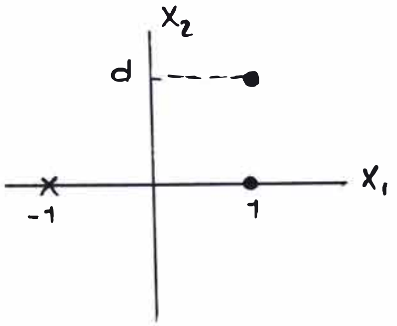
1. (15 MARKS) Consider a binary linear classification where the data points are two dimensional, i.e., $\mathbf{x} \in \mathbb{R}^2$ and the labels $y \in \{-1, 1\}$. The training set consists of the following points:

$$\mathbf{x}_1 = (1,0)^T, \qquad y_1 = +1$$
$$\mathbf{x}_2 = (-1,0)^T, \qquad y_2 = -1$$
$$\mathbf{x}_3 = (1,d)^T, \qquad y_3 = +1$$

Assume that $d > 0$ is some constant. **For the perceptron algorithm please treat the points that fall exactly on the decision boundary as mistakes and do the update accordingly. Assume that the initial weight is the all-zero vector.**

1 mark (a) Is the data-set linearly separable? Sketch the points in the $x_1 - x_2$ plane and show the labels.



Yes, data-set is linearly separable

⇒ Either full or zero grade

6 marks (b) Run the perceptron algorithm in the following order $(1,2,3), (1,2,3), \ldots$ until it converges. Show the output of the perceptron algorithm in each step, sketch the resulting decision boundary, and find the resulting margin (distance of the decision boundary to the nearest training point). s

⇒ 1 mark per step & 1 mark for sketch and margin

⇒ Not accounting for bias or not initializing to zeros → -3

$$w^{(0)} = [0 \quad 0 \quad 0]^T$$

Step 1 : $\hat{y}_1 = w^{(0)T} x_1 = [0 \; 0 \; 0] \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = 0 \rightarrow$ misclassified

$$w^{(1)} \leftarrow w^{(0)} + y_1 x_1 \Longrightarrow w^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Step 2 : $\hat{y}_2 = w^{(1)T} x_2 = [1 \; 1 \; 0] \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = 0 \longrightarrow$ misclassified

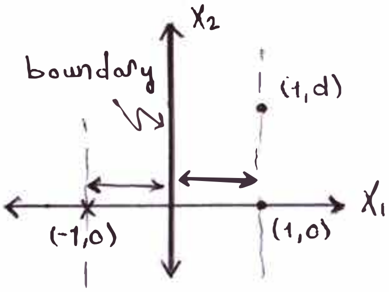
$$w^{(2)} \leftarrow w^{(1)} + y_2 x_2 \Longrightarrow w^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$$

Step 3 : $\hat{y}_3 = w^{(2)T} x_3 = [0 \; 2 \; 0] \begin{bmatrix} 1 \\ 1 \\ d \end{bmatrix} = 2 \Rightarrow +ve \longrightarrow$ correct

Step 4 : $\hat{y}_1 = w^{(2)T} x_1 = 2 \Rightarrow +ve \longrightarrow$ correct

Step 5 : $\hat{y}_2 = w^{(2)T} x_2 = -2 \Rightarrow -ve \longrightarrow$ correct

Margin = 1



7 marks (c) Repeat part (b) when the order is (3, 2, 1), (3, 2, 1, ), .... Also comment on what the decision boundary approaches when $d \to \infty$

⇒ 1 mark per step, 2 marks for margin, 1 for $d \to \infty$

$w^{(c)T} = [0 \quad 0 \quad 0]$ ⇒ Not accounting for bias or not initializing to zero → $\underline{\underline{-3}}$

Step 1: $\hat{y}_3 = w^{(c)T} x_3 = [0 \; 0 \; 0] \begin{bmatrix} 1 \\ 1 \\ d \end{bmatrix} = 0 \longrightarrow$ misclassified

