Answer Key and Marking Scheme

Q1

(1) The value of β can be found out by deriving x_3 by appealing to the update steps of gradient descent with momentum term. Recall that the update step of gradient descent with momentum on a function $\mathbf{f}(\mathbf{z})$ where $\mathbf{z} \in \mathbb{R}^{\mathbf{2}}$ had the following form:

$$\mathbf{z_{i+1}} = \mathbf{z_i} - \alpha \nabla \mathbf{f}(\mathbf{z_i}) + \mathbf{v_i}$$

$$\text{where } \mathbf{v_i} = \beta(\mathbf{z_i} - \mathbf{z_{i-1}}) \text{ and } \mathbf{v_0} = \mathbf{0}$$

Approach 1 (using x updates)

By using the above formula only on first variable x, we can build a formula for x_3 by treating β as an unknown constant. The resultant expression is derived as follows. Recall that $\nabla f = \begin{bmatrix} 6x \\ 4y \end{bmatrix}$ and $x_0 = 2, y_0 = 4$

(a) $x_1 = x_0 - \alpha 6 x_0 = 2 - \frac{1}{2}.12 = -4$ (0.5 mark)
(b) $x_2 = x_1 - \alpha 6 x_1 + \beta(x_1 - x_0) == 8 - 6\beta$ (1 mark)
(c) $x_3 = x_2 - \alpha 6 x_2 + \beta(x_2 - x_1) = -6\beta^2 + 24\beta - 16$ (1.5 mark)

From above $x_3 = -6\beta^2 + 24\beta - 16 = -7.36$.
Solving above quadratic equation in β, we get $\beta = 0.4$ or $\beta = 3.6$. Since $\beta \in (0, 1)$, we get $\beta = 0.4$
(1 mark)

Approach 2 leading to same answer (using y updates instead of x updates)

By using the above formula only on second variable y, we can build a formula for y_3 by treating β as an unknown constant.

(a) $y_1 = y_0 - \alpha 4 y_0 = -4$ (0.5 mark)
(b) $y_2 = y_1 - \alpha 4 y_1 + \beta(y_1 - y_0) = (4 - 8\beta)$ (1 mark)
(c) $y_3 = y_2 - \alpha 4 y_2 + \beta(y_2 - y_1) = -8\beta^2 + 16\beta - 4$ (1.5 mark)

From above $y_3 = -8\beta^2 + 16\beta - 4 = 1.12$
Solving above quadratic equation in β, we get $\beta = 0.4$ or $\beta = 1.6$. Since $\beta \in (0, 1)$, we get $\beta = 0.4$ (1 mark)

(2) (i) Given matrix is positive definite, so it has a Cholesky decomposition $A = LL^T$. The entries of L above principal diagonal will all be zero by definition of lower triangular matrix. The other entries can be obtained by the formulas given in the image screenshot from Lecture 3 slides. The same formulas can also be obtained by simply multiplying L and L^T and equate to A.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

We can solve for the elements of the lower triangular matrix to get $l_{11} = \sqrt{a_{11}}$, $l_{22} = \sqrt{a_{22} - l_{21}^2}$, $l_{33} = \sqrt{a_{33} - (l_{31}^2 + l_{32}^2)}$.
For the elements below the diagonal we have $l_{21} = \frac{a_{21}}{l_{11}}$, $l_{31} = \frac{a_{31}}{l_{11}}$ and $l_{32} = \frac{a_{32} - l_{31} l_{21}}{l_{22}}$.

FIGURE 1. Cholesky Decomposition

The lower triangular matrix has the following structure: $L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}$

The lower triangular matrix L is obtained as follows from the above formula:

$$l_{11} = 2, l_{21} = 1 \quad (0.5\ mark)$$
$$l_{22} = 4, l_{31} = 3 \quad (1\ mark)$$
$$l_{32} = 1, l_{33} = 5 \quad (1\ mark)$$

(ii) The eigenvalues of L can be found out by constructing the characteristic equation

$$|L - \lambda I| = (\lambda - 2)(\lambda - 4)(\lambda - 5) = 0$$

The 3 eigenvalues of L are $\lambda_1 = 2, \lambda_1 = 4, \lambda_1 = 5$ (0.5 mark)

