1. Recall that in the analysis of perceptron's mistake bound, we had used $\|x^t\|_2 \leq R, \forall t = 1, 2, 3, \ldots$ and note that for $x = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$, we have $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$, $\|x\|_1 = \sum_{i=1}^d |x_i|$ and $\|x\|_\infty = \max_{i \in \{1,2,\ldots,d\}} |x_i|$ (recall that for a real number z, $|z|$ denotes its absolute value). Answer the following.

(a) [**4 marks**] Find suitable values $\beta > 0, \mu > 0$ such that $\|x\|_2 \leq \beta\|x\|_1$ and $\|x\|_2 \leq \mu\|x\|_\infty$. Justify if your bounds are tight. (**Hint:** To show that your bound is tight, you can check if $\|x\|_2 \leq \beta\|x\|_1$ for every $x \in \mathbb{R}^d$ and if there exists a particular choice for x such that $\|x\|_2 = \beta\|x\|_1$ holds.)

(b) [**4 marks**] Consider the training data $\{(x^t, y^t)\}$, $t = 1, 2, \ldots$, where $x^t \in \mathbb{R}^d$, $\forall t = 1, 2, \ldots$, and $d = 3m$ where m is a positive integer. Denote the first $d/3$ components of x^t as a^t and denote the next $d/3$ components as b^t and the last $d/3$ components of x^t as q^t. Suppose $\max_{t=1,2,\ldots} \|a^t\|_1 \leq 1$ and $\max_{t=1,2,\ldots} \|b^t\|_\infty \leq 1$ and $\max_{t=1,2,\ldots} \|q^t\|_2 \leq 1$. Using the answer you obtained for part (a), find a suitable choice for R used in perceptron mistake bound analysis.

Question 1

a) $$\|x\|_1^2 = \left(\sum_{i=1}^{d} |x_i|\right)^2$$

$$= \sum_{j=1}^{d} \sum_{i=1}^{d} |x_i| \cdot |x_j|$$

$$\geq \sum_{i=1}^{d} |x_i| \cdot |x_i| \qquad (\because |x_i| \cdot |x_j| \geq 0 \ \forall i, j)$$

$$= \|x\|_2^2$$

$$\Rightarrow \|x\|_1^2 \geq \|x\|_2^2$$

$$\Rightarrow \|x\|_2 \leq 1 \cdot \|x\|_1 \qquad \beta > 0, \quad \beta = 1$$

To show that bound is tight

take $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\|x\|_2 = \sqrt{1+0} = 1$$

$$\|x\|_1 = 1 + 0 = 1$$

$$\|x\|_2 \leq \mu \|x\|_\infty$$

let $m = \max_i |x_i| = \|x\|_\infty$

$$\|x\|_2^2 = \sum_{i=1}^{d} (x_i)^2$$

$$\leq \sum_{i=1}^{d} m^2$$

$$= d \cdot m^2$$

$$\Rightarrow \|x\|_2^2 \leq d \cdot m^2$$

$$\Rightarrow \|x\|_2 \leq \sqrt{d}\, \|x\|_\infty \quad \text{so, } \mu = \sqrt{d}$$

To prove that bound is tight

consider $x = (a, a)$

then $\|x\|_2 = \sqrt{a^2 + a^2} = \sqrt{2}|a|$

$$\|x\|_\infty = |a|$$

so, $\|x\|_2 = \sqrt{2} \cdot \|x\|_\infty$

b) R is a bound on 2-norm of x^t

i.e., $\|x^t\| \leq R \quad \forall t$

Now $x^t = [a^t \; b^t \; q^t]$

$$\|a^t\|_2^2 = a_1^2 + a_2^2 + \cdots + a_m^2$$
$$\|b^t\|_2^2 = b_1^2 + b_2^2 + \cdots + b_m^2$$
$$\|q^t\|_2^2 = q_1^2 + q_2^2 + \cdots + q_m^2$$

$$\|x^t\|_2^2 = a_1^2 + \cdots + a_m^2 + b_1^2 + \cdots + b_m^2 + q_1^2 + \cdots + q_m^2$$
$$= \|a^t\|_2^2 + \|b^t\|_2^2 + \|q^t\|_2^2 \quad \text{—— (1)}$$

