(1) i) A researcher used the gradient descent algorithm with momentum term to find the minimum of function $f(x, y) = 4x^2 + 3y^2$. Let the momentum/friction parameter used in this algorithm be referred to as β. Find the value of β if you are given the following information about the algorithm:

(i) Initial point of algorithm is $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

(ii) The iterates obtained after 3 iterations is given as $\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} -40.32 \\ -14.01 \end{bmatrix}$

(iii) A fixed step size is used for all iterations and its value is $\alpha = 0.5$

(4 marks)

ii) Consider a quadratic function $f(x_1, x_2) = x_1^2 + \beta x_2^2$ where $\beta \in \mathbb{R}$ is an unknown constant. Also assume that $\beta > 0$. Consider the problem of minimizing this function using basic gradient descent algorithm. Assume that the gradient descent is initialized as follows:

$$\mathbf{x}^{(0)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

The step size value α at each iteration is obtained by applying a decay algorithm. Assume that $\alpha_0 = .8$ was used in the initial iteration $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha_0 \nabla f(\mathbf{x}^{(0)})$. Derive the value of the step size α_1, α_2 for the next two iterations if we use

(i) exponential decay algorithm with parameter $K = 3\beta$

(ii) inverse decay algorithm with parameter $K = 4\beta$

(2 marks)

i) 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 $\nabla f = \begin{bmatrix} 8x \\ 6y \end{bmatrix}$ and $x_0 = 2, y_0 = 3$

(i) $x_1 = x_0 - \alpha 8 x_0 = 2 - \frac{1}{2}.16 = -6$ (0.5 mark)
(ii) $x_2 = x_1 - \alpha 8 x_1 + \beta(x_1 - x_0) = 18 - 8\beta$ (1 mark)
(iii) $x_3 = x_2 - \alpha 8 x_2 + \beta(x_2 - x_1) = -8\beta^2 + 48\beta - 54$ (1.5 mark)

From above $x_3 = -8\beta^2 + 48\beta - 54 = -40.32$.
Solving above quadratic equation in β, we get $\beta = 0.3$ or $\beta = 5.7$.
Since $\beta \in (0, 1)$, we get $\beta = 0.3$
(1 mark)

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

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.

(i) $y_1 = y_0 - \alpha 6 y_0 = -6$ (0.5 mark)
(ii) $y_2 = y_1 - \alpha 6 y_1 + \beta(y_1 - y_0) = (12 - 9\beta)$ (1 mark)
(iii) $y_3 = y_2 - \alpha 6 y_2 + \beta(y_2 - y_1) = -9\beta^2 + 36\beta - 24$ (1.5 mark)

From above $y_3 = -9\beta^2 + 36\beta - 24 = -14.01$
Solving above quadratic equation in β, we get $\beta = 0.3$ or $\beta = 3.7$.
Since $\beta \in (0, 1)$, we get $\beta = 0.3$ (1 mark)

ii) The initial value $\alpha_0 = 0.8$.

(i) $\alpha_1 = \alpha_0 e^{-K*1} = 0.8e^{-3\beta}$ (0.5 marks)

$\alpha_2 = \alpha_0 e^{-K*2} = 0.8e^{-6\beta}$ (0.5 marks)

(ii) $\alpha_1 = \frac{\alpha_0}{1+K*1} == \frac{0.8}{1+4\beta}$ (0.5 marks)

$\alpha_2 = \frac{\alpha_0}{1+K*2} == \frac{0.8}{1+8\beta}$ (0.5 marks)

(2) Consider the 5 dimensional data matrix given by

$$\boldsymbol{X} = \sqrt{3}\begin{bmatrix} 1 & 1 & -2 \\ 1 & -1 & 0 \\ 2 & -2 & 0 \\ -1 & 1 & 0 \\ -2 & 2 & 0 \end{bmatrix}$$

where each row in $\boldsymbol{X}$ represent 1 feature.

i) Help the data scientist to find out the unit vector along each of the first 2 principal components. [6 Marks]

Solution **(Kindly award marks for any alternate correct method appropriately)**

i) Clearly along each dimension the mean is zero.

Therefore, $\boldsymbol{S} = \frac{1}{3}\boldsymbol{X}\boldsymbol{X}^T$. But, consider

$$\frac{1}{3}\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 11 & -9 & -2 \\ -9 & 11 & -2 \\ -2 & -2 & 4 \end{bmatrix}$$

$$\Rightarrow \quad \det\left(\frac{1}{3}\boldsymbol{X}^T\boldsymbol{X} - \lambda\boldsymbol{I}\right) = 0$$

$$\Rightarrow \quad -\lambda^3 + 26\lambda^2 - 120\lambda = 0$$

$$\Rightarrow \quad \lambda = 0, 6, 20$$

$\Rightarrow$ The first 2 largest eigenvalues of $\boldsymbol{S}$ are $20, 6$.

