Answer Key and Marking Scheme

Q1 Alternative approaches would also be accepted and awarded marks

a The data scientist has the following plan in mind: the $D \times N$ data matrix X is modified to RX where R is a diagonal matrix with $R_{ii} = \alpha_i$ to get the modified data matrix. The original covariance matrix XX^T could be written as $S\Lambda S^T$ using the spectral theorem. Now the covariance matrix for the modified data matrix RX can be written as RXX^TR^T which can be written as $RS\Lambda S^TR^T$. This might give the impression that the spectral decomposition of RXX^TR^T is $RS\Lambda S^TR^T$ where RS plays the role of the eigenvector matrix. However, we see that RS is not an orthogonal matrix since $RSS^TR^T = RR^T \neq I$. If RS had been an orthogonal matrix, the data scientist's claims would have been true, but this is not the case.

Marking Scheme: 3 Marks $\rightarrow$ factoring the modified covariance matrix, 2 Marks $\rightarrow$ the rest of the argument.

b The $D \times N$ data matrix X after transformation becomes RX where R is a diagonal matrix with $R_{ii} = \alpha_i$. The covariance matrix for this modified matrix becomes RXX^TR^T. Now XX^T is the previously computed covariance matrix, and we need to post-multiply it with a diagonal matrix R^T. The first column of the product XX^TR^T can be computed by taking a linear combination of the columns of XX^T with the combining coefficients coming from the first column of R^T. But the first column of R^T has a non-zero in only one place, ie the first location, so this linear combination can be computed in $O(D)$ time. Similarly the other columns of XX^TR^T can each be computed in $O(D)$ time, so that computing the whole matrix XX^TR^T takes $O(D^2)$ time. Computing RXX^TR^T means pre-multiplying the matrix XX^TR^T by the diagonal matrix R. The first row of RXX^TR^T can be obtained by taking a linear combination of the rows of XX^TR^T where the combining coefficients come from the first row of R. Since the first row of R has only a single non-zero entry, this boils down to a $O(D)$ computation. The other rows of XX^TR^T can be similarly computed, so the computation of RXX^TR^T can be seen to be of complexity $O(D^2) + O(D^2) = O(D^2)$.

Marking Scheme: 2 Marks $\rightarrow$ recognizing the data matrix can be written as RX and XX^TR^T takes $O(D^2)$ time, 1 Mark $\rightarrow$ the rest of the argument. If the student has pursued a different argument, partial marks to be awarded as necessary.

c The Lagrangian expressed in terms of only the parameters α_i is the following $\alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}(\sum_i \sum_j \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i.\boldsymbol{x}_j)$. In this case we see that this becomes $L = \alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}(2\alpha_1\alpha_2(-1)(8) + 2\alpha_1\alpha_3(-1)(4) + 2\alpha_2\alpha_3(1)(2) + 10\alpha_1^2 + 10\alpha_2^2 + 2\alpha_3^2)$. We need to maximize this subject to the criterion that $\alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3 = 0$. This allows us to substitute for α_1 in terms of the other two variables α_2 and α_3 to get an expression in two variables. We simplify the expression L to $2(\alpha_2 + \alpha_3) + 8(\alpha_2 +$

$\alpha_3)\alpha_2 + 4(\alpha_2+\alpha_3)\alpha_3 - 2\alpha_2\alpha_3 - 5(\alpha_2+\alpha_3)^2 - 5\alpha_2^2 - 1\alpha_3^2$. This expression finally simplifies to $L = -2\alpha_2^2 - 2\alpha_3^2 + 2\alpha_2 + 2\alpha_3$. This has four terms in it and two variables. If you substitute for α_2 or α_3 in terms of the other variables, you still end up with four terms and two variables.