from part a

we know $\|y\|_2 \leq \|y\|_1$

so $\|a^t\|_2^2 \leq \|a^t\|_1^2 \leq 1$

Also, $\|y\|_2 \leq \sqrt{d}\, \|y\|_\infty$

so, $\|b^t\|_2^2 \leq d \cdot 1$

and $\|q^t\|_2^2 \leq 1$

Put this in (1)

$$\|x^t\|_2^2 = \|a^t\|_2^2 + \|b^t\|_2^2 + \|q^t\|_2^2$$

$$\leq 1 + d + 1$$

$$\Rightarrow \|x^t\|_2 \leq \sqrt{2+d}$$

So, a suitable choice for R is $\sqrt{2+d}$

2. Consider a finite 2 dimensional data set D with $2m$ samples where the i-th sample is of the form $(i, i+1)$, $\forall i \in \{1, 2, \ldots, 2m\}$. Suppose the first m samples are labeled $+1$ and the remaining m samples are labeled -1. Answer the following:

   (a) **[3 marks]** Construct a suitable hyperplane (w, b) which makes D linearly separable. Justify why your choice of hyperplane is valid.

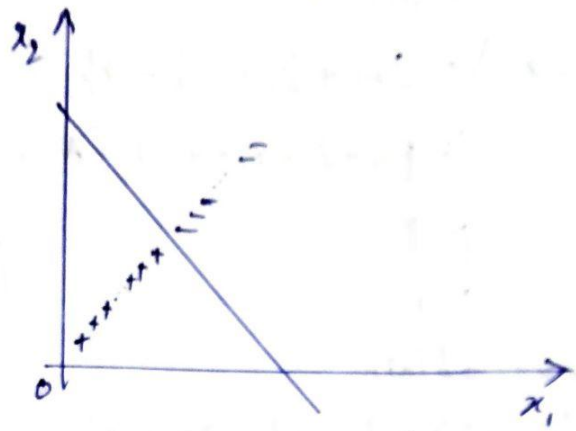
   (b) **[4 marks]** Transform the data set D into a new 2 dimensional data set $\widetilde{D}$ which is linearly separable by a hyperplane of the form $(\widetilde{w}, 0)$. Describe the procedure using which you transformed the data set and hence discuss if there is any relation between w and $\widetilde{w}$.

   (c) **[5 marks]** Would your transformation idea in part (b) help in transforming any d dimensional data set which has a separating hyperplane of the form (w, b) with $b \neq 0$ into another d dimensional data set for which the separating hyperplane is $(\widetilde{w}, 0)$? Justify with proper reasons. (Assume appropriate dimensions for w, $\widetilde{w}$ and b.)

Q2(a) Data set D contains $2m$ samples

$(1,2), (2,3), \dots, (m, m+1) \rightarrow$ Label $+1$

$(m+1, m+2), (m+2, m+3), \dots, (2m, 2m+1) \rightarrow$ Label -1



From the graph it is clear that data set D is linearly separable. A suitable hyperplane (w, b) which separates D can be found by a line which bisects $(m, m+1)$ and $(m+1, m+2)$ and is perpendicular to the line joining both the points.

$$\frac{x_2 - \frac{m+1+m+2}{2}}{x_1 - \frac{m+m+1}{2}} = -\frac{1}{\frac{m+2-(m+1)}{m+1-m}}$$

$$\Rightarrow x_1 + x_2 = 2m+2$$

Now as we want $\langle w, x \rangle \geq b$ for labels $+1$ and $\langle w, x \rangle < b$ for labels -1, we can choose

$$w = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \text{ and } b = -(2m+2)$$

For γ, we can find half of the distance between $(m, m+1)$ and $(m+1, m+2)$:

$$\gamma = \sqrt{(m+1-m)^2 + (m+2-m-1)^2} / 2$$

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

In order to verify that hyperplane (w, b) separates D, we consider $\bar{w} = \begin{pmatrix} 2m+2 \\ -1 \\ -1 \end{pmatrix}$ and $\bar{x}^i = \begin{pmatrix} 1 \\ x_1^i \\ x_2^i \end{pmatrix}$

