0 & -3 & 3 \\ \text{E5} & 0 & -4 & 0 & 0 & 4 \\ \text{E6} & 0 & -3 & 3 & 0 & 0 \\ \text{E7} & -9 & 0 & 0 & 0 & 9 \\ \text{E8} & 0 & 0 & -8 & 0 & 8 \\ \text{E9} & -8 & 0 & 8 & 0 & 0 \\ \hline & \text{B. Collie} & \text{B. Shorthair} & \text{G. Dane} & \text{Siamese} & \text{S. Husky} \end{array}\right)$$

This question is similar to Exercise 1.1 of Coursework 1.

(c) The corresponding graph Laplacian then reads

$$L_w = M_w^\top M_w = \left(\begin{array}{c|ccccc} \text{B. Collie} & 210 & -64 & -64 & -1 & -81 \\ \text{B. Shorthair} & -64 & 170 & -9 & -81 & -16 \\ \text{G. Dane} & -64 & -9 & 137 & 0 & -64 \\ \text{Siamese} & -1 & -81 & 0 & 91 & -9 \\ \text{S. Husky} & -81 & -16 & -64 & -9 & 170 \\ \hline & \text{B. Collie} & \text{B. Shorthair} & \text{G. Dane} & \text{Siamese} & \text{S. Husky} \end{array}\right)$$

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

$$v = \begin{pmatrix} v_{\text{Border Collie}} \\ v_{\text{British Shorthair}} \\ v_{\text{Great Dane}} \\ v_{\text{Siamese}} \\ v_{\text{Siberian Husky}} \end{pmatrix}$$

we know the third and fourth entry; the third belongs to the class "dogs" and therefore takes on the value $v_{\text{Great Dane}} = 1$, whereas the fourth entry belongs to the class "cats", hence $v_{\text{Siamse}} = 0$. Thus, for $y = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top$ we have

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

 Continue to next page

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

$$\begin{pmatrix} 210 & -64 & -81 \\ -64 & 170 & -16 \\ -81 & -16 & 170 \end{pmatrix} \tilde{v} = \begin{pmatrix} 64 \\ 9 \\ 64 \end{pmatrix} ,$$

Solving this linear system leads to the (approximate) solution

$$\tilde{v} \approx \begin{pmatrix} 0.7157 \\ 0.3934 \\ 0.7545 \end{pmatrix} .$$

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

$$v = \begin{pmatrix} v_{\text{Border Collie}} \\ v_{\text{British Shorthair}} \\ v_{\text{Great Dane}} \\ v_{\text{Siamese}} \\ v_{\text{Siberian Husky}} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} .$$

This question is similar to Question 1.3 of Coursework 1.

 Continue to next page

Question 3 [26 marks].

(a) Perform k-means clustering by hand for the five data points $x_1 = \begin{pmatrix}1\\0\end{pmatrix}$, $x_2 = \begin{pmatrix}-7\\3\end{pmatrix}$, $x_3 = \begin{pmatrix}0\\-1\end{pmatrix}$, $x_4 = \begin{pmatrix}13\\15\end{pmatrix}$, and $x_5 = \begin{pmatrix}-11\\17\end{pmatrix}$. Assume $k = 2$ clusters and initialise your centroids as

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

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:

$$\begin{pmatrix} 1 & -2 & d & -7 \\ -3 & 6 & ? & 21 \\ ? & -10 & ? & -35 \end{pmatrix}.$$

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

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

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

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

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

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 $D_f(\tilde{z}, x) - \langle \tilde{z}, y \rangle$ 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) Consider the following modification of k-means clustering with uncertainty:

$$(\hat{z}, \hat{\mu}) = \arg \min_{z \in \mathbb{R}^{s \times k},\, \mu \in \mathbb{R}^{n \times k}} \left\{ \sum_{i=1}^s \sum_{j=1}^k z_{ij} \|x_i - \mu_j\|^2 \text{ s.t. } z_{ij} \in [0,1], \sum_{j=1}^k z_{ij} = 1, \forall i, j \right\},$$

for s data points $\{x_i\}_{i=1}^s$, where every $x_i \in \mathbb{R}^n$ is n-dimensional, centroids $\mu \in \mathbb{R}^{n \times k}$ and assignments $z \in \mathbb{R}^{s \times k}$. What is the difference to traditional k-means clustering? Formulate an algorithm to computationally solve this modified k-means clustering problem. **[9]**

 Continue to next page

Solution:

(a) We show the solution for $d = 5$; other solutions are computed in similar fashion. We compute the Euclidean distances of the data points with respect to the initial centroids:

$$\begin{pmatrix} 4 & \sqrt{153} & \sqrt{26} & 17 & \sqrt{545} \\ \sqrt{26} & \sqrt{113} & 4 & \sqrt{569} & \sqrt{605} \end{pmatrix},$$

leading to the following assignment:

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

Hence, we update the centroids to

$$\mu_1^1 = \frac{x_1 + x_4 + x_5}{3} = \begin{pmatrix} 1 \\ \frac{32}{3} \end{pmatrix}$$

$$\mu_2^1 = \frac{x_2 + x_3}{2} = \begin{pmatrix} -\frac{7}{2} \\ 1 \end{pmatrix}$$