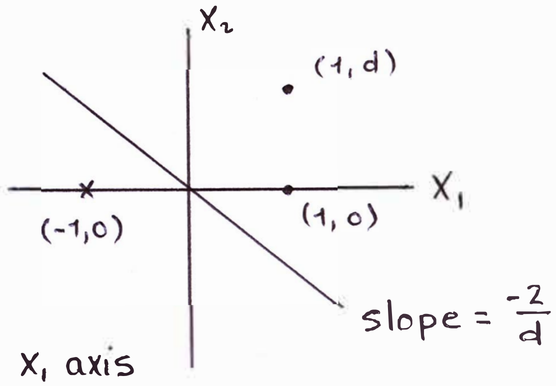
$w^{(1)} \leftarrow w^{(0)} + y_3 x_3 = \begin{bmatrix} 1 \\ 1 \\ d \end{bmatrix}$

Step 2: $\hat{y}_2 = w^{(1)T} x_2 = [1 \; 1 \; d] \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = 0 \longrightarrow$ misclassified

$w^{(2)} \leftarrow w^{(1)} + y_2 x_2 = \begin{bmatrix} 1 \\ 1 \\ d \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ d \end{bmatrix}$

Step 3: $\hat{y}_1 = w^{(2)T} x_1 = [0 \; 2 \; d] \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = 2 \Rightarrow +ve \longrightarrow$ correct

Step 4: $\hat{y}_3 = w^{(2)T} x_3 = [0 \; 2 \; d] \begin{bmatrix} 1 \\ 1 \\ d \end{bmatrix} = +2 \Rightarrow +ve \longrightarrow$ correct



∴ Boundary is $2X_1 + dX_2 = 0$

$$X_2 = \frac{-2}{d} X_1$$

$$\text{Margin} = \frac{|(2 \times 1) + (d \times 0)|}{\sqrt{(2)^2 + (d)^2}} = \frac{2}{\sqrt{4 + d^2}}$$

As $d \to \infty$: slope $\to 0$ ∴ Boundary is $X_1$ axis ($X_2 = 0$)

1 mark (d) Between parts (b) and (c) which solution will be preferred? Briefly explain.

(b) is preferred as it has bigger margin

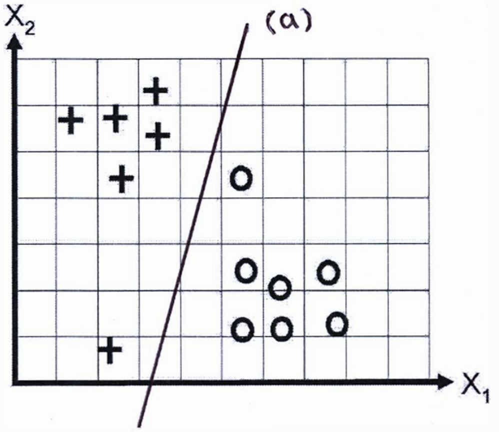
$$\left( \frac{2}{\sqrt{4 + d^2}} < 1 \right)$$

⇒ Either full or zero grade

total/8

10 marks

**2.** Consider a binary linear classification problem where $\mathbf{x} \in \mathbb{R}^2$ and $y \in \{-1,+1\}$. We illustrate the training dataset below. The '+' label refers to $y = +1$ and the 'o' label refers to $y = -1$. We would like to construct a classifier $h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(w_0 + w_1x_1 + w_2x_2)$ where sign(·) is the *sign* function as discussed in class.



In the figure above, the adjacent vertical (and horizontal) lines are 1 unit apart from each other. Assume that the training points are above are $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ (with $N = 13$). We consider the classification loss

$$L(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(y_i \neq h_{\mathbf{w}}(\mathbf{x}_i))$$

where $\mathbb{I}$ denotes the indicator function.

2 marks (a) Draw a decision boundary in the figure above that achieves zero training error.