For $y^i = +1$, we obtain

$$\langle \bar{w}, \bar{x}^i \rangle = 2m + 2 - x_1^i - x_2^i$$

$$\geq 2m+2-m-m-1 \quad \left[\text{As max of } x_1^i + x_2^i = 2m+1 \text{ for } y^i = +1\right]$$

$$\geq 1$$

$$> \gamma$$

For $y^i = -1$, we obtain

$$\langle \bar{w}, \bar{x}^i \rangle = 2m+2 - x_1^i - x_2^i$$

$$\leq 2m+2-m-1-m-2 \quad \left[\text{As min of } x_1^i + x_2^i = 2m+3 \text{ for } y^i = -1\right]$$

$$\leq -1$$

$$< \gamma$$

$$\therefore \underline{y^i \langle \bar{w}, \bar{x}^i \rangle > \gamma}$$

(b) In order to transform the dataset D to $\tilde{D}$ s.t. the separating hyperplane becomes $(\tilde{w}, 0)$. This can be obtained by shifting the origin s.t. it lies on hyperplane (w, b). This shifting will still result in $\tilde{D}$ which is linearly separable. Hence,

for $y^i = +1$,

$$\langle w, x^i \rangle \geq b$$

$$\Rightarrow \langle w, x^i \rangle - b \geq 0$$

$$\Rightarrow \langle w, x^i \rangle - \frac{\langle w, w \rangle b}{\|w\|^2} \geq 0$$

$$\Rightarrow \left\langle w, x^i - \frac{b\,w}{\|w\|^2} \right\rangle \geq 0$$

Similarly, for $y^i = -1$,

$$\langle w, x^i \rangle < b$$

$$\Rightarrow \langle w, x^i \rangle - b < 0$$

$$\Rightarrow \left\langle w, x^i - \frac{bw}{\|w\|^2} \right\rangle < 0$$

$$\therefore \tilde{x}^i = x^i - \frac{bw}{\|w\|^2} \qquad \text{and } \tilde{w} = w = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$$= x^i + \frac{2m+2}{2} \begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$$= x^i - (m+1) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

(c) The same transformation used in (b) will work for any d-dimensional dataset $\{(x^i, y^i)\}_{i=1}^{m}$ which has a separating hyperplane (w, b), where $b \neq 0$.

Let, the dataset satisfy $y^i \langle \tilde{w}, \tilde{x}^i \rangle > \gamma$ (say).

Consider $\tilde{x}^i = x^i - \frac{bw}{\|w\|^2}$

$$\therefore y^i \left\langle \begin{pmatrix} -b \\ w \end{pmatrix}, \begin{pmatrix} 1 \\ x^i \end{pmatrix} \right\rangle > \gamma$$

$$\Rightarrow y^i \left(-b + \langle w, x^i \rangle \right) > \gamma$$

$$\Rightarrow y^i \left(-b + \langle w, \tilde{x}^i \rangle \right) > \gamma$$

$$\Rightarrow y^i \left(-b + \left\langle w, \tilde{x}^i + \frac{bw}{\|w\|^2} \right\rangle \right) > \gamma$$

$$\Rightarrow y^i \left(-\cancel{b} + \cancel{b} + \langle w, \tilde{x}^i \rangle \right) > \gamma$$

$$\Rightarrow y^i \left\langle \begin{pmatrix} 0 \\ w \end{pmatrix}, \begin{pmatrix} 1 \\ \tilde{x}^i \end{pmatrix} \right\rangle > \gamma$$

We note γ remains the same.

3. Assume a fully connected feed forward neural network where the output layer has a single neuron and every neuron in the hidden layers and output layer uses logistic sigmoid activations. Your classmate suggests that to avoid vanishing gradient issue in this feed forward network, the following idea can be tried: in addition to the usual connections already available between layers ℓ and $\ell + 1$, the activations output by layer ℓ can also be connected to the output layer neuron directly thus introducing additional weights in the network. Answer the following.

(a) **[3 marks]** Describe pictorially the architecture of the neural network suggested by your classmate with appropriate weight notations.