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

$$\begin{pmatrix} \frac{32}{3} & \frac{\sqrt{1105}}{3} & \frac{\sqrt{1234}}{3} & \frac{\sqrt{1465}}{3} & \frac{\sqrt{1657}}{3} \\ \frac{\sqrt{85}}{2} & \frac{\sqrt{65}}{2} & \frac{\sqrt{65}}{2} & \frac{\sqrt{1873}}{2} & \frac{\sqrt{1249}}{2} \end{pmatrix},$$

leading to the following assignment:

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

We update the centroids to

$$\mu_1^1 = \frac{x_4 + x_5}{2} = \begin{pmatrix} 1 \\ 16 \end{pmatrix}$$

$$\mu_2^1 = \frac{x_1 + x_2 + x_3}{3} = \begin{pmatrix} -2 \\ \frac{2}{3} \end{pmatrix}$$

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

$$\begin{pmatrix} 16 & \sqrt{233} & \sqrt{290} & \sqrt{145} & \sqrt{145} \\ \frac{\sqrt{85}}{3} & \frac{\sqrt{274}}{3} & \frac{\sqrt{61}}{3} & \frac{\sqrt{3874}}{3} & \frac{\sqrt{3130}}{3} \end{pmatrix},$$

leading to the assignment:

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

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.

 Continue to next page

(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 5. Hence, we set

$$\begin{pmatrix} 1 & -2 & d & -7 \\ -3 & 6 & -3d & 21 \\ 5 & -10 & 5d & -35 \end{pmatrix}$$

to complete the matrix such that it has rank one.

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 $z = \text{softmax}\,(\log(x)y)$ is a global minimiser of our objective. We do this by reformulating

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

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

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

and, thus, obtain

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

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

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

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) The classical k-means clustering model as introduced in the lecture notes is

$$(\hat{z}, \hat{\mu}) = \arg \min_{z \in \mathbb{R}^{s \times k}, \mu \in \mathbb{R}^{n \times k}} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{k} z_{ij} \|x_i - \mu_j\|^2 \text{ subject to } z_{ij} \in \{0,1\}\,, \sum_{j=1}^{k} z_{ij} = 1,\ \forall i, j \right\},$$

for s data points $\{x_i\}_{i=1}^{s}$, where every $x_i \in \mathbb{R}^n$ is n-dimensional, centroids $\mu \in \mathbb{R}^{n \times k}$ and assignments $z \in \mathbb{R}^{s \times k}$. The difference of the new model is the change of the constraint $z_{ij} \in \{0,1\}$ to $z_{ij} \in [0,1]$; hence, we have

$$(\hat{z}, \hat{\mu}) = \arg \min_{z \in \mathbb{R}^{s \times k}, \mu \in \mathbb{R}^{n \times k}} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{k} z_{ij} \|x_i - \mu_j\|^2 \text{ subject to } z_{ij} \in [0,1]\,, \sum_{j=1}^{k} z_{ij} = 1,\ \forall i, j \right\}.$$

This part of the question requires knowledge about the definition of k-means clustering as introduced in Section 3.1.1 in the lecture notes.

Based on Question 3(c), we modify the unconstrained k-means clustering objective to

$$L^l(\mu, z) = \sum_{i=1}^{s} \sum_{j=1}^{k} z_{ij} \|x_i - \mu_j\|^2 + D_f(z, z^l)\,,$$

for $f(z) = \sum_{i=1}^{s} \sum_{j=1}^{k} z_{ij} \log(z_{ij})$ and propose an iterative, alternating minimisation algorithm of the form

$$z^{l+1} = \arg \min_{z \in \mathbb{R}^{s \times k}} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{k} z_{ij} \|x_i - \mu_j^l\|^2 + D_f(z, z^l) \text{ subject to } z_{ij} \in [0,1]\,, \sum_{j=1}^{k} z_{ij} = 1,\ \forall i, j \right\},$$

$$\mu^{l+1} = \arg \min_{\mu \in \mathbb{R}^{n \times k}} \left\{ \sum_{i=1}^{s} \sum_{j=1}^{k} z_{ij}^{l+1} \|x_i - \mu_j\|^2 \right\},$$

where $l = 0, 1, 2, \ldots$ denotes the iteration index. Thanks to Question 3(c), we know that the solution of the first problem simply reads

$$z_{ij}^{l+1} = \frac{z_{ij}^l \exp\left(-\|x_i - \mu_j^l\|^2\right)}{\sum_{r=1}^{k} z_{ir}^l \exp\left(-\|x_i - \mu_r^l\|^2\right)}\,, \text{ for all } i \in \{1, \ldots, s\} \text{ and } j \in \{1, \ldots, k\}\,,$$

while the update for the centroids is unchanged and is given in the lecture notes as

$$\mu_j^{l+1} = \frac{\sum_{i=1}^{s} z_{ij}^{l+1} x_i}{\sum_{i=1}^{s} z_{ij}}\,, \quad \text{for all} \quad j \in \{1, \ldots, k\}\,.$$

This part of the question requires understanding of k-means clustering as introduced in Section 3.1.1 in the lecture notes and its numerical implementation, as well as the ability to transfer knowledge about the solution of Question 3(c) *into a new context.*

End of Paper.