Using Calculus we can set $\frac{\partial L}{\partial \alpha_2} = 0$ to get $2 - 4\alpha_2 = 0$ or $\alpha_2 = 0.5$. Similarly $\frac{\partial L}{\partial \alpha_2} = 0$ gives $2 - 4\alpha_3 = 0$ which gives $\alpha_3 = 0.5$. From $\alpha_1 = \alpha_2 + \alpha_3$ we conclude $\alpha_1 = 1$. Now the vector $w = \alpha_1 y_1 \boldsymbol{x}_1 + \alpha_2 y_2 \boldsymbol{x}_2 + \alpha_3 y_3 \boldsymbol{x}_3 = 1 * 1[-1,3]^T + 0.5 * -1 * [1,3]^T + 0.5 * -1 * [-1,1]^T = [-1,1]$. Solving for b from $\boldsymbol{w}^T\boldsymbol{x} + b = 1$ at $\boldsymbol{x} = [-1,3]^T$ gives $b = -3$. Thus the separating hyperplane is $-x_1 + x_2 - 3 = 0$.

Marking Scheme: 3 Marks $\rightarrow$ obtaining the expression for the Lagrangian in terms of the α_is. 2 Marks $\rightarrow$ substituting for one of the α_is in terms of the others and obtaining the simplest expression. 3 Marks $\rightarrow$ taking partial derivatives, setting them to zero and computing the parameters.

Q2 Alternative approaches would also be accepted and awarded marks

2.1 (a)

$$\|\lambda x + (1-\lambda)y\|_1 \leq \|\lambda x\|_1 + \|(1-\lambda)y\|_1$$

In this step we used triangle inequality of ℓ_1 norm (1 marks).
Now rhs can be seen to satify the following

$$\|\lambda x\|_1 + \|(1-\lambda)y\|_1 \leq |\lambda|\|x\|_1 + |(1-\lambda)|\|y\|_1$$

In this step we used homogeneity of norms. (0.5 marks)
Hence, it can be seen that

$$\|\lambda x + (1-\lambda)y\|_1 \leq \lambda\|x\|_1 + (1-\lambda)\|y\|_1$$

This is the definition of convexity. Hence ℓ_1 norm is a convex function . (0.5 marks)

(b) We need to show that $h(x)$ satisfy the definition of convex function as discussed in course.
It can be seen by definition of $h(x)$ that

$$h\big(\lambda x + (1-\lambda)y\big) = g\Big(A\big(\lambda x + (1-\lambda)y\big) + b\Big) \quad (0.25 marks)$$

we can simplify the expression as

$$g\Big(A\big(\lambda x + (1-\lambda)y\big) + b\Big) = g\Big(\lambda(Ax+b) + (1-\lambda)(Ay+b)\Big) \quad (0.5 marks)$$

Since $g(x)$ is a convex function it satifies the inequality

$$g\Big(\lambda(Ax+b) + (1-\lambda)(Ay+b)\Big) \leq \lambda.g(Ax+b) + (1-\lambda).g(Ay+b) \quad (0.5 marks)$$

From definition of $h(x)$ we know that

$$\lambda.g\big(Ax+b\big) + (1-\lambda).g\big(Ay+b\big) = \lambda.h(x) + (1-\lambda).h(y) \quad (0.5 marks)$$

Together from the previous 4 expressions, we can conclude that

$$h\big(\lambda x + (1-\lambda)y\big) \leq \lambda h(x) + (1-\lambda)h(y) \quad (0.25 marks)$$

Hence h(x) is a convex function . (2 marks)

2.2 Let the singular values of A in diagonal entries of Σ be named $\sigma_1, \sigma_2, ..., \sigma_n$.

(a) By defintion of frobenius norm we get

$$\gamma = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 \ \ (0.5 marks)$$

Now, we can see the following

$$\alpha = Tr(B) = Tr(A^T A) = Tr(V\Sigma^T \Sigma V^T) = Tr(\Sigma^T \Sigma) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 = \gamma \ \ (0.5 marks)$$

We have used property that $Tr(M_1 M_2) = Tr(M_2 M_1)$.
In conclusion claim made by G1 is True

(b) By defintion of frobenius norm we get

$$\gamma = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 \ \ (0.5 marks)$$

