IE643: Deep Learning - Theory and Practice

Quiz 1 Solution

July-Dec 2024

1. Consider the following data set: $D = \{(x^1, y^1), (x^2, y^2), (x^3, y^3), (x^4, y^4)\}$, where $(x^1, y^1) = ((0, 1), -1)$, $(x^2, y^2) = ((1, 0), -1)$, $(x^3, y^3) = ((0, 3), +1)$ and $(x^4, y^4) = ((3, 0), +1)$.

   (a) **[3 marks]** Find a linear separator for dataset D. Clearly write the expression for the linear separator and justify why your construction is in fact a linear separator for D.

let, take a linear seperator $x_1 w_1 + x_2 w_2 - b = 0$

x_1	x_2	$\hat{y}$	y	
0	1	$\text{sign}(w_2 - b)$	-1	$\rightarrow w_2 - b < 0$ — (1)
1	0	$\text{sign}(w_1 - b)$	-1	$\rightarrow w_1 - b < 0$ — (2)
0	3	$\text{sign}(3w_2 - b)$	+1	$\rightarrow 3w_2 - b > 0$ — (3)
3	0	$\text{sign}(3w_1 - b)$	+1	$\rightarrow 3w_1 - b > 0$ — (4)

from eq (1) and (2)

$w_1 + w_2 - 2b < 0$

$\Rightarrow \frac{1}{2}(w_1 + w_2) < b$ — (6)

from eq (3) and (4)

$3w_1 + 3w_2 - 2b > 0$

$\Rightarrow \frac{3}{2}(w_1 + w_2) > b$ — (5)

from eq (5) and (6)

$$\boxed{\frac{1}{2}(w_1 + w_2) < b < \frac{3}{2}(w_1 + w_2)}$$

from eq (1) & (3)

$3w_2 - b > 0$

$-w_2 + b > 0$

$2w_2 > 0$

$\boxed{w_2 > 0}$

from (2) and (4)

$3w_1 - b > 0$

$-w_1 + b > 0$

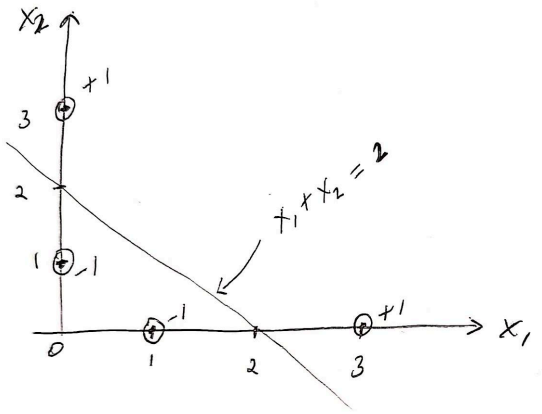
$2w_1 > 0$

$\boxed{w_1 > 0}$

let take $w_1 = w_2 = 1$, then $\boxed{1 < b < 3}$

$\therefore$ Linear seperator will be, taking $b = 2$

$x_1 + x_2 - 2 = 0$



1. Consider the following data set: $D = \{(x^1, y^1), (x^2, y^2), (x^3, y^3), (x^4, y^4)\}$, where $(x^1, y^1) = ((0,1), -1)$, $(x^2, y^2) = ((1,0), -1)$, $(x^3, y^3) = ((0,3), +1)$ and $(x^4, y^4) = ((3,0), +1)$.

   (a) **[3 marks]** Find a linear separator for dataset D. Clearly write the expression for the linear separator and justify why your construction is in fact a linear separator for D.

   (b) **[6 marks]** Suppose the linear separator you found in part 1a is represented as (w, b) such that for any $x \in \mathbb{R}^2$ on the linear separator, $\langle w, x\rangle - b = 0$ holds. Transform D into a new dataset $D' = \{(q^1, y^1), (q^2, y^2), (q^3, y^3), (q^4, y^4)\}$ where the features of i-th sample x^i in D are transformed into q^i so that the feature $q_1^i = \alpha_1 x_1^i + \beta_1$ and the feature $q_2^i = \alpha_2 x_2^i + \beta_2$ for suitable real numbers $\alpha_1, \beta_1, \alpha_2, \beta_2$, such that the linear separator of the new dataset D' is of the form $(v, 0)$ (that is for any point $x \in \mathbb{R}^2$ on the linear separator, $\langle v, x\rangle = 0$). Give details of the resultant new dataset D' and the corresponding new linear separator $(v, 0)$. How do v and w relate to each other?

In this part we have to transform the point such that the linear seperator will pass through origin.

for some $x \in R^2$, and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ the transformation is applied such that

$$\left.\begin{aligned} q_1 &= \alpha_1 x_1 + \beta_1 \\ q_2 &= \alpha_2 x_2 + \beta_2 \end{aligned}\right\} \text{—— (1)}$$

let $q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}$

Now, as per question we have to make sure that the linear seperator for transformed dataset is of the form $\langle v, q \rangle = 0$

$$\Rightarrow v_1 q_1 + v_2 q_2 = 0 \quad \text{—— (2)}$$

we know from part (a) that seperator is $x_1 + x_2 - 2 = 0$

i.e. of the form $w_1 x_1 + w_2 x_2 - b = 0$

replacing the values of x_1 and x_2 based upon eq ①.

$$\Rightarrow w_1 \left(\frac{q_1 - \beta_1}{\alpha_1} \right) + w_2 \left(\frac{q_2 - \beta_2}{\alpha_2} \right) = b$$

Simplifying,

$$\frac{w_1}{\alpha_1} q_1 + \frac{w_2}{\alpha_2} q_2 - \left(\frac{w_1 \beta_1}{\alpha_1} + \frac{w_2 \beta_2}{\alpha_2} \right) = b \quad \text{— (3)}$$

In new seperator we don't need bias term

So, $\frac{w_1 \beta_1}{\alpha_1} + \frac{w_2 \beta_2}{\alpha_2} + b = 0$

[take values of w_1, w_2 and b from part (a) i.e. $w_1 = w_2 = 1$ and $b = 2$]

$$\frac{\beta_1}{\alpha_1} + \frac{\beta_2}{\alpha_2} = -2 \quad \text{— (4)}$$

eq (3) becomes

$$\frac{w_1}{\alpha_1} q_1 + \frac{w_2}{\alpha_2} q_2 = 0 \longrightarrow \frac{1}{\alpha_1} q_1 + \frac{1}{\alpha_2} q_2 = 0$$

comparing with $v_1 q_1 + v_2 q_2 = 0$

$$v_1 = \frac{w_1}{\alpha_1} \quad \text{and} \quad -v_2 = \frac{w_2}{\alpha_2}$$