(3) (a) The gradient of $f(x, y)$ at (x_0, y_0) is as follows

$$\nabla f(x_0, y_0) = \begin{bmatrix} 2x_0 \\ 2\beta y_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 2\beta \end{bmatrix}$$

The optimal step size is obtained as solution of following 1 dimensional optimization problem

$$\operatorname{argmin}_\alpha f\left(\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \alpha \nabla f(x_0, y_0) \right)$$

(0.5 mark)

after substituting $x_0 = 1$ and $y_0 = 1$, we get that, this is equivalent to solving

$$\operatorname{argmin}_\alpha f(1 - 2\alpha, 1 - 2\alpha\beta)$$

In other words we need to find the minimum of

$$g(\alpha) = (1 - 2\alpha)^2 + \beta(1 - 2\alpha\beta)^2$$

(0.5 mark)

To find minimum of $g(\alpha)$ with respect to α, we find $g'(\alpha)$ and equate it to zero

$$g'(\alpha) = 2\beta^3\alpha - \beta^2 + 2\alpha - 1 = 0$$

(1 mark)

Hence the closed form expression for α is given by

$$\alpha = \frac{1 + \beta^2}{2 + 2\beta^3}$$

(0.5 mark)

(b) If it is given that $\alpha = 0.5$, substitutng in the repvious expression we get the equation

$$0.5 = \frac{1 + \beta^2}{2 + 2\beta^3}$$

Rearranging we get the expression $\beta^2(\beta - 1) = 0$. Hence potential value of β is 0 or 1. Its given in question that $\beta \neq 0$.
Hence final answer is $\beta = 1$. (0.5 mark)

Q2

(1) Suitable transformation would be

$\phi(x) = x \mod 2$ (1 mark)

$\phi((7,0) = 7 \mod 2 = 1$ and $\phi((9,0) = 9 \mod 2 = 1$ (1 mark)

$\phi((8,0) = 8 \mod 2 = 0$ and $\phi((10,0) = 10 \mod 2 = 0$ (1 mark)

The decision boundary is $x = 0.5$ (2 marks)

(2) To compute the Kernel matrix K using the feature transformation $\phi(x) = [x_1, x_2, ||x||]$, we need to first calculate the transformed features for each data point in the dataset X, and then compute the dot product of these transformed features to obtain the entries of the Kernel matrix.

Given the dataset $X = [(4, -3), (0, 1)]$, let's compute the transformed features for each data point:

For $(4, -3)$: - $\phi((4, -3)) = [4, -3, ||(4, -3)||] = [4, -3, 5]$ (1 mark)

For $(0, 1)$: - $\phi((0, 1)) = [0, 1, ||(0, 1)||] = [0, 1, 1]$ (1 mark)

Now, let's compute the dot product of these transformed features to obtain the entries of the Kernel matrix K:

$$K_{ij} = \phi(x_i) \cdot \phi(x_j)$$

Where x_i and x_j are data points in the dataset X.

For our dataset, the Kernel matrix will be a 2×2 matrix. Let's calculate it:

For the entry K_{11}:

$$K_{11} = \phi((4,-3))\cdot\phi((4,-3)) = [4,-3,5]\cdot[4,-3,5] = 4\times4+(-3)\times(-3)+5\times5 = 16+9+25 = 50$$

For the entry K_{12}:

$$K_{12} = \phi((4,-3))\cdot\phi((0,1)) = [4,-3,5]\cdot[0,1,1] = 4\times0+(-3)\times1+5\times1 = -3+5 = 2$$

For the entry K_{21}:

$$K_{21} = \phi((0,1))\cdot\phi((4,-3)) = [0,1,1]\cdot[4,-3,5] = 0\times4+1\times(-3)+1\times5 = -3+5 = 2$$

For the entry K_{22}:

$$K_{22} = \phi((0,1)) \cdot \phi((0,1)) = [0,1,1] \cdot [0,1,1] = 0 \times 0 + 1 \times 1 + 1 \times 1 = 1 + 1 = 2$$

So, the Kernel matrix K is:

$$K = \begin{bmatrix} 50 & 2 \\ 2 & 2 \end{bmatrix}$$

(1 mark)

(3) The hinge loss for each data sample is given by:

$$\text{Hinge Loss} = \max(0, 1 - y \cdot y')$$

where: y is the true label. y' is the predicted label.