Now, we can see the following

$$\alpha = Tr(B) = Tr(A^T A) = Tr(V\Sigma^T \Sigma V^T) = Tr(\Sigma^T \Sigma) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 = \gamma \ \ (0.5 marks)$$

We have used property that $Tr(M_1 M_2) = Tr(M_2 M_1)$.
In conclusion claim made by G2 is False

(c) By defintion of frobenius norm we get

$$\gamma = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 \ \ (0.5 marks)$$

Now, we can see the following

$$\beta = Tr(C) = Tr(AA^T) = Tr(U\Sigma^T \Sigma U^T) = Tr(\Sigma^T \Sigma) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 = \gamma \ \ (1 marks)$$

.We have used property that $Tr(M_1 M_2) = Tr(M_2 M_1)$.
In conclusion $\beta \neq \gamma^2$(0.5 marks)

2.3 (a) First observe that

$$\|Ax - b\|_2^2 = x^T A^T Ax - 2x^T A^T b + b^T b \ \ (1 marks)$$

The gradient of the above part alone is easily derived based on identities learned in course as

$$2A^T Ax - 2A^T b \ \ (0.5 marks)$$

Now the gradient of the remaining part ie $c^T x + d$ is obtained as c (0.5 marks)
Hence $\nabla f(x) = 2A^T Ax - 2A^T b + c$

(b) First observe that

$$\|A_1^T A_1 x\|_2^2 = x^T A_1^T A_1 A_1^T A_1 x \ \ (0.5 marks)$$

The gradient of the above part alone is easily derived based on identities learned in course as

$$2A_1^T A_1 A_1^T A_1 x \ \ (0.5 marks)$$

Next we can see that

$$\|A_2^T x\|_2^2 = x^T A_2 A_2^T x \ \ (0.5 marks)$$

The gradient of the above part alone is easily derived based on identities learned in course as

$$2A_2A_2^Tx \ \ (0.5marks)$$

In summary $\nabla g(x) = 2A_1^TA_1A_1^TA_1x + 2A_2A_2^Tx$

Q3 Alternative approaches would also be accepted and awarded marks

a i)

$$\begin{aligned} & \det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0 \\ \Rightarrow \quad & (1-\lambda)(\lambda^2 - 2\lambda + (1-2\rho^2)) = 0 \text{ (0.5 marks)} \\ \Rightarrow \quad & \lambda = 1, 1 \pm \rho\sqrt{2} \text{ (1.5 marks)} \end{aligned}$$

Now for $\boldsymbol{A}$ to be positive definite all eigenvalues need to be positive as $\boldsymbol{A} = \boldsymbol{A}^T$. Therefore, we have

$$\begin{aligned} & 1 \pm \rho\sqrt{2} > 0 \\ \Rightarrow \quad & 1 + \rho\sqrt{2} > 0 \text{ and } 1 - \rho\sqrt{2} > 0 \\ \Rightarrow \quad & \frac{-1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}} \text{ (1 mark)} \end{aligned}$$

Thus, $\langle \mathbf{x}, \mathbf{y} \rangle_{\boldsymbol{A}} = \mathbf{x}^T\boldsymbol{A}\mathbf{y}$ is an innerproduct defined on $\mathbb{R}^3$ if $\frac{-1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$.

a ii)

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_{\boldsymbol{A}} &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ (0.5 mark)} \\ &= 0. \text{ (0.5 mark)} \end{aligned}$$

Let $\mathbf{z} = [z_1, z_2, z_3]^T$. $\mathbf{z}$ perpendicular to $\mathbf{x}$ will give us

$$\begin{aligned} \langle \mathbf{x}, \mathbf{z} \rangle_{\boldsymbol{A}} &= 0 \\ \Rightarrow \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} &= 0 \\ z_1 + \rho z_3 &= 0 \text{ (0.5 marks)} \end{aligned}$$