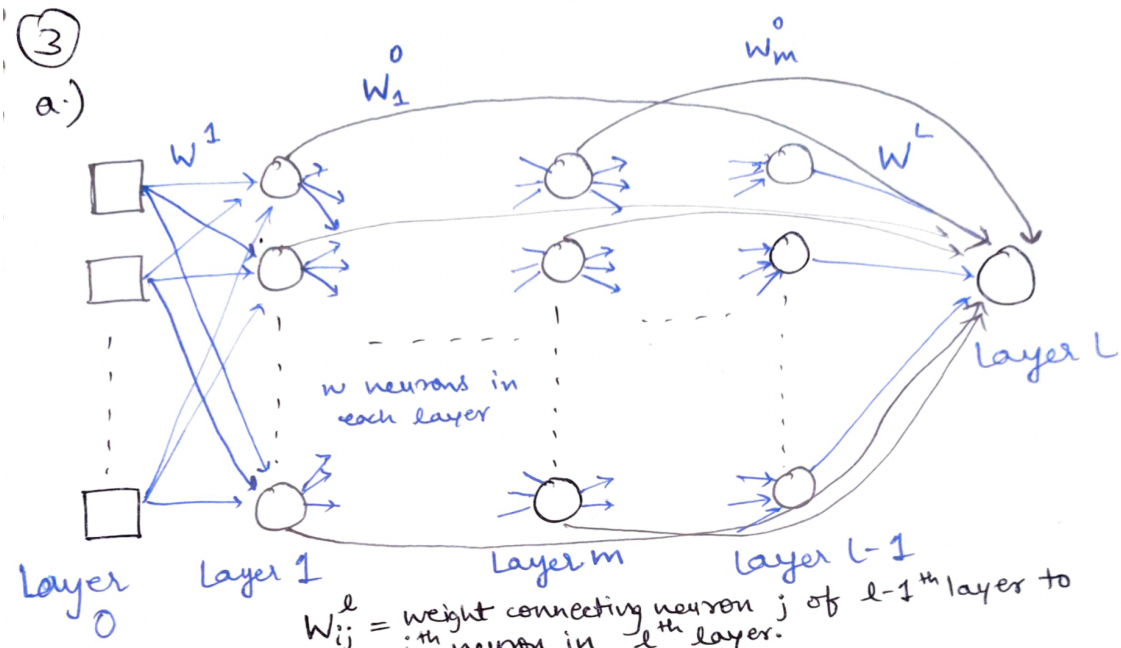
(b) **[3 marks]** Describe a suitable loss function which can be used for the network suggested.

(c) **[6 marks]** Using appropriate derivations of the error gradients for the network architecture suggested by your classmate, check and justify if vanishing gradient problem can indeed be solved or not.

(d) **[3 marks]** Comment on the issue of exploding gradients for the proposed network.

3

a.)



W_{ij}^{l} = weight connecting neuron j of $l-1^{th}$ layer to i^{th} neuron in l^{th} layer.

(b.) The suitable loss function that can be used for the suggested network can be Mean Squared Error loss.

(c.) error gradients at last layer:-

$$\nabla w_{11}^{L} = \frac{\partial e}{\partial z_1^L} \cdot \frac{\partial z_1^L}{\partial w_{11}^L} = \frac{\partial e}{\partial a_1^L} \cdot \frac{\partial a_1^L}{\partial z_1^L} a^{L-1} = \frac{\partial e}{\partial \hat{y}} \phi'(z_1^L)(a_1^{L-1})^T$$

Similarly

$$\nabla w_{12}^{L} = \frac{\partial e}{\partial \hat{y}} \phi'(z)(a_1^{L-1})^T$$

So $\nabla_{w^L} e = \frac{\partial e}{\partial \hat{y}} \phi'(z)(a^{L-1})^T$

error gradients for hidden layers:-

At layer l, we know error gradients are:-

$$\nabla w_l e = \text{Diag}\cdot(\phi'^{l})\, \delta^{l} (a^{l-1})^T$$

but for this network structure, an additional term of error gradient with respect to weight connecting directly to output neuron for every neuron in the hidden layers will come.