[1 Marks] for steps and [1 Mark] for correct eigenvalues.

By solving $\left(\frac{1}{3}\boldsymbol{X}^T\boldsymbol{X} - 20\boldsymbol{I}\right)\boldsymbol{y} = 0$

we get $\boldsymbol{y} = [-t, t, 0]^T \ \forall t \neq 0$.

Therefore, $\boldsymbol{v}_1 = [-1, 1, 0]^T$ is an eigenvector corresponding to 20.

So, $\boldsymbol{X}\boldsymbol{v}_1 = [0, -2, -4, 2, 4]^T$ is an eigenvector of $\boldsymbol{S}$ corresponding to 20.

Therefore unit vector along first principal component is $[0, -1/\sqrt{10}, -2/\sqrt{10}, 1/\sqrt{10}, 2/\sqrt{10}]^T$.

Similarly, by solving $\left(\frac{1}{3}\boldsymbol{X}^T\boldsymbol{X} - 6\boldsymbol{I}\right)\boldsymbol{y} = 0$

we get $\boldsymbol{y} = [-t/2, -t/2, t]^T \ \forall t \neq 0$.

Therefore, $\boldsymbol{v}_2 = [-1, -1, 2]^T$ is an eigenvector corresponding to 6.

So, $\boldsymbol{X}\boldsymbol{v}_2 = [-6, 0, 0, 0, 0]^T$ is an eigenvector of $\boldsymbol{S}$ corresponding to 6.

Therefore unit vector along second principal component is $[-1, 0, 0, 0, 0]^T$.

[2 Marks] for steps and [2 Marks] for correct eigenvectors of unit norm. (Kindly note that the eigenvectors with unit norm are not unique.)

(3) Consider the function $f(x) = 9x^5 - 15x^3$. From first and second derivative tests one can find that the points of local maximum and local minimum are $x = -1$ and $x = 1$ respectively. Suppose you use Gradient Descent algorithm to find the point x at which the function $f(x)$ takes local minimum with initial value $x_0 = 0.5$ and the learning parameter $\alpha = 0.05$.
   (a) Find x_1, x_2, and x_3 using first three iterations. Is there any problem in convergence? [6 Marks]
   (b) For the same problem use adapt AdaGrad method to find x_1. What is your observation? [2 Marks]

   Solution:
   (a) Find x_1, x_2, and x_3 using first three iterations. Is there any problem in convergence? [6M]
   $f'(x) = 45x^4 - 45x^2 = 45x^2\,(x^2 - 1)$ [0.5M]
   $x_1 = 0.5 - 0.05\,(45 * 0.5^2)\,(0.5^2 - 1) = 0.5 + 0.42175 = 0.92175$[1.5M]
   $x_2 = 0.92175 - 0.05\,(45 * 0.92175^2)\,(0.92175^2 - 1) = 0.92175 + 0.280918 = 1.20918$[1.5M]
   $x_3 = 1.20918 - 0.05\,(45 * 1.20918^2)\,(1.20918^2 - 1) = 1.20918 - 1.52027 = -0.31109$ [1.5M]
   We observe the problem of overshooting due to large step size. Also, from the third iteration we observe that the values are oscillating around the optimal solution. [1M]

   (b) $A_1 = f'\,(x_0)^2 = 8.435^2$

$$x_1 = 0.5 - \frac{0.05}{\sqrt{8.435^2}}\left(45 * 0.5^2\right)\left(0.5^2 - 1\right) = 0.5 - \frac{0.05}{8.435}\left(45 * 0.5^2\right)\left(0.5^2 - 1\right) = 0.5\frac{0.42175}{8.435}$$
$$= 0.5 + .05 = 0.55[1.5\text{M}]$$

We observe that the problem of large step size is avoided. [0.5M]

(4) Let A be a normal matrix i.e $A^H A = AA^H$ over $\mathcal{C}$(i.e complex numbers) . Show that if $\mathbf{x}$ is an eigenvector of A with eigenvalue λ, then $\mathbf{x}$ is an eigenvector of A^H with eigenvalue $\bar{\lambda}$. **Note that you are not allowed to use the fact that normal matrices are unitarily diagonalizable to solve this problem**. [6 Marks]
   Solution:
   Since $AA^H = A^H A$ and $A^{HH} = A$, we have

$$(Ax)^H A\mathbf{x} = x^H A^H A\mathbf{x} = \mathbf{x}^H AA^H\mathbf{x} = (A^H\mathbf{x})^H A^H\mathbf{x}$$