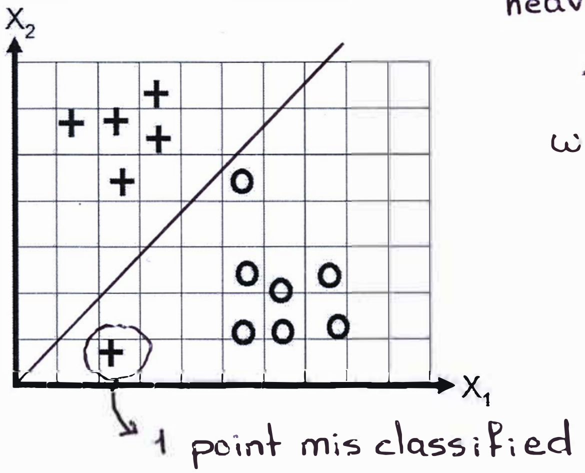
4 marks (b) Suppose that we attempt to minimize the following loss function over $\mathbf{w}$ : $J(\mathbf{w}) = L(\mathbf{w}) + \lambda w_0^2$, where $\lambda = 10^7$ is a huge constant. Sketch a possible decision boundary in the figure below. How many points are mis-classified.

heavily penalizing $w_0$

$\therefore w_0 \rightarrow 0$

$w_1 x_1 + w_2 x_2 = 0$

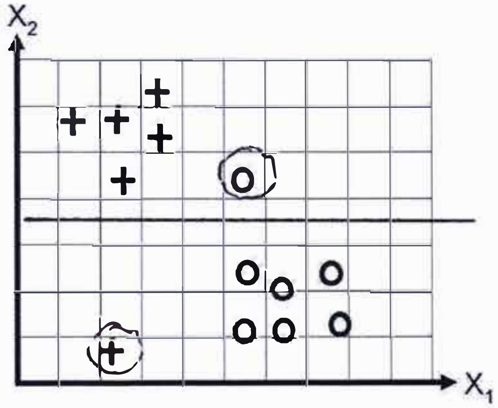
is the boundary



total/6

2 marks

(c) Suppose that we attempt to minimize the following loss function over $\mathbf{w}$ : $J(\mathbf{w}) = L(\mathbf{w}) + \lambda w_1^2$, where $\lambda = 10^7$ is a huge constant. Sketch a possible decision boundary in the figure below. How many points are mis-classified.



heavily penalizing $w_1$

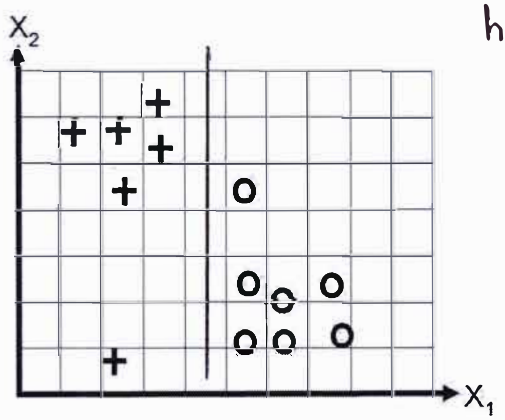
$\therefore w_1 \to 0$

$w_0 + w_2 x_2 = 0$ is the boundary

2 points misclassified

2 marks

(d) Suppose that we attempt to minimize the following loss function over $\mathbf{w}$ : $J(\mathbf{w}) = L(\mathbf{w}) + \lambda w_2^2$, where $\lambda = 10^7$ is a huge constant. Sketch a possible decision boundary in the figure below. How many points are mis-classified.



heavily penalizing $w_2$

$\therefore w_2 \to 0$

$w_0 + w_1 x_1 = 0$ is the boundary

No points misclassified

$\Rightarrow$ For parts (b), (c) & (d), half the question grade for sketching & the other half for number of misclassified points

total/4

25 marks **3**. Suppose that a data vector $\mathbf{y} \in \mathbb{R}^n$ and a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ are given. Assume that $n > p$ and that $\mathbf{A}^T\mathbf{A}$ is an $p \times p$ invertible matrix. We would like to approximate $\mathbf{y}$ using a vector of the form $\hat{\mathbf{y}}_{\mathbf{w}} = \mathbf{A}\mathbf{w}$.