Now for new dataset D'

Let $\alpha_1 = -\beta_1$ and $\alpha_2 = -\beta_2$ satisfying eq (4)

applying transformation to each point in D

X_1	q_1	X_2	q_2	Y
0	$0 + \beta_1 = \beta_1$	1	$-\beta_2 + \beta_2 = 0$	-1
1	$-\beta_1 + \beta_1 = 0$	0	$0 + \beta_2 = \beta_2$	-1
0	$0 + \beta_1 = \beta_1$	3	$-3\beta_2 + \beta_2 = -2\beta_2$	1
3	$-3\beta_1 + \beta_1 = -2\beta_1$	0	$0 + \beta_2 = \beta_2$	1

Verifying linear seperator

$$V_1 q_1 + V_2 q_2 = 0$$

$$\frac{1}{\alpha_1} q_1 + \frac{1}{\alpha_2} q_2 = 0$$

$$\text{Sign}\left(\frac{\beta_1}{-\beta_1} + \frac{0}{-\beta}\right) = -1 = \hat{y}'$$

$$\text{Sign}\left(0 + \frac{\beta_2}{-\beta_2}\right) = -1$$

$$\text{Sign}\left(\frac{\beta_1}{-\beta_1} + \frac{(-2\beta_2)}{-\beta_2}\right) = 1$$

$$\text{Sign}\left(\frac{(-2\beta_1)}{-\beta_1} + \frac{\beta_2}{-\beta_2}\right) = 1$$

Let take $\alpha_1 = \alpha_2 = 1$ and $\beta_1 = \beta_2 = -1$

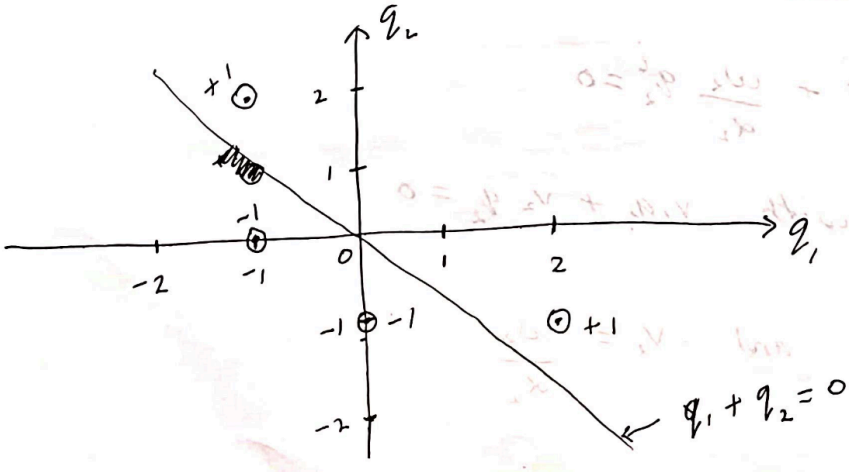
applying transformation to each point in D

X_1	q_1	X_2	q_2	Y
0	$0 - 1 = -1$	1	$1 - 1 = 0$	-1
1	$1 - 1 = 0$	0	$0 - 1 = -1$	-1
0	$0 - 1 = -1$	3	$3 - 1 = 2$	1
3	$3 - 1 = 2$	0	$0 - 1 = -1$	1

$\therefore D' = \{((-1, 0), -1), ((0, -1), -1), ((-1, 2), 1), ((2, -1), 1)\}$

$V_1 = 1$ and $V_2 = 1$

$\therefore$ linear seperator $\rightarrow$ $q_1 + q_2 = 0$



1. Consider the following data set: $D = \{(x^1, y^1), (x^2, y^2), (x^3, y^3), (x^4, y^4)\}$, where $(x^1, y^1) = ((0,1), -1)$, $(x^2, y^2) = ((1,0), -1)$, $(x^3, y^3) = ((0,3), +1)$ and $(x^4, y^4) = ((3,0), +1)$.

   (a) **[3 marks]** Find a linear separator for dataset D. Clearly write the expression for the linear separator and justify why your construction is in fact a linear separator for D.

   (b) **[6 marks]** Suppose the linear separator you found in part 1a is represented as (w, b) such that for any $x \in \mathbb{R}^2$ on the linear separator, $\langle w, x\rangle - b = 0$ holds. Transform D into a new dataset $D' = \{(q^1, y^1), (q^2, y^2), (q^3, y^3), (q^4, y^4)\}$ where the features of i-th sample x^i in D are transformed into q^i so that the feature $q_1^i = \alpha_1 x_1^i + \beta_1$ and the feature $q_2^i = \alpha_2 x_2^i + \beta_2$ for suitable real numbers $\alpha_1, \beta_1, \alpha_2, \beta_2$, such that the linear separator of the new dataset D' is of the form $(v, 0)$ (that is for any point $x \in \mathbb{R}^2$ on the linear separator, $\langle v, x\rangle = 0$). Give details of the resultant new dataset D' and the corresponding new linear separator $(v, 0)$. How do v and w relate to each other?

   (c) **[4 marks]** Consider the dataset $A = \{(x^t, y^t)\}_{t=1,2,3,\ldots}$ where $x^t \in \mathbb{R}^2$, and $y^t \in \{+1, -1\}$. Recall that this dataset A is linearly separable if there exist a non-zero $w^* \in \mathbb{R}^2$ and $\gamma > 0$ such that $\langle w^*, x^t\rangle > \gamma$ when $y^t = 1$ and $\langle w^*, x^t\rangle < -\gamma$ when $y^t = -1$. Justify if this definition directly fits for the dataset D. Also justify if this definition fits directly for the dataset D'. Using your analysis in part 1b, explain why such a general definition of linear separability for A is sufficient.

$\langle w^*, x^t \rangle > \gamma$ when $y^t = 1$

$\langle w^*, x^t \rangle < -\gamma$ when $y^t = -1$ } — (1)

where $w^* \in R^2$, $x^t \in R^2$ and $\gamma > 0$

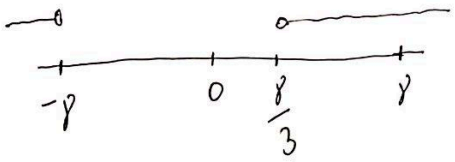
Checking condition (1) for dataset D

X_1	X_2	$\langle w^*, x^t \rangle$	y	
0	1	w_2^*	−1	$\rightarrow w_2^* < -\gamma$ — (1)
1	0	w_1^*	−1	$\rightarrow w_1^* < -\gamma$ — (2)
0	3	$3w_2^*$	+1	$\rightarrow 3w_2^* > \gamma$ — (3)
3	0	$3w_1^*$	+1	$\rightarrow 3w_1^* > \gamma$ — (4)

from eq (1) and (3)

$+w_2^* < -\gamma$

$3w_2^* > \gamma \rightarrow w_2^* > \frac{\gamma}{3}$ } contradiction



Similarly from eq (2) and (4) it can be shown the inequalities contradicts.