It follows that

$$\begin{aligned}
0 &= (A\mathbf{x} - \lambda\mathbf{x})^H(A\mathbf{x} - \lambda\mathbf{x}) \\
&= (\mathbf{x}^H A^H - \bar{\lambda}\mathbf{x}^H)(A\mathbf{x} - \lambda\mathbf{x}) \\
&= \mathbf{x}^H A^H A\mathbf{x} - \bar{\lambda}\mathbf{x}^H A\mathbf{x} - \lambda\mathbf{x}^H A^H\mathbf{x} + \lambda\bar{\lambda}\mathbf{x}^H\mathbf{x} \\
&= \mathbf{x}^H AA^H\mathbf{x} - \bar{\lambda}\mathbf{x}^H A\mathbf{x} - \lambda\mathbf{x}^H A^H\mathbf{x} + \lambda\bar{\lambda}\mathbf{x}^H\mathbf{x} \\
&= (A^H\mathbf{x} - \bar{\lambda}\mathbf{x})^H(A^H\mathbf{x} - \bar{\lambda}\mathbf{x}).
\end{aligned}$$

So, we have an eigenvalue equation $A^H\mathbf{x} = \bar{\lambda}\mathbf{x}$, which implies that $\mathbf{x}$ is an eigenvector of A^H corresponding to eigenvalue $\bar{\lambda}$.

Suggested Marking Scheme 2 Marks $\rightarrow$ recognizing that $0 = (A\mathbf{x} - \lambda\mathbf{x})^H(A\mathbf{x} - \lambda\mathbf{x})$, 4 Marks $\rightarrow$ completing the rest of the argument

(5) Consider the primal optimization problem

$$\begin{aligned}
\min x + y \text{ subject to} \\
2x + 9y &\leq 12 \\
3x + 5y &\leq 8 \\
5x + 14y &\geq \alpha
\end{aligned}$$

(a) What is the smallest integer value of α in the above formulation so that $\min_{x,y} \max_{\lambda_1,\lambda_2\lambda_3\geq 0} L(x, y, \lambda_1, \lambda_2, \lambda_3) = \infty$? Note that $L(x, y, \lambda_1, \lambda_2, \lambda_3)$ is the Lagrangian of the formulation and $\lambda_i \geq 0$ are Lagrange multipliers. Give detailed justification for your answer. (4 Marks)

(b) Let $\alpha = 1$ in the above formulation. Using Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3 \geq 0$, set up the dual of the formulation and show that the objective function of the dual can be written as $c_1 + c_2\lambda_3$ where c_1 and c_2 are numerical constants and λ_3 is the Lagrange multiplier for the third constraint above. (4 Marks)

Solution

(a) The Lagrangian of the problem can be set up as $L(x, y, \lambda_1, \lambda_2, \lambda_3) = x + y + \lambda_1(2x + 9y - 12) + \lambda_2(3x + 5y - 8) + \lambda_3(-5x - 14y + \alpha)$. For a given x, y, if one or more of the expressions $(2x + 9y - 12$, $3x + 5y - 8$, $-5x - 14y + \alpha$ is positive then the associated Lagrange multiplier can be made to go to infinity so that $L(x, y, \lambda_1, \lambda_2, \lambda_3)$ becomes infinity. Thus for $\min_{x,y} \max_{\lambda_1,\lambda_2\lambda_3\geq 0} L(x, y, \lambda_1, \lambda_2, \lambda_3)$ to be infinity we need to ensure that for every x, y one of the expressions $2x + 9y - 12$, $3x + 5y - 8$, $-5x - 14y + \alpha$ be positive. This can only happen if the given set of constraints in the problem are infeasible. Adding the first two constraints gives $5x + 14y \leq 20$. If we make the last constraint $5x + 14y \geq 21$, we will ensure infeasibility. $\alpha = 21$ is the smallest integer that will ensure this condition.

Suggested Marking Scheme: 2 Marks $\rightarrow$ recognizing that for $\min_{x,y} \max_{\lambda_1,\lambda_2\lambda_3\geq 0} L(x, y, \lambda$ ∞, we need to have one or more of the expressions $(2x + 9y - 12$, $3x + 5y - 8$, $-5x - 14y + \alpha$ become positive for each x, y, 2 Marks $\rightarrow$ rest of the argument.