2 marks (a) Let $\mathbf{w}_{\text{LS}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{y}_{\mathbf{w}}\|^2$, be the least square estimate of $\mathbf{y}$. Provide an expression for $\mathbf{w}_{\text{LS}}$ and the associated $\hat{\mathbf{y}}_{\text{LS}} = A\mathbf{w}_{\text{LS}}$. No calculation is required in this part.

$$W_{LS} = (A^TA)^{-1} A^T y$$

$$\hat{y}_{LS} = A(A^TA)^{-1} A^T y$$

⇒ 1 point for each

4 marks (b) Show that for any $\mathbf{w} \in \mathbb{R}^p$ the following identity holds:

$$\|\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{w}}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}_{\text{LS}}\|^2 + \|\hat{\mathbf{y}}_{\mathbf{w}} - \hat{\mathbf{y}}_{\text{LS}}\|^2$$

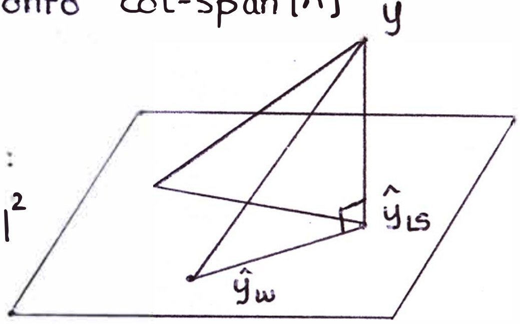
No calculation is needed for this part, just a clear geometric argument and an accompanying figure that uses properties of $\hat{\mathbf{y}}_{\text{LS}}$ is sufficient.

$\hat{y}_{LS}$ is the projection of $y$ onto col-span{A}

$\therefore$ $y - \hat{y}_{LS}$ is $\perp$ to $\hat{Y}_w - \hat{Y}_{LS}$

Using pythagoras theorem:

$$\|y - \hat{y}_w\|^2 = \|y - \hat{y}_{LS}\|^2 + \|\hat{y}_w - \hat{y}_{LS}\|^2$$



⇒ 2 points for sketch

⇒ 2 points for explanation

6 marks (c) **For parts (c) and (d) of this question assume that $\mathbf{A}^T\mathbf{A}$ is a diagonal matrix and $\lambda > 0$ is a constant.** Let $a_1, a_2, \ldots, a_p$ be the elements on the diagonal of $\mathbf{A}^T\mathbf{A}$, assumed to be non-zero. Let $\mathbf{w}^* = \arg\min_{\mathbf{w}} \{\|\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{w}}\|^2 + \lambda\|\mathbf{w}\|^2\}$. Using part (b) show that

$$w_i^* = \arg\min_w \{a_i(w - w_{\text{LS},i})^2 + \lambda w^2\}, \quad i = 1, 2, \ldots, p$$

where $w_i^*$ and $w_{\text{LS},i}$ denote the $i$-th elemenet of $\mathbf{w}^*$ and $\mathbf{w}_{\text{LS}}$ respectively. Hence express $w_i^*$ in terms of $w_{\text{LS},i}$, $a_i$ and $\lambda$.

⇒ 3 points for $w^*$

$$w^* = \arg\min_w \{\|y - \hat{y}_w\|^2 + \lambda\|w\|^2\} \quad \text{(using part (b))}$$

$$= \arg\min_w \{\|y - \hat{y}_{LS}\|^2 + \|\hat{y}_w - \hat{y}_{LS}\|^2 + \lambda\|w\|^2\} \quad \left(\|y - \hat{y}_{LS}\|^2 \text{ is independant of } w\right)$$

$$= \arg\min_w \{\|\hat{y}_w - \hat{y}_{LS}\|^2 + \lambda\|w\|^2\}$$

$$= \arg\min_w \{(w - w_{LS})^T A^TA (w - w_{LS}) + \lambda w^T w\}$$

⇒3 point for getting $w_i^*$

$\because$ $A^TA$ is a diagonal matrix

$$\therefore w_i^* = \arg\min_w \{a_i(w_i - w_{LS_i})^2 + \lambda w_i^2\} = \arg\min_w \{h(w_i)\}$$

$$h'(w_i) = 0 \longrightarrow 2a_i(w_i - w_{LS_i}) + 2\lambda w_i = 0$$

$$\therefore w_i = \frac{a_i}{a_i + \lambda} w_{LS_i}$$

total/12

4 marks

(d) Suppose we wish to compute:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{w}}\|^2 + \lambda \sum_{i=1}^{p} |w_i| \right\}$$