So, for layer l:-

$$\nabla w_l e = \text{Diag}\cdot(\phi'^{l})\, \delta^{l} (a^{l-1})^T + \frac{\partial e}{\partial \hat{y}} \phi'(z)^{L} (a^{l-1})$$

where:-

$$\text{Diag}(\phi'^{l}) = \begin{bmatrix} \phi'(z_1)^{l} & & \\ & \ddots & \\ & & \phi'(z_{n_l}^{l}) \end{bmatrix}, \quad \delta = \begin{bmatrix} \frac{\partial e}{\partial a_1^{l}} \\ \vdots \\ \frac{\partial e}{\partial a_{n_l}^{l}} \end{bmatrix}$$

and $a^{l-1} = \begin{bmatrix} a_1^{l-1} \\ \vdots \\ a_{n_{l-1}}^{l-1} \end{bmatrix}$

So,

$$\nabla_{W_l} e = \mathrm{Diag}(\phi'^{l}) V^{l+1} V^{l+2} \cdots V^{L} \delta^{L} (a^{l-1})^T + \frac{\partial e}{\partial \hat{y}} \phi'(z)^{L} (a^{l-1})^T$$

where $V^{l+1} = (W^{l+1})^T \mathrm{Diag}(\phi'^{l+1})$

Comparing the above expression with the expression that we get in a standard feed forward neural network, we can see that for every hidden layer l, where $0 < l \leq L-1$, an additional term of error gradient with respect to weight for output neuron is there.

From here, we can deduce that, when error gradients for standard hidden layers will be very less (close to 0), than the additional term i.e. $\frac{\partial e}{\partial \hat{y}} \phi'(z)^{L} (a^{l-1})^T$ for every neuron will balance the effects of neglible gradients. i.e. the additional term can handle "vanishing gradient".

d.) For the proposed network, we only care about the problem of vanishing gradient, and no other thing to handle exploding gradient is done, so, the problem of exploding gradient may arise which needs to be taken care of. If the additional term grows much, than this problem can arise, which can be tackled by reducing the direct connections of every hidden layer neurons with the output neuron.

4. Recall that a hyperplane (w, b) can be considered as a separating hyperplane for a data set $D = \{(x^t, y^t)\}$, $t = 1, 2, 3, \ldots$ if $y^t = \text{sign}(\langle w, x^t \rangle + b)$ for every $t = 1, 2, 3, \ldots$ (Assume an appropriate definition of sign function).

   (a) **[3 marks]** If linear separability assumption holds for D, does it imply the existence of a separating hyperplane in the sense of the discussion above? Discuss with proper justification.

   (b) **[4 marks]** If the existence of a separating hyperplane for D holds based on the discussion above, would it imply that the linear separability assumption holds for D? Discuss with proper justification.

Page No.
Date

Sol^n 4 $\Rightarrow$ Hyperplane (w, b) be considered as a separating hyperplane for a data set $D = \{(x^t, y^t)\}$, $t = 1, 2, 3, \ldots$ if $y^t = \text{sign}(\langle w, x^t \rangle + b)$ for every $t = 1, 2, 3, \ldots$

(a) Given that linear separability assumption hold for D.

i.e. there exist $w^* \in \mathbb{R}^d$, $\neq 0$ and $\gamma > 0$

such that

Ineq ① $\rightarrow$ $\langle w^*, x^t \rangle > \gamma$ where $y^t = 1$

Ineq ② $\rightarrow$ $\langle w^*, x^t \rangle < -\gamma$ where $y^t = -1$

by inequality (1) $\rightarrow$

$$\langle w^*, x^t \rangle - \gamma > 0, \text{ where } y^t = 1$$

$$\Rightarrow \text{sign}(\langle w^*, x^t \rangle - \gamma) = y^t \quad - ③$$

and by Inequality (2) $\Rightarrow$

$$\langle w^*, x^t \rangle < -\gamma \quad \text{where } y^t = -1$$

since $\gamma > 0$

$$\Rightarrow \langle w^*, x^t \rangle < -\gamma < \gamma$$

$\Rightarrow \langle w^*, x^b \rangle < \gamma$, where $y^b = -1$

$\Rightarrow \langle w^*, x^b \rangle - \gamma < 0$, where $y^b = -1$

$\Rightarrow \text{sign}(\langle w^*, x^b \rangle - \gamma) = y^b$ — (4)

by both equation (3) and (4) we can claim

~~that~~ there exist an Hyperplane $(w^*, -\gamma)$ so that the data set D is separable.