Therefore condition (1) does not fits on dataset D it is due to the fact that, for seperating dataset D we need a bias term in the linear seperator.

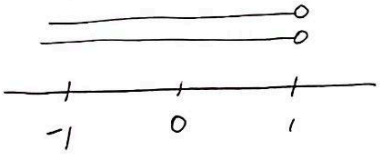
checking condition (1) for dataset D'

q_1	q_2	y	$\langle v^*, q^t \rangle$
β_1	0	-1	$v_1\beta_1 = \frac{\beta_1}{\alpha_1} = -1 \longrightarrow < -\gamma$ — (1)
0	β_2	-1	$v_2\beta_2 = \frac{\beta_2}{\alpha_2} = -1 \longrightarrow < -\gamma$ — (2)
β_1	$-2\beta_2$	+1	$v_1\beta_1 - 2v_2\beta_2 = \frac{\beta_1}{\alpha_1} - \frac{2\beta_2}{\alpha_2} = 1 \longrightarrow > \gamma$ — (3)
$-2\beta_1$	β_2	+1	$-2v_1\beta_1 + v_2\beta_2 = \frac{-2\beta_1}{\alpha_1} + \frac{\beta_2}{\alpha_2} = 1 \longrightarrow > \gamma$ — (4)

from eq (1) and (3)

$-1 < -\gamma \rightarrow 1 > \gamma$

$1 > \gamma$



from eq (2) and (4)

$-1 < -\gamma \rightarrow 1 > \gamma$

$1 > \gamma$

Same as previous

It can be seen that condition (1) fits directly to Dataset D', it is due to the fact that linear seperator $\langle v^*, q^t \rangle = 0$ passes from origin therefore not need of bias term.

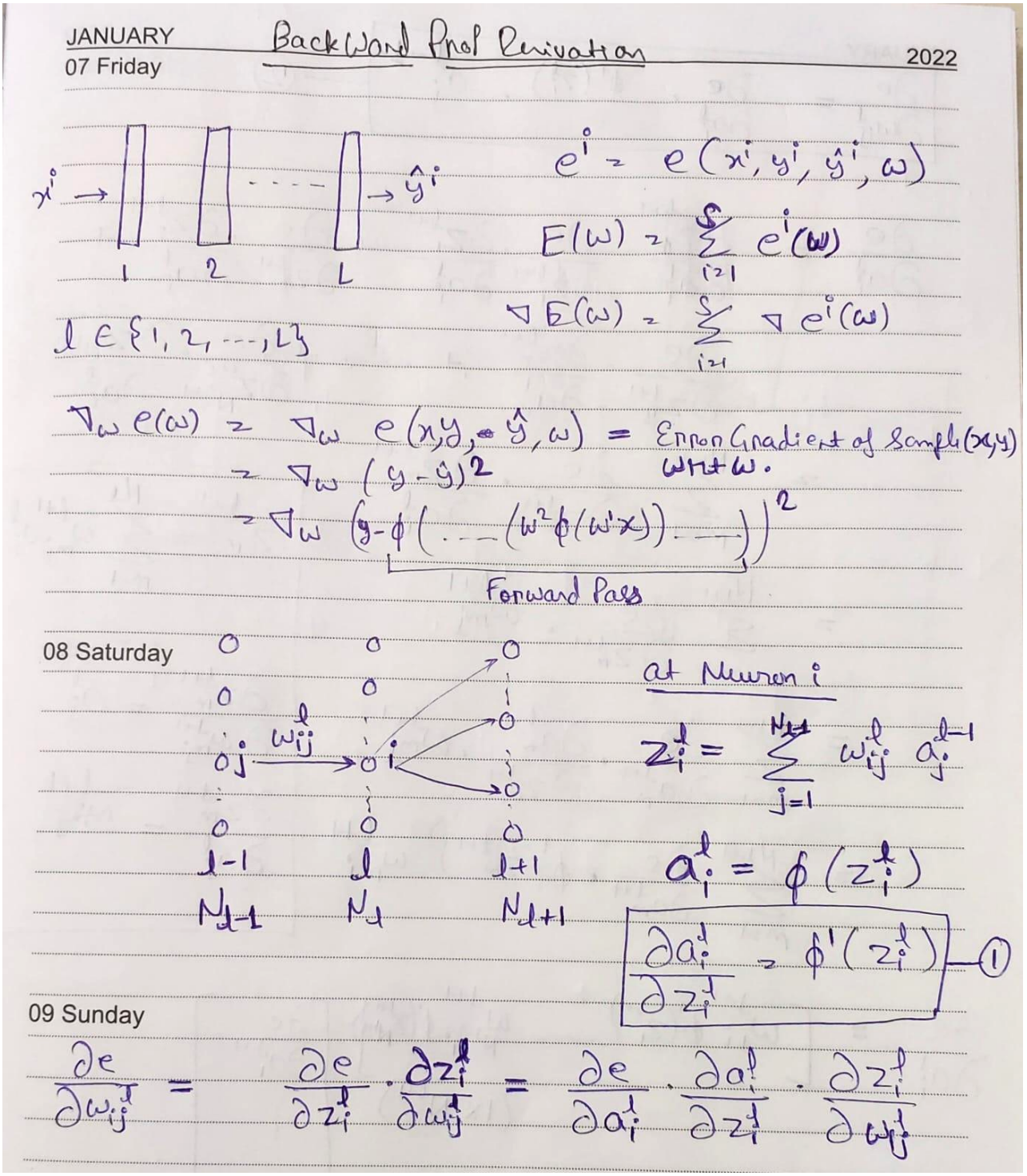
But dataset D' is formed by transforming the dataset D, maintaing the relative position of points,

So any dataset can be transform into the dataset like O', on which the linear seperability condition ① i.e. $y^t \langle w^*, x^t \rangle > \gamma$ can fit.

Therefor such general definition of linear seperability is sufficient.