Given the data samples:

1. $y = 0.5$ and $y' = 1$ 2. $y = 1$ and $y' = -1$

Let's calculate the hinge loss for each sample:

1. For the first sample:

$$\text{Hinge Loss} = \max(0, 1 - 0.5 \cdot 1) = \max(0, 0.5) = 0.5$$

(1 mark)

2. For the second sample:

$$\text{Hinge Loss} = \max(0, 1 - 1 \cdot (-1)) = \max(0, 1 + 1) = \max(0, 2) = 2$$

(1 mark)

The misclassified sample is the one with a higher non-zero hinge loss. In this case, it's the second sample, where $y = 1$ and $y' = -1$.

Q3

(A) (i) Given $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{R_1^T} \\ \vdots \\ \boldsymbol{R_m^T} \end{bmatrix}$.

Therefore, $\boldsymbol{Ax} = \begin{bmatrix} \boldsymbol{R_1^T x} \\ \vdots \\ \boldsymbol{R_m^T x} \end{bmatrix} = \boldsymbol{0}$, since $\langle \boldsymbol{R_i}, \boldsymbol{x} \rangle = 0, i = 1, \cdots, m$.

$\Rightarrow \boldsymbol{b} = \boldsymbol{0}$. (1 mark)

(ii)

$$\begin{aligned} S &= \{\boldsymbol{x} \in \mathbb{R}^m | \langle \boldsymbol{R_i}, \boldsymbol{x} \rangle = 0, i = 1, \cdots, m\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^m | \boldsymbol{R_i^T x} = 0, i = 1, \cdots, m\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^m | \boldsymbol{Ax} = \boldsymbol{0}\} \\ &= N(\boldsymbol{A}). \end{aligned}$$

$N(\boldsymbol{A})$ is a subspace of $\mathbb{R}^m$ when $\boldsymbol{A}$ is of order $m \times m$. (2 marks)
(Kindly give marks for alternate correct approach as well.)

(iii) Now, rank$(\boldsymbol{A}) = m$. Therefore by rank nullity theorem, we get
dim S = dim $N(\boldsymbol{A}) = m-$ rank$(\boldsymbol{A}) = 0$. (0.5 marks)
$\Rightarrow S = \{\boldsymbol{0}\}$ (0.5 marks)

(B) (i) Given $\sigma(z) = (1 + e^{-z})^{-1}$.

$$\begin{aligned} \frac{d\sigma}{dz} &= (-1)(1 + e^{-z})^{-2}(e^{-z})(-1) \\ &= (\sigma(z))^2(e^{-z}) && \text{(0.5 marks)} \\ &= (\sigma(z))^2((1 + e^{-z}) - 1) \\ &= (\sigma(z))^2(\frac{1}{(1 + e^{-z})^{-1}} - 1) \\ &= (\sigma(z))^2(\frac{1}{(\sigma(z))} - 1) \\ &= \sigma(z)(1 - \sigma(z)) && \text{(0.5 marks)} \end{aligned}$$

(ii)

$$\begin{aligned}
f(x,y) &= \alpha \ln\left(\frac{1}{\sigma(x+\beta y)}\right) + (1-\alpha)\ln\left(\frac{1}{1-\sigma(x+\beta y)}\right)\\
&= -\alpha \ln(\sigma(x+\beta y)) + (1-\alpha)\ln\left(\frac{1}{1-(1+e^{-(x+\beta y)})^{-1}}\right)\\
&= -\alpha \ln(\sigma(x+\beta y)) + (1-\alpha)\ln\left(\frac{1+e^{-(x+\beta y)}}{1+e^{-(x+\beta y)}-1)}\right)\\
&= -\alpha \ln(\sigma(x+\beta y)) + (1-\alpha)\ln\left(\frac{1+e^{-(x+\beta y)}}{e^{-(x+\beta y)})}\right)\\
&= -\alpha \ln(\sigma(x+\beta y)) + (1-\alpha)\ln\left(\frac{e^{(x+\beta y)}}{\sigma(x+\beta y)}\right)\\
&= -\alpha \ln(\sigma(x+\beta y)) + (1-\alpha)[(x+\beta y) - \ln(\sigma(x+\beta y))]\\
&= (1-\alpha)(x+\beta y) - \ln(\sigma(x+\beta y))
\end{aligned}$$