$\Rightarrow$ If data set holds linear separability assumption ~~then it is also there always exist an~~ then it imply the existance of a separating hyperplane.

Note $\Rightarrow$ sign function $\rightarrow$

$$\text{sign}(x) = f(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases}$$

(b) Given that separating hyperplane exist for Data D.

i.e there exist (w, b) such that.

$$\text{sign}(\langle w, x \rangle + b) = y^t$$

by definition of sign function $\Rightarrow$

$\langle w, x \rangle + b \geq 0$, where $y^t = 1$

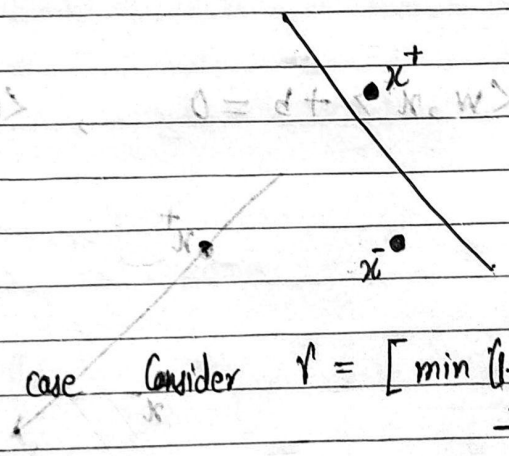
and $\langle w, x \rangle + b < 0$, where $y^t = -1$

Now Consider x^+ to be the closest point with label $+1 =: y^+$ to the separating hyperplane and let x^- be the closest point with label $-1 =: y^-$ to the separating hyperplane.

Now Consider the following Cases :-

<u>Case 1</u> $\rightarrow$ $\langle w, x^+ \rangle + b > 0$ and $\langle w, x^- \rangle + b < 0$

(that is x^+ and x^- are away from the separating hyperplane)



In this case Consider $\gamma = \left[\min \left(|\langle w, x^+ \rangle + b|, |\langle w, x^- \rangle + b| \right) \right] - \varepsilon$

(for some $\varepsilon > 0$)

small quantity almost close to 0.

Now Note that $y^+(\langle w, x^+\rangle + b)$

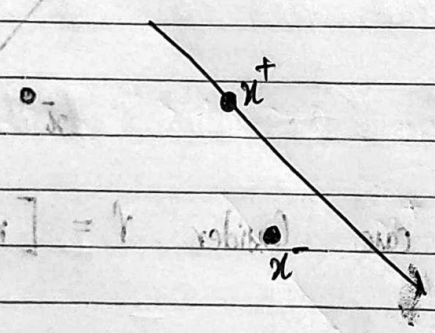
$$= |\langle w, x^+\rangle + b| > [\min(|\langle w, x^+\rangle + b|, |\langle w, x^-\rangle + b|) - \varepsilon]$$

$$\Rightarrow |\langle w, x^+\rangle + b| > \gamma$$

Similarly $y^-(\langle w, x^-\rangle + b) = |\langle w, x^-\rangle + b| > \gamma$

hence linear separability assumption hold in this case

Case 2 $\Rightarrow$ $\langle w, x^+\rangle + b = 0$, $\langle w, x^-\rangle + b < 0$



now consider $\alpha = \langle w, x^-\rangle + b$, since $\langle w, x^-\rangle + b < 0$

we have $\alpha < 0$ (ie α is negative)

we also have $y^-(\langle w, x^-\rangle + b) > 0$.

or $y^-\alpha > 0$, $(y^-\alpha = |\alpha|)$

Page No.
Date

now consider $y^-\left(\alpha + \frac{|\alpha|}{2}\right) = y^-\alpha + y^-\frac{|\alpha|}{2} = y^-\alpha - \frac{|\alpha|}{2}$