2. [**5 marks**] Consider a multi-layer perceptron (MLP) with L layers and the layers are numbered as $\ell = 1, 2, \ldots, L$. Let W^ℓ denote the weights connecting to the ℓ-th layer of the MLP where $\ell \geq 2$. Let e denote the error with respect to a sample (x, y). Derive a suitable expression to compute the error gradients $\nabla_{W^\ell} e$ and show that it is of the form $\nabla_{W^\ell} e = \text{Diag}(\phi')^\ell \delta^\ell (a^{\ell-1})^\top$ where the j-th diagonal element of $\text{Diag}(\phi')^\ell$ is $\phi'(z_j^\ell)$, δ^ℓ denotes the error gradients with respect to the activations at ℓ-th layer and $a^{\ell-1}$ denotes the activations at layer $\ell - 1$. (**Note:** You can reuse the expressions for δ^ℓ, derived in class.)

2nd Answer:



JANUARY

10 Monday

$$\frac{\partial e}{\partial w_{ij}^{l}} = \frac{\partial e}{\partial a_i^{l}} \cdot \phi'(z_i^{l}) \cdot a_j^{l-1} \quad \text{---(2)}$$

$$\frac{\partial e}{\partial a_i^{l}} = \frac{\partial e}{\partial z_1^{l+1}} \cdot \frac{\partial z_1^{l+1}}{\partial a_i^{l}} + \frac{\partial e}{\partial z_2^{l+1}} \cdot \frac{\partial z_2^{l+1}}{\partial a_i^{l}} + \frac{\partial e}{\partial z_3^{l+1}} \cdot \frac{\partial z_3^{l+1}}{\partial a_i^{l}} + \cdots + \frac{\partial e}{\partial z_{N^{l+1}}^{l+1}} \cdot \frac{\partial z_{N^{l+1}}^{l+1}}{\partial a_i^{l}}$$

$$= \sum_{m=1}^{N_{l+1}} \frac{\partial e}{\partial z_m^{l+1}} \cdot \frac{\partial z_m^{l+1}}{\partial a_i^{l}}$$

$$= \sum_{m=1}^{N_{l+1}} \frac{\partial e}{\partial z_m^{l+1}} \cdot w_{mi}^{l+1}$$

$$= \sum_{m=1}^{N_{l+1}} \frac{\partial e}{\partial a_m^{l+1}} \cdot \frac{\partial a_m^{l+1}}{\partial z_m^{l+1}} \cdot w_{mi}^{l+1}$$

$$= \sum_{m=1}^{N_{l+1}} \frac{\partial e}{\partial a_m^{l+1}} \cdot \phi'(z_m^{l+1}) \cdot w_{mi}^{l+1}$$

$$z_m^{l+1} = \sum_{m=1}^{N_l} w_{mj}^{l+1} \cdot a_j^{l}$$

11 Tuesday

$$\frac{\partial z_m^{l+1}}{\partial w_{mj}^{l+1}} = a_j^{l}$$

$$\frac{\partial z_m^{l+1}}{\partial a_j^{l}} = w_{mj}^{l+1}$$

$$\frac{\partial e}{\partial a_i^{l}} = \begin{bmatrix} w_{1i}^{l+1} \phi'(z_1^{l+1}) & \cdots & w_{N^{l+1}1}^{l+1} \cdot \phi'(z_{N^{l+1}}^{l+1}) \end{bmatrix}_{(1 \times N^{l+1})} \begin{bmatrix} \frac{\partial e}{\partial a_1^{l+1}} \\ \vdots \\ \frac{\partial e}{\partial a_{N^{l+1}}^{l+1}} \end{bmatrix}_{(N^{l+1} \times 1)}$$

JANUARY 2022

M	T	W	T	F	S	S
31					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

JANUARY 2022

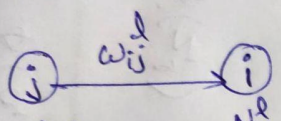
12 Wednesday

$$\begin{bmatrix} \frac{\partial e}{\partial a_1^l} \\ \vdots \\ \frac{\partial e}{\partial a_{N^l}^l} \end{bmatrix}_{(N^l \times 1)} = \begin{bmatrix} w_{11}^{l+1}\phi'(z_1^{l+1}) & \cdots & w_{N^{l+1}1}^{l+1}\phi'(z_{N^{l+1}}^{l+1}) \\ w_{12}^{l+1}\phi'(z_1^{l+1}) & \cdots & \\ \vdots & & \\ w_{1N^l}^{l+1}\phi'(z_1^{l+1}) & \cdots & w_{N^{l+1}N^l}^{l+1}\phi'(z_{N^{l+1}}^{l+1}) \end{bmatrix}_{(N^l \times N^{l+1})} \begin{bmatrix} \frac{\partial e}{\partial a_1^{l+1}} \\ \vdots \\ \frac{\partial e}{\partial a_{N^{l+1}}^{l+1}} \end{bmatrix}_{(N^{l+1} \times 1)}$$

$$= \begin{bmatrix} w_1^{l+1} & w_{21}^{l+1} & \cdots & w_{N^{l+1}1}^{l+1} \\ w_{12}^{l+1} & w_{22}^{l+1} & \cdots & w_{N^{l+1}2}^{l+1} \\ \vdots & & & \\ w_{1N^l}^{l+1} & w_{2N^l}^{l+1} & \cdots & w_{N^{l+1}N^l}^{l+1} \end{bmatrix}_{(N^l \times N^{l+1})} \begin{bmatrix} \phi'(z_1^{l+1}) & & & \\ & \phi(z_2^{l+1}) & & \\ & & \ddots & \\ & & & \phi(z_{N^{l+1}}^{l+1}) \end{bmatrix}_{(N^{l+1} \times N^{l+1})} \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{(N^{l+1} \times 1)}$$

13 Thursday

$$\boxed{\delta^l = (w^{l+1})^T \; \text{Diag}\left((\phi'(z))^{l+1}\right) \delta^{l+1}} \quad \text{— (3)}$$



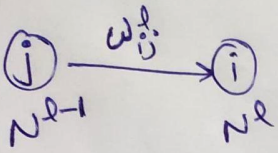
$\Rightarrow [w^l]$ have shape $(N^l \times N^{l-1})$

$\Rightarrow \left[\frac{\partial e}{\partial w^l}\right]$ have shape $(N^l \times N^{l-1})$

$i \in \{1, 2, \cdots, N^l\}$, N^l = Number of Neuron in l^{th} layer.