(2 marks)

(iii)

$$\begin{aligned}
\frac{\partial f}{\partial x} &= \frac{\partial}{\partial x}[(1-\alpha)(x+\beta y) - \ln(\sigma(x+\beta y))] \text{ (from (ii))}\\
&= (1-\alpha) - \frac{1}{\sigma(x+\beta y)}\sigma(x+\beta y)(1-\sigma(x+\beta y))(1) \text{ (from (i))}\\
&= 1-\alpha-1+\sigma(x+\beta y)\\
&= -\alpha+\sigma(x+\beta y)
\end{aligned}$$

(1 mark)

$$\begin{aligned}
\frac{\partial f}{\partial y} &= \frac{\partial}{\partial y}[(1-\alpha)(x+\beta y) - \ln(\sigma(x+\beta y))] \text{ (from (ii))}\\
&= (1-\alpha)\beta - \frac{1}{\sigma(x+\beta y)}\sigma(x+\beta y)(1-\sigma(x+\beta y))(\beta) \text{ (from (i))}\\
&= \beta-\alpha\beta-\beta+\sigma(x+\beta y)\beta\\
&= \beta(-\alpha+\sigma(x+\beta y))
\end{aligned}$$

(1 mark)

(iv) The Taylor's polynomial of degree 1 of f at $(0,0)$ is given by

$$\begin{aligned}
T_1(x,y) &= f(0,0) + \left[\frac{\partial f}{\partial x}(0,0), \frac{\partial f}{\partial y}(0,0)\right][x,y]^T\\
&= \ln(2) + \left[-\alpha+\frac{1}{2}, \beta(-\alpha+\frac{1}{2})\right][x,y]^T\\
&= \ln(2) + (-\alpha+\frac{1}{2})(x+\beta y)
\end{aligned}$$

(0.5 marks) for formula (0.5 marks) for answer.

Q4

The figure below shows 4 points, representing some data in $\mathbb{R}^2$

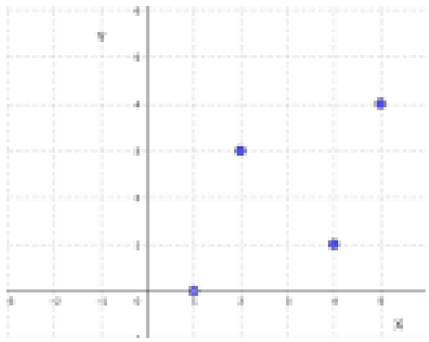

FIGURE 1. PCA

(A) The four points are $(2,3)$, $(4,1)$, $(5,4)$ and $(1,0)$ This data X is represented as:

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & 5 & 1 \\ 3 & 1 & 4 & 0 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 12/4 \\ 8/4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

(1 mark)

$$\mathbf{X} - \mu = \begin{bmatrix} -1 & 1 & 2 & -2 \\ 1 & -1 & 2 & -2 \end{bmatrix}$$

$$\mathbf{cov}(\mathbf{X}) = \frac{\mathbf{1}}{\mathbf{4}} \begin{bmatrix} -1 & 1 & 2 & -2 \\ 1 & -1 & 2 & -2 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix} = \frac{\mathbf{1}}{\mathbf{4}} \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

(2 marks)

To find eigen values and eigen vectors:
eigen values:

$$\lambda^2 - 5\lambda + 4 = 0 \implies \lambda = 4, 1$$

eigen vectors:

$$\lambda = 4 \implies v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\lambda = 1 \implies v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Principal component directions:

$$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

and

$$\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

(2 marks)

(B) The components along first PC are:

$$\hat{x_1} = x_1^T * e_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 3 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{5}{\sqrt{2}}$$

$$\hat{x_2} = \frac{5}{\sqrt{2}}$$

$$\hat{x_3} = \frac{9}{\sqrt{2}}$$

$$\hat{x_4} = \frac{1}{\sqrt{2}}$$

(2 marks)

(C) Percentage variance captured by first component $= \frac{4}{4+1} = 0.8$ (i.e., 80%)

(1 mark)

(D) Rotating all the points by same angle does not affect the components along the principal component. It will be smae as the answer in part (B)

(2 marks)