where $|\cdot|$ is the absolute value. Using the method analogous to part (c) express $w_i^*$ (the $i$-th component of $\mathbf{w}^*$) in terms of $w_{\mathrm{LS},i}$, $a_i$ and $\lambda$. *Hint: Consider the function* $f(x) = (x-c)^2 + \eta|x|$, *where* $\eta > 0$. *The minimizing value of* $f(x)$, *say* $x_0$, *can be expressed as follows: If* $c > \eta/2$ *then* $x_0 = c - \eta/2$, *if* $c < -\eta/2$ *then* $x_0 = c + \eta/2$ *and if* $|c| \leq \eta/2$ *then* $x_0 = 0$.

from part (c) : $w_i^* = \text{argmin}\{ a_i(w_i - w_{LS_i})^2 + \lambda|w_i|$

$= \text{argmin}\{(w_i - w_{LS_i})^2 + \frac{\lambda}{a_i}|w_i|\}$

(Using the hint given)

$$= \begin{cases} w_{LS_i} - \lambda/2a_i &, \ w_{LS_i} > \lambda/2a_i \\ w_{LS_i} + \lambda/2a_i &, \ w_{LS_i} < -\lambda/2a_i \\ 0 &, \ \text{otherwise} \end{cases}$$

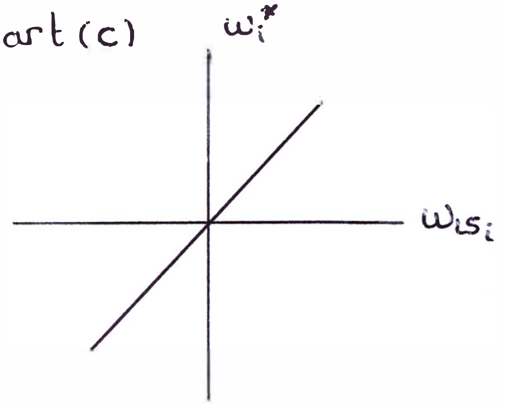
⇒ 2 points for putting $w_i^*$ in the correct format

⇒ 2 points for using the hint

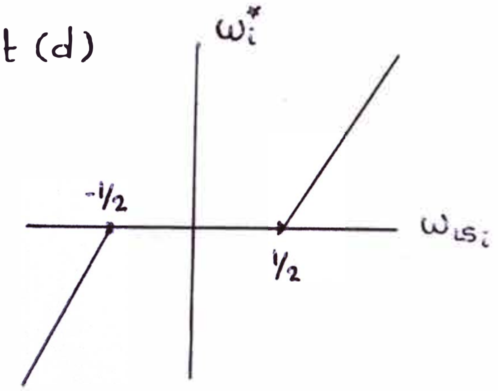
4 marks

(e) Sketch $w_i^*$ as a function of $w_{\mathrm{LS},i}$ for $\lambda = 1$ and $a_i = 1$ for both parts (c) and (d). Also explain qualitatively how does your answer in the two cases differ for large values of $\lambda$?

part (c)



part (d)



for large $\lambda$ : $\because w_i^* = \frac{a_i}{a_i + \lambda} w_{LS_i}$

$\lambda \to \infty \quad w_i^* \to 0$

for large $\lambda$ : $w_i^* = 0$

⇒ 1 point for sketch & 1 point for large $\lambda$ in each part

5 marks

(f) Suppose that we have $N$ training examples: $(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_N, y_N)$ where $\mathbf{x}_i \in \mathbb{R}^p$ and we wish to minimize the following loss function:

$$J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2 + \frac{\lambda}{N}\sum_{j=1}^{p}|w_j|,$$

where $\mathbf{w} = (w_1, \ldots, w_p)^T$. Provide a SGD update rule for minimizing $J(\mathbf{w})$. In each step assume that one example is selected at random in each update, and the gradient from the regularization term is added in each step.