$j \in \{1, 2, \cdots, N^{l-1}\}$, N^{l-1} = Number of Neuron in $(l-1)^{th}$ layer.

$$\frac{\partial e}{\partial w_{ij}^l} = \frac{\partial e}{\partial a_i^l} \cdot \phi'(z_i^l) \cdot a_j^{l-1}$$



$\Rightarrow [w^l]$ have shape $(N^l \times N^{l-1})$

$\Rightarrow \left[\frac{\partial e}{\partial w^l}\right]$ have shape $(N^l \times N^{l-1})$

$i \in \{1, 2, \ldots, N^l\}$

$j \in \{1, 2, \ldots, N^{l-1}\}$

$$\left[\frac{\partial e}{\partial w_{1j}^l} \quad \frac{\partial e}{\partial w_{2j}^l} \quad \frac{\partial e}{\partial w_{3j}^l} \quad \cdots \quad \frac{\partial e}{\partial w_{N^l j}^l}\right] = \frac{\partial e}{\partial a_1^l} \quad \frac{\partial e}{\partial a_2^l} \cdots$$

$$= \left[\frac{\partial e}{\partial a_1^l} \phi'(z_1^l) \quad \frac{\partial e}{\partial a_2^l} \phi'(z_2^l) \cdots \frac{\partial e}{\partial a_{N^l}^l} \phi'(z_{N^l}^l)\right] a_j^{l-1}$$

$$\begin{bmatrix} \frac{\partial e}{\partial w_{11}^l} & \frac{\partial e}{\partial w_{21}^l} & \cdots & \frac{\partial e}{\partial w_{N^l 1}^l} \\ \frac{\partial e}{\partial w_{12}^l} & \frac{\partial e}{\partial w_{22}^l} & \cdots & \frac{\partial e}{\partial w_{N^l 2}^l} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial e}{\partial w_{1N^l}^l} & \frac{\partial e}{\partial w_{2N^l}^l} & \cdots & \frac{\partial e}{\partial w_{N^l N^{l-1}}^l} \end{bmatrix}_{(N^{l-1} \times N^l)} = \begin{bmatrix} a_1^{l-1} \\ a_2^{l-1} \\ \vdots \\ a_{N^{l-1}}^{l-1} \end{bmatrix}_{N^{l-1} \times 1} \begin{bmatrix} \frac{\partial e}{\partial a_1^l} \phi'(z_1^l) & \frac{\partial e}{\partial a_2^l} \phi'(z_2^l) & \cdots & \frac{\partial e}{\partial a_{N^l}^l} \phi'(z_{N^l}^l) \end{bmatrix}_{1 \times N^l}$$

$$\left(\nabla_{(W^l)^T} e\right) \quad \nabla_{W^l} e \quad = \begin{bmatrix} a_1^{l-1} \\ a_2^{l-1} \\ \vdots \\ a_{N^{l-1}}^{l-1} \end{bmatrix}_{N^{l-1} \times 1} \begin{bmatrix} \frac{\partial e}{\partial a_1^l} & \frac{\partial e}{\partial a_2^l} & \cdots & \frac{\partial e}{\partial a_{N^l}^l} \end{bmatrix}_{1 \times N^l} \begin{bmatrix} \phi'(z_1^l) & \cdots & 0 & \cdots & 0 \\ \vdots & & 0 & & \\ 0 & \cdots & 0 & \cdots & \phi'(z_{N^l}^l) \end{bmatrix}_{N^l \times N^l}$$

$$\nabla_{(W^l)^T} e = \quad = a^{l-1} \left(\delta^l\right)^T \text{Diag}(\phi'(z^l))$$

Take Transpose on Both Side,

$$\left(\nabla_{(W^l)^T} e\right)^T = \quad = \left(a^{l-1} \left(\delta^l\right)^T \text{Diag}(\phi'(z^l))\right)^T = \text{Diag}(\phi'(z^l)) \, \delta^l \left(a^{l-1}\right)^T$$

$$\boxed{\nabla_{W^l} e = \text{Diag}(\phi'(z^l)) \, \delta^l \left(a^{l-1}\right)^T}$$

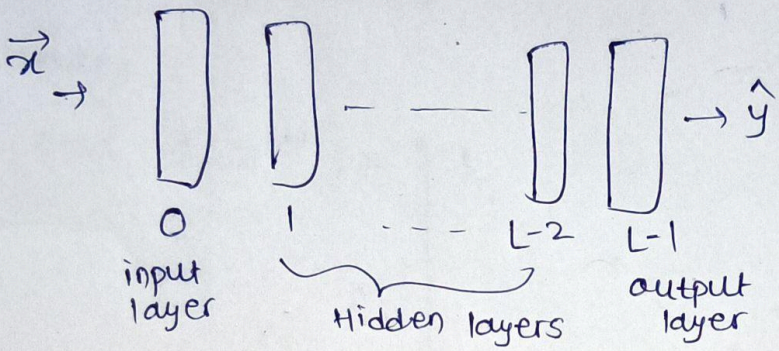
3. *Consider a MLP (assume an input layer, an output layer and $L-2$ hidden layers amounting to a total of L layers in the MLP) such that at the hidden layers and at the output layer, the linear activation function given by $\phi(z) = z$ is used. Consider that a sample (x, y) is used during training.*

(a) ***[2 marks]*** *Given the sample (x, y), write the expression to denote a prediction $\hat{y}$ from the MLP.*

(b) ***[3 marks]*** *Let c be the error for the sample (x, y). Write the expression to compute the error gradients with respect to the activations at the ℓ-th layer. Compare and contrast this expression with that derived in class.*

(c) ***[4 marks]*** *Discuss based on the expression obtained in part 3b, when the network can experience vanishing and exploding gradient problems.*

There are total L layers indexing from 0 to L-1.



a) $\hat{y} = \phi^{L-1}(W^{L-1}\phi^{L-2}(W^{L-2}\phi^{L-3}(W^{L-3} \dots \phi^{1}(W^{1}\vec{x})\dots)))$

since $\phi^{l}(z^{l}) = z^{l}$ for all hidden layers & output layer,

$\hat{y} = W^{L-1}.W^{L-2}W^{L-3} \dots W^{1}\vec{x}$

(all the weight matrices are pre-multiplied to their next term)

---

b) $\delta^{l} = V^{l+1}V^{l+2} \dots V^{L-1}\delta^{L-1}$

where $L-1$ = ~~old~~ index of output layer for given question.