$= |\alpha| - \frac{|\alpha|}{2} = \frac{|\alpha|}{2}$

$> \frac{|\alpha|}{2} - \varepsilon$ —(1)

(for some $\varepsilon > 0$)

small quantity close to 0.

now consider $\langle w, x^+ \rangle + b = 0$

$\Rightarrow \langle w, x^+ \rangle + b + \frac{|\alpha|}{2} = \frac{|\alpha|}{2} > \frac{|\alpha|}{2} + \varepsilon$ (for some $\varepsilon > 0$)

—(2)

Thus with $\tilde{b} = b + \frac{|\alpha|}{2}$ and $\gamma = \frac{|\alpha|}{2} - \varepsilon$

we have from (1) and (2)

$y^+ (\langle w, x^+ \rangle + \tilde{b}) > \gamma$

$y^- (\langle w, x^- \rangle + \tilde{b}) > \gamma$

hence linear separability assumption hold in this case.

Case 3 $\Rightarrow$ $\langle w, x^+ \rangle + b = 0$, $\langle w, x^- \rangle + b = 0$

This case can not happen, cause sign function definition not hold.

As in this starting we have make an assumption that x^+ to be the closest point to the ~~by~~ separating hyperplane.

Now let consider x^- be the closest point to the separating hyperplane. In this manner we have again three cases -

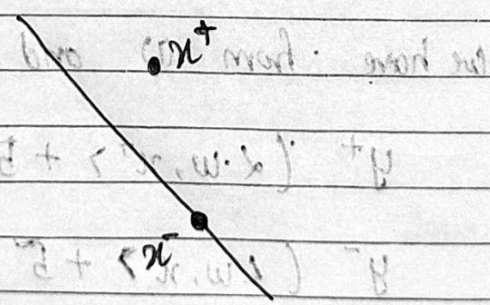
Case 1 $\Longrightarrow$ $\langle w, x^+ \rangle + b > 0$ and $\langle w, x^- \rangle + b < 0$

this case same as in the previous part.

So in this case linear separability holds.

Case 2 $\Longrightarrow$ $\langle w, x^+ \rangle + b > 0$ and $\langle w, x^- \rangle + b = 0$

i.e $\Rightarrow$



now consider $\beta = \langle w, x^+ \rangle + b$

since $\langle w, x^+ \rangle + b > 0 \Rightarrow \beta > 0$

now $y^+ \left(\langle w, x^+ \rangle + b - \frac{\beta}{2} \right) =$

$$= y^+\left(\beta - \frac{\beta}{2}\right) = y^+\beta - \frac{y^+\beta}{2} = \frac{\beta}{2} > \frac{\beta}{2} - \varepsilon \quad -(1)$$

(for some $\varepsilon > 0$)

small quantity close to 0

now $\langle w, x^- \rangle + b = 0$

$$\langle w, x^- \rangle + b = \frac{\beta}{2} = -\frac{\beta}{2} < -\left(\frac{\beta}{2} - \varepsilon\right) \quad -(2)$$

(for some $\varepsilon > 0$)

now by equation (1) and (2)

and by letting $\tilde{b} = b - \frac{\beta}{2}$ and $\gamma = \frac{\beta}{2} - \varepsilon$

$$y^+\left(\langle w, x^+ \rangle + \tilde{b}\right) > \gamma$$

and $$y^-\left(\langle w, x^- \rangle + \tilde{b}\right) > \gamma$$

hence the linear separability assumption hold.

<u>Case 3</u> $\Longrightarrow$ $\langle w, x^+ \rangle + b = 0$, $\langle w, x^- \rangle + b = 0$

$\Rightarrow$ Again this can not possible, by the definition of sign function.

Hence by the all conditions and cases we conclude that in each case if for D; any separating hyperplane exist, then D always hold linear separability condition..

The main ~~notice~~ attractive part is if (w, b) is a separating hyperplane of D, then it is not necessary that linear separability hold with (w, b). But ~~we it~~ ~~it guanteed~~ it is guaranteed that there might exist some other corresponding hyperplane $(w, \tilde{b})$ that satisfy linear separability.