$\mathbf{z}$ perpendicular to $\mathbf{y}$ will give us

$$\begin{aligned} \langle \mathbf{y}, \mathbf{z} \rangle_{\boldsymbol{A}} &= 0 \\ \Rightarrow \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} &= 0 \\ z_2 + \rho z_3 &= 0 \text{ (0.5 marks)} \end{aligned}$$

The augmented matrix corresponding to the system of two equations can be written as $\begin{pmatrix} 1 & 0 & \rho & | & 0 \\ 0 & 1 & \rho & | & 0 \end{pmatrix}$. (0.5 marks)
Now putting $z_3 = t$, we get $z_1 = z_2 = -\rho t$.
Therefore $\mathbf{z} = \begin{pmatrix} -\rho t \\ -\rho t \\ t \end{pmatrix}$, $\forall t \in \mathbb{R}$. (0.5 marks)

b i) Now $\nabla_{\mathbf{x}} f = \begin{pmatrix} \mathbf{x}^T(\boldsymbol{Q} + \boldsymbol{Q}^T) \\ \boldsymbol{b}^T \end{pmatrix} = \begin{pmatrix} 2\mathbf{x}^T\boldsymbol{Q} \\ \boldsymbol{b}^T \end{pmatrix}$ as $\boldsymbol{Q}$ is symmetric. (1 mark)

b ii)

$$\begin{aligned} f(0,0,0) &= \begin{pmatrix} [0,0,0]\boldsymbol{Q}[0,0,0]^T \\ \boldsymbol{b}^T[0,0,0]^T \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \text{ (0.5 marks)} \\ \nabla_{\mathbf{x}} f(0,0,0) &= \begin{pmatrix} 2[0,0,0]\boldsymbol{Q} \\ \boldsymbol{b}^T \end{pmatrix} = \begin{pmatrix} \mathbf{0}^T \\ \boldsymbol{b}^T \end{pmatrix}. \text{ (0.5 marks)} \end{aligned}$$

where $\mathbf{0}^T = (0,0,0)$.
The linear approximation of f about $(0,0,0)$ is given by

$$\begin{aligned} T_1 f(\mathbf{x}) &= f(0,0,0) + \nabla_{\mathbf{x}} f(0,0,0)\mathbf{x} \\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{0}^T \\ \boldsymbol{b}^T \end{pmatrix}\mathbf{x} = \begin{pmatrix} 0 \\ \boldsymbol{b}^T\mathbf{x} \end{pmatrix} \end{aligned}$$

(1 mark)

c Let $\boldsymbol{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$
So, the problem is to $\min(\|\boldsymbol{M} - \boldsymbol{A}\|_2)^2 = (a-1)^2 + b^2 + c^2 + (d-2)^2$ such that $a + d = 0$. (0.5 marks)

Therefore, the Lagrangian of the constrained optimization problem is given by

$$L(a,b,c,d,\lambda) = (a-1)^2 + b^2 + c^2 + (d-2)^2 + \lambda(a+d) \text{ (0.5 marks)}$$

Partially differentiating with respect to a, b, c, d, λ, we get

$$\begin{aligned} \frac{\partial L}{\partial a} &= 2a - 2 + \lambda = 0 \Rightarrow a = \frac{-\lambda + 2}{2}, \\ \frac{\partial L}{\partial b} &= 2b = 0 \Rightarrow b = 0, \\ \frac{\partial L}{\partial c} &= 2c = 0 \Rightarrow c = 0, \\ \frac{\partial L}{\partial d} &= 2d - 4 + \lambda = 0 \Rightarrow d = \frac{-\lambda + 4}{2} \\ \frac{\partial L}{\partial \lambda} &= a + d = 0 \Rightarrow \lambda = 3 \end{aligned}$$

(0.5 marks for a, d in terms of λ, 0.5 marks for $b = c = 0$, 0.5 marks for finding value of λ)

$$\text{Thus, } \boldsymbol{M} = \begin{pmatrix} \frac{-1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \qquad \text{(0.5 marks)}$$