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

where, $\mathrm{Diag}(\phi'(z)^{l+1}) = \begin{bmatrix} \phi'(z_1^{l+1}) & & & \\ & \phi'(z_2^{l+1}) & & \\ & & \phi'(z_3^{l+1}) & \\ & & & \ddots \\ & & & & \phi'(z_{N_{l+1}}^{l+1}) \end{bmatrix}$

Expression derived in the class

Since $\phi(z^l) = z^l$ (linear activation)

$\therefore \phi'(z^l) = 1$

Hence,

$$\text{Diag}(\phi'(z)^{l+1}) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}_{N_{l+1} \times N_{l+1}}$$ (Identity matrix)

$\therefore \text{Diag}(\phi'(z)^{l+1}) = I_{N_{l+1}}$ (identity matrix $N_{l+1} \times N_{l+1}$)

Hence,

$$V^{l+1} = (W^{l+1})^T \text{Diag}((\phi'(z))^{l+1}) = (W^{l+1})^T I_{N_{l+1}}$$

$$\therefore V^{l+1} = (W^{l+1})^T$$

putting in the expression of δ^l.

$$\therefore \delta^l = V^{l+1} V^{l+2} \cdots V^{L-1} \delta^{L-1}$$

$$\delta^l = (W^{l+1})^T (W^{l+2})^T \cdots (W^{L-1})^T \delta^{L-1}$$

where δ^{L-1} = error gradient wrt activation of output layer $= \frac{\partial e}{\partial a^{L-1}}$

c) error gradient wrt weight =

$$\frac{\partial e}{\partial w^l} = \nabla_{w^l} e = \underbrace{\text{Diag}(\phi'(z^l))}_{\because \phi'(z^l)=1} \delta^l (\vec{a}^{l-1})^T$$

$$= I_{N_l} \delta^l (\vec{a}^{l-1})^T$$

$$\nabla_{w^l} e = \delta^l (\vec{a}^{l-1})^T$$

putting δ^l from part b)

$$\nabla_{w^l} e = (w^{l+1})^T (w^{l+2})^T \cdots (w^{L-1})^T \delta^{L-1} (\vec{a}^{l-1})^T$$

Exploding gradient problem will occur →

① when $(\vec{a}^{l-1})^T$ takes very high values.
since we are using linear activation ($\phi(z^l) = z^l$),

$\vec{a}^{l-1} = \vec{z}^{l-1}$

Hence, there is no limit for the values which $\vec{a}^{l-1}$ can take. ② when the term $(w^{l+1})^T (w^{l+2})^T \cdots (w^{L-1})^T$ takes very high values.

Vanishing gradient problem will occur →

when the term $(w^{l+1})^T (w^{l+2})^T \cdots (w^{L-1})^T$ takes very small values.

4. Consider a data set $D = \{(x^t, y^t)\}$, $t = 1, 2, 3, \ldots$, where $x^t \in \mathbb{R}^d$ and $y^t \in \{+1, -1\}$, $\forall t = 1, 2, 3 \ldots$. Note that D is linearly separable if there exists a hyperplane $H = (w^*, b^*)$ (with $w^* \neq \mathbf{0}$) and $\gamma > 0$ such that $y^t(\langle w^*, x^t \rangle - b^*) \geq \gamma$ for every $t = 1, 2, 3, \ldots$.

(a) **[4 marks]** Consider a point x^p in D with $y^p = +1$. Find a point u on H closest to x^p and derive an expression for $\text{dist}(x^p, u)$, the distance between x^p and u. Explain how you constructed u.

(b) **[3 marks]** Consider a point x^q in D with $y^q = -1$. Find a point v on H closest to x^q and derive an expression for $\text{dist}(x^q, v)$, the distance between x^q and v. Explain how you constructed v.

(c) **[6 marks]** Let $\hat{x}$ denote the point with label $+1$ which is closest to H and let $\tilde{x}$ denote the point with label -1 which is closest to H. Suppose that the distance of $\hat{x}$ to the linear separator H is α and the distance of $\tilde{x}$ to the linear separator H is also α. Justify with proper reasons if and when the relations $\alpha = \gamma$, $\alpha > \gamma$ and $\alpha < \gamma$ can hold. (**Note:** The distance of a point $x \in \mathbb{R}^d$ to H is given by the distance $\text{dist}(x, u)$ where u is a point on H and is closest to x.)

Q4) Given:

