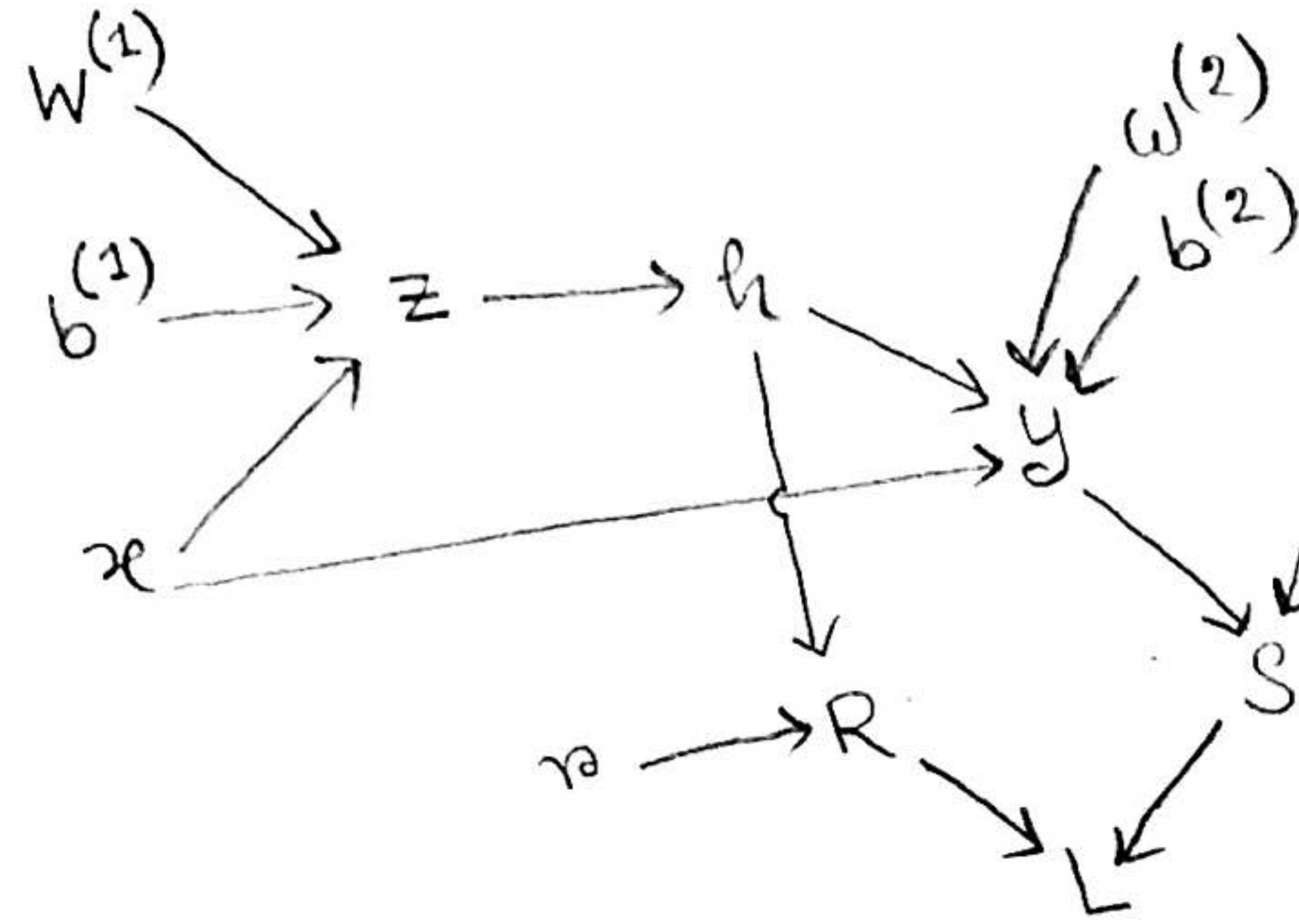
Asadullah Hill Galib

CSE-891 : Homework 2

October, 11

(1) (1)



(1) (2) $\frac{dL}{dx} = x' = ?$

$L' = 1$

$S' = L'$

$R' = L'$

$y' = S'(y - s)$

$h' = y' \cdot w^{(2)} + R' \cdot r^T$

$z' = h' \cdot \sigma'(z)$

$x' = z' w^{(1)} + y' \cdot 1$

(Ans)

(2) (1)

$f(x) = vv^T x$

$J = vv^T$

$n = 3,\ v^T = [1, 2, 3]$

$$J = vv^T = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

(2) (2)

The time and memory cost of evaluating the Jacobian is $O(n^2)$. [There are $n \times n$ calculation & $n \times n$ grids ~~in~~ in memory are needed]

(2) (3)

$z = J^T y$

$= vv^T y$ [transpose of $J = J$]

Here, instead of evaluating $(v.v^T).y$, one should do this : $\underbrace{v.\underbrace{(v^T.y)}_{\text{first}}}_{\text{second}}$

first → linear in n [1×1]

second → linear in n [3×1]

So, $z = J^T y$

~~So,~~ z ~~=~~ $v.(v^T y)$

$$= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} [1 \cdot 2\ 3] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot 6 = \begin{bmatrix} 6 \\ 12 \\ 18 \end{bmatrix} \text{(Ans)}$$

(3)(1)

$$\text{Loss}, \mathcal{L} = \frac{1}{n} \| X\hat{\omega} - t \|_2^2$$

$$\frac{d\mathcal{L}}{d\hat{\omega}} = \cancel{\frac{2}{n}(X\omega} \; \frac{2}{n} X^T (X\hat{\omega} - t)$$

$$= \frac{2}{n} \left(X^T X \hat{\omega} - X^T t \right)$$

(3)(2) Underparameterized:

(a)

$$\frac{d\mathcal{L}}{d\hat{\omega}} = 0$$

$$\Rightarrow \frac{2}{n} (X^T X \hat{\omega} - X^T t) = 0$$

$$\Rightarrow X^T X \hat{\omega} = X^T t$$

$$\therefore \hat{\omega} = (X^T X)^{-1} X^T t$$ [as $X^T X$ is invertible for $n > d$]

(3)(2)(b) $t_i = \omega^{*T} x_i$.

$\cancel{t = X w^*}$ $t = X\omega^*$

So, $\hat{\omega} = (X^T X)^{-1} X^T X \omega^* = \omega^*$

Therefore, $\forall x \in \mathbb{R}^d$, $(\omega^{*T} x - \hat{\omega}^T x)^2 = 0$ [when $d < n$]

and $\hat{\omega}$ achieves perfect generalization.

(3) (3) Overparameterized Model: 2D

(a)

$n = 1,\ d = 2,\ x_1 = [2\ \ 1] \qquad y_1 = 2$

$\hat{\omega}^T x_1 = y_1$. Let, $\hat{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}$