$$\underline{W}(t+1) \leftarrow \underline{W}(t) - \eta_t \nabla_{\underline{W}(t)} e_n$$

$$e_n = (X_n^T W - y_n)^2 + \lambda \sum_{j=1}^{P} |w_j|$$

$$\frac{\partial e_n}{\partial w_j} = 2(X_n^T W - y_n)\, x_j + \lambda \operatorname{sign}(w_j)$$

$$\nabla_{\underline{W}(t)} e_n = 2(X_n^T \underline{W}(t) - y_n)\, \underline{X}_n + \lambda \operatorname{sign}(\underline{W}(t))$$

⇒ 2 points for update rule (Knowing that N = 1)

⇒ 3 points for derivatives

total/5

10 marks **4.** Consider a logistic regression binary classification model that given $\mathbf{x} \in \mathbb{R}^2$ outputs:

$$P_{\mathbf{w}}(y=1|\mathbf{x}) = \frac{1}{1+e^{-(w_1x_1+w_2x_2)}}, \qquad P_{\mathbf{w}}(y=-1|\mathbf{x}) = \frac{1}{1+e^{(w_1x_1+w_2x_2)}}.$$

Assume that the bias term in the model is set to zero for simplicity in this problem. Suppose that we have only two training points:

$$\mathbf{x}_1 = (1,1), \quad y_1 = 1$$
$$\mathbf{x}_2 = (1,0), \quad y_2 = -1$$

We intend to select $\mathbf{w}$ that minimizes the following regularized loss function:

$$J(\mathbf{w}) = -\sum_{i=1}^{2} \log P_{\mathbf{w}}(y_i|\mathbf{x}_i) + \lambda\|\mathbf{w}\|^2$$

5 marks (a) Suppose that $\lambda = 0$. What will the optimal choice of $\mathbf{w}$ be? What will the resulting value of $J(\mathbf{w})$ be?

$$\because \omega_0 = 0, \ \lambda = 0 \ : \ J(\omega) = -\log\left(\frac{1}{1+e^{-(\omega_1+\omega_2)}}\right) - \log\left(\frac{1}{1+e^{\omega_1}}\right)$$

$$= \log(1+e^{-(\omega_1+\omega_2)}) + \log(1+e^{\omega_1})$$

To minimize $J(\omega) \Rightarrow \omega_1 = -K, \ \omega_2 = 2K$ & $K \to \infty$

$$\therefore J(\omega) = 0$$

$\Rightarrow$ 2 points for writing eq. of $J(\omega)$

$\Rightarrow$ 2 points for $\omega \to \infty$, 1 point for $J(\omega) = 0$

5 marks (b) Suppose that $\lambda$ is a large constant such that it only suffices to consider $\|\mathbf{w}\| \ll 1$ when minimizing $J(\mathbf{w})$. In this case we can approximate

$$\log(1+e^{-y_i \cdot \mathbf{w}^T\mathbf{x}_i}) \approx \log 2 - \frac{1}{2}y_i \cdot \mathbf{w}^T\mathbf{x}_i$$

Assuming that the above approximation is exact find $\mathbf{w}$ that minimizes $J(\mathbf{w})$.

$$J(\omega) = \log(1+e^{-(\omega_1+\omega_2)}) + \log(1+e^{\omega_1}) + \lambda(\omega_1^2+\omega_2^2)$$

$$= \log 2 - \tfrac{1}{2}(\omega_1+\omega_2) + \log 2 + \tfrac{1}{2}\omega_1 + \lambda(\omega_1^2+\omega_2^2)$$

$$\left.\begin{array}{l} \dfrac{\partial J(\omega)}{\partial \omega_1} = -\tfrac{1}{2} + \tfrac{1}{2} + 2\lambda\omega_1 = 0 \longrightarrow \omega_1 = 0 \\[2ex] \dfrac{\partial J(\omega)}{\partial \omega_2} = -\tfrac{1}{2} + 2\lambda\omega_2 = 0 \longrightarrow \omega_2 = \dfrac{1}{4\lambda} \end{array}\right\} \ \omega^* = \begin{bmatrix} 0 \\ \frac{1}{4\lambda} \end{bmatrix}$$

$\Rightarrow$ 3 points for writing eq. of $J(\omega)$

$\Rightarrow$ 2 points for getting $\omega_1$ & $\omega_2$

total/10