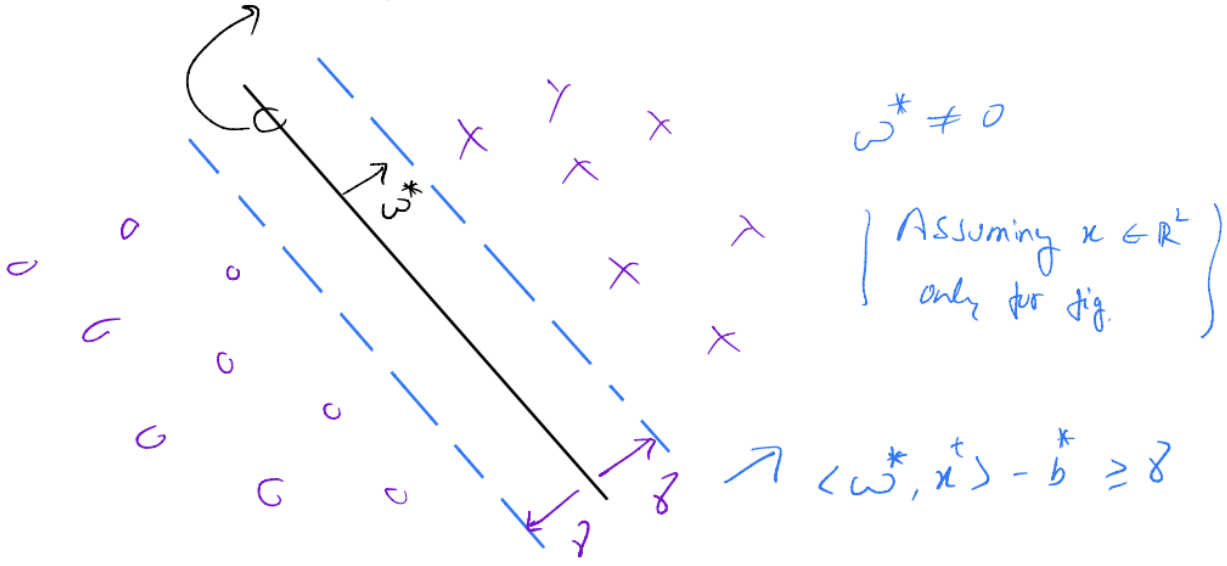
Dataset : $D = \{(x^t, y^t)\}$

feature vectors : $x^t \in \mathbb{R}^d$

class labels : $y^t \in \{+1, -1\}$

D is linearly separable if

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



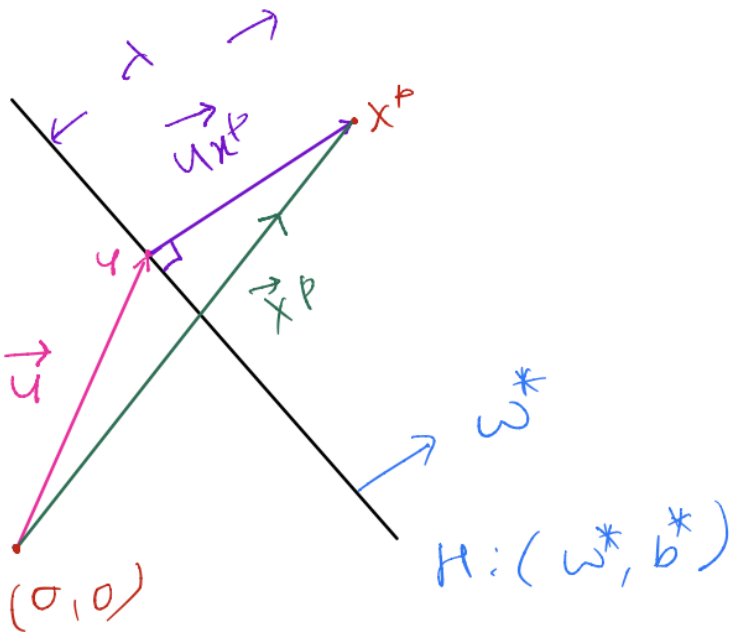
$w^* \neq 0$

Assuming $x \in \mathbb{R}^2$ only for fig.

$\langle w^*, x^t \rangle - b^* \geq \gamma$

$\langle w^*, x^t \rangle - b^* \leq -\gamma$

4. a)



→ Assuming (0,0) not lie on H

→ $\vec{u}$: closest point to $\vec{x^p}$ on H

→ $\vec{u}$ is chosen as the point intersecting H when a $\perp^r$ line is dropped on H from $\vec{x^p}$

→ $\vec{ux^p}$ is vector joining point u & point x^p with magnitude λ (λ to find)

Since, u is in the hyperplane H,

$$\langle w^*, u \rangle - b^* = 0 \qquad - (1)$$

By triangle law of vector addition ($\vec{x^p} = \vec{u} + \vec{ux^p}$)

$$\vec{x}^p = \vec{u} + \lambda \frac{\vec{w}^*}{\|\vec{w}^*\|} \quad - (2)$$

eq - (2) $\times (w^*)^T$

$$\vec{w}^{*T} \vec{x}^p = w^{*T} \cdot \vec{u} + \lambda \frac{\vec{w}^{*T} \cdot \vec{w}^*}{\|\vec{w}^*\|}$$

$$\langle \vec{w}^*, \vec{x}^p \rangle = \langle \vec{w}^*, \vec{u} \rangle + \frac{\lambda \langle \vec{w}^*, \vec{w}^* \rangle}{\|\vec{w}^*\|}$$

$$\langle \vec{w}^*, \vec{x}^p \rangle = b^* + \frac{\lambda \|\vec{w}^*\|^2}{\|\vec{w}^*\|} \quad (\text{from } ①)$$

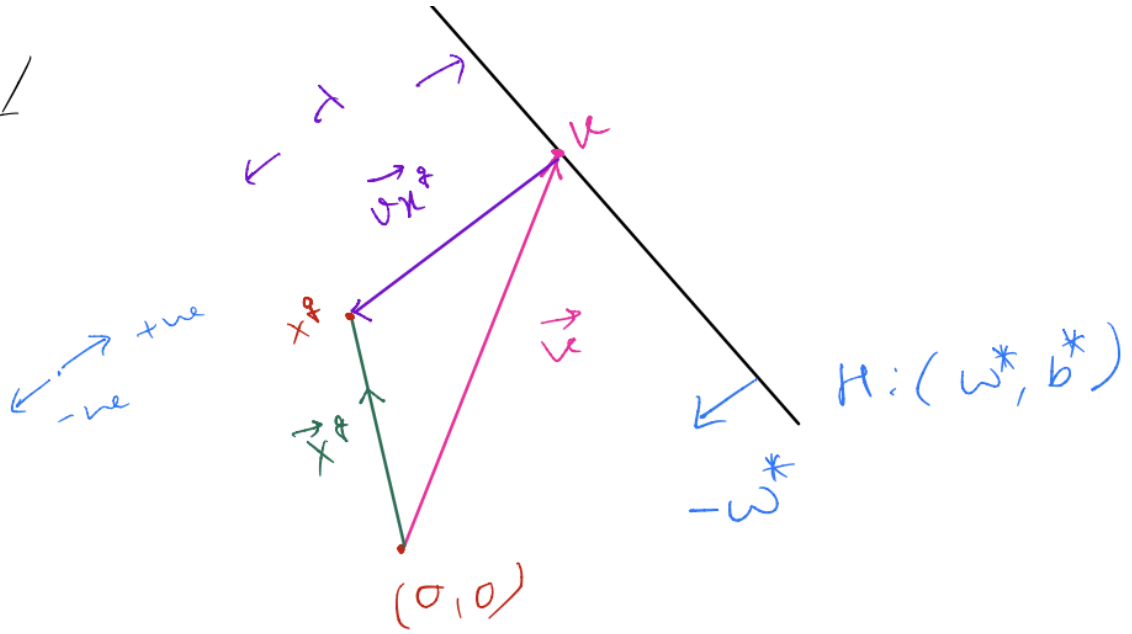
$$\lambda = \frac{\langle w^*, \vec{x}^p \rangle - b^*}{\|\vec{w}^*\|}$$

$$\Rightarrow \lambda = \left| \frac{\langle w^*, \vec{x}^p \rangle - b^*}{\|\vec{w}^*\|} \right|$$ Since λ is distance & cannot be negative.