(b) Multiplying the last constraint by -1, we can put the given primal problem in standard form. The dual formulation then becomes

$$\max -\lambda^T b \text{ subject to}$$

$$\begin{bmatrix} 2 & 3 & -5 \\ 9 & 5 & -14 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0$$

$$\lambda \geq 0$$

Solving the equation $A^T\lambda = -c$ gives us

$$\begin{bmatrix} 2 & 3 & -5 \\ 9 & 5 & -14 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

We find the particular solution $\begin{bmatrix} 2/17 \\ -7/17 \\ 0 \end{bmatrix}$ and the homogeneous solution $\lambda_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, so that $\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \lambda_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2/17 \\ -7/17 \\ 0 \end{bmatrix}$ so that the objective function $\lambda^T b = -\lambda_3 \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 12 \\ 8 \\ -1 \end{bmatrix} + \begin{bmatrix} 2/17 & -7/17 & 0 \end{bmatrix} \begin{bmatrix} 12 \\ 8 \\ -1 \end{bmatrix} = -32/17 - 19\lambda_3$.

Suggested Marking Scheme: 1 Mark $\rightarrow$ coming up with the dual formulation, 3 Marks $\rightarrow$ writing the objec
tive function in the correct form.

(6) Let $(x_1, 1), (x_2, 1), \ldots (x_N, 1)$ be N tuples, where the x_i represent points in D-dimensional space, and the points are all positively labeled. Similarly let $(z_1, -1), (z_2, -1), \ldots (z_N, -1)$ be N tuples which are negatively labeled. We form the point $(x_{N+1}, 1)$ where $x_{N+1} = \alpha_1 x_{i_1} + \alpha_2 x_{i_2} + \ldots \alpha_k x_{i_k}$, $0 \leq \alpha_j \leq 1, 1 \leq j \leq k, \sum_{j=1}^{j=k} \alpha_j = 1$ where the points $x_{i_1}, x_{i_2}, \ldots x_{i_k}$ are k points that are all positively labeled. Assume that the dataset is linearly separable.

(a) Calculate the hinge loss of the point x_{N+1} for the optimal SVM separating hyperplane of the form $w^T x + b = 0$. [4 Marks]

(b) What condition needs to be true on the points $x_{i1}, x_{i2}, \ldots x_{ik}$ if we find that x_{N+1} is a support vector? Justify your answer mathematically. [4 Marks]

Solution

(1) The hinge loss for the point x_{N+1} is $\max(0, 1 - y_{N+1}(w^T x_{N+1} + b))$. We can rewrite this as $\max(0, 1 - y_{N+1}(w^T(\sum_{j=1}^{j=k} \alpha_j x_{i_j}) + (\sum_{j=1}^{j=k} \alpha_j)b = 1 - y_{N+1}(\sum_{j=1}^{j=k} \alpha_j (w^T x_{i_j} + b_j)$. Since the x_{i_j} are all positively labeled it must be the case that $w^T x_{i_j} + b \geq 1$ for $x_{i_1}, x_{i_2}, \ldots x_{i_k}$. Then we see that $1 - y_{N+1}(\sum_{j=1}^{j=k} \alpha_j (w^T x_{i_j} + b_j) \leq 1 - y_{N+1} \sum_{j=1}^{j=K} \alpha_j \leq 1 - y_{N+1} \leq 0$. Here we use the fact that the label y_{N+1} for the $(N+1)th$ point is $+1$. Thus

the hinge loss turns out to be zero for the $(N+1)th$ point.
Suggested Marking Scheme: 2 Marks $\rightarrow$ recognizing that the hinge loss for x_{N+1} can be written in terms of the points of which it is a convex combination. 2 Marks $\rightarrow$ remaining argument.

(2) For x_{N+1} to be a support vector, we need $w^T x_{N+1} + b = 1$, but since $x_{N+1} = \sum \alpha_j x_{i_j}$ we can rewrite this condition as $w^T(\sum \alpha_j x_{i_j}) + (\sum \alpha_j) b = 1$ which further means $\sum \alpha_j (w^T x_{i_j} + b) = 1$. But we know that $w^T x_{i_j} + b \geq 1$, which means that the only way the given condition is true is for all x_{i_j} to be support vectors because $w^T x_{i_j} + b$ must be equal to 1 for all the given points.

Suggested Marking Scheme: 1 Mark $\rightarrow$ recognizing the condition on x_{N+1} for it to be a support vector, 3 Marks $\rightarrow$ showing that the points on which x_{N+1} depends must also be support vectors.