Now substitute value of λ in eq^n (2) we get $\vec{u}$ as follows :

$$\vec{u} = \vec{x}^p - \left| \frac{\langle \vec{w}^*, \vec{x}^p \rangle - b^*}{\|\vec{w}^*\|} \right| * \frac{\vec{w}^*}{\|\vec{w}^*\|}$$

4.b)



→ Assuming (0,0) not lie on H

→ $\vec{v}$: closest point to $\vec{x^*}$ on H

→ $\vec{v}$ is chosen as the point intersecting H when a $\perp^r$ line is dropped on H from $\vec{x^*}$

→ $\vec{vx^*}$ is vector joining point v & point x^* with magnitude λ (λ to find)

Since, v is in the hyperplane H,

$$\langle w^*, v \rangle - b^* = 0 \qquad - (1)$$

By triangle law of vector addition ($\vec{x^*} = \vec{v} + \vec{vx^*}$)

$$\vec{x}^* = \vec{v} + \lambda^* \frac{(-\vec{w}^*)}{\|\vec{w}^*\|} \quad - (2)$$

eq - (2) $\times (w^*)^T$

$$\vec{w}^{*T} \vec{x}^* = w^{*T} \cdot \vec{v} - \lambda^* \frac{(\vec{w}^{*T} \cdot \vec{w}^*)}{\|\vec{w}^*\|}$$

$$\langle \vec{w}^*, \vec{x}^* \rangle = \langle \vec{w}^*, \vec{v} \rangle - \frac{\lambda \langle \vec{w}^*, \vec{w}^* \rangle}{\|\vec{w}^*\|}$$

$$\langle \vec{w}^*, \vec{x}^* \rangle = b^* - \frac{\lambda \|\vec{w}^*\|^2}{\|\vec{w}^*\|} \quad (\text{from } ①)$$

$$\lambda = \frac{-\langle w^*, \vec{x}^* \rangle + b^*}{\|\vec{w}^*\|}$$

$$\Rightarrow \lambda = \left| \frac{\langle w^*, \vec{x}^* \rangle - b^*}{\|\vec{w}^*\|} \right|$$ since λ is distance & cannot be negative.

Now substitute value of λ in eq^n (2) we get $\vec{v}$ as follows :

$$\vec{v} = \vec{x}^* + \left| \frac{\langle \vec{w}^*, \vec{x}^* \rangle - b^*}{\|\vec{w}^*\|} \right| * \frac{\vec{w}^*}{\|\vec{w}^*\|}$$

4.c)

using eq^{n} of δ from 4.a & 4.b

$$\alpha = \frac{|\langle w^*, \hat{x}\rangle - b|}{\|w^*\|} = \frac{|\langle w^*, \tilde{x}\rangle - b|}{\|w^*\|} \quad -(1)$$

⇒ From given linear separability condition:

$$y^t(\langle w^*, x^t\rangle - b^*) \geq \delta$$

If x^t has label $y^t = +1$,

then, $(\langle w^*, x^t\rangle - b^*) \geq \delta$ — (2)

If x^t has label $y^t = -1$,

then, $-(\langle w^*, x^t\rangle - b^*) \geq \delta$

$\Rightarrow (\langle w^*, x^t\rangle - b^*) \leq -\delta$ — (3)

from (2) & (3) we can write,

$$|\langle w^*, x^t\rangle - b^*| \geq \delta \quad \forall\ x^t \in D$$

$\therefore \hat{x}$ & $\tilde{x} \in D$

Thus $|\langle w^*, \hat{x}\rangle - b^*| \geq \delta$ — (4)

(At place of $\hat{x}$, $\tilde{x}$ can also be used)

In eq-n ④ Dividing $\|w^*\|$ on both side

$$\frac{|\langle w^*, \hat{x} \rangle - b^*|}{\|w^*\|} \geq \frac{\gamma}{\|w^*\|}$$

$$\boxed{\alpha \geq \frac{\gamma}{\|w^*\|}} \quad \text{(From eq-n ①)} \quad - ⑤$$

Now we will analyze the case for $\|w^*\|$ in 3 intervals $\|w^*\| = 1$, $\|w^*\| < 1$ & $\|w^*\| > 1$

Case 1: $\|w^*\| = 1$: From eq-⑤

$$\alpha \geq \frac{\gamma}{\|w^*\|} \Rightarrow \alpha \geq \gamma$$

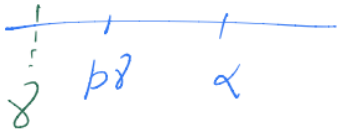
{when $\alpha = \gamma$}: This holds when the point lies exactly on the margin defined by γ, meaning the distance from the point to the hyperplane is exactly γ.

{when $\alpha > \gamma$}: This happens for closest points ($\hat{x}$ & $\tilde{x}$) that lie farther from the hyperplane than the margin γ but still satisfy the linear separability condition.

Case 2 : $0 < \|w^*\| < 1$

from eq-n ② : $\alpha \geq \frac{\gamma}{\|w^*\|}$ ⇒ let $\|w^*\| = 1/p$ $(p > 1)$

so, $\alpha \geq \frac{\gamma}{1/p}$ ⇒ $\alpha \geq p\gamma$



⇒ $\boxed{\alpha > \gamma}$

{when $\alpha > \gamma$}: This happens for closest points ($\hat{x}$ & $\tilde{x}$) that lie farther from the hyperplane than the margin γ but still satisfy the linear separability condition.

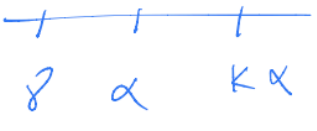
Case 3 : $\|w^*\| > 1$

from eq-n ⑤ : $\alpha \geq \frac{\gamma}{\|w^*\|}$ ⇒ let $\|w^*\| = K$ $(K > 1)$

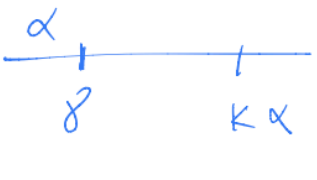
$K\alpha \geq \gamma$ ⇒ $\gamma \leq K\alpha$

Now 3 case can be possible

$\alpha < \gamma$



$\alpha > \gamma$



$\alpha = \gamma$

Here $\alpha < \gamma$, $\alpha > \gamma$ & $\alpha = \gamma$ all 3 are possible

{when $\alpha = \gamma$}: This holds when the point lies exactly on the margin defined by γ, meaning the distance from the point to the hyperplane is exactly γ

{when $\alpha > \gamma$}: This happens for closest points ($\hat{x}$ & $\tilde{x}$) that lie farther from the hyperplane than the margin γ but still satisfy the linear separability condition.

{when $\alpha < \gamma$}: happens when a points lies inside the margin, meaning it's closer to the hyperplane that the margin threshold. This is possible for points that are still correctly classified but do not respect the margin condition.

Conclusion:

- when $\|w^*\| = 1$, Possible case: $\alpha = \gamma$, $\alpha > \gamma$
- when $\|w^*\| > 1$, Possible case: $\alpha > \gamma$
- when $\|w^*\| < 1$, Possible case: $\alpha = \gamma$, $\alpha > \gamma$, $\alpha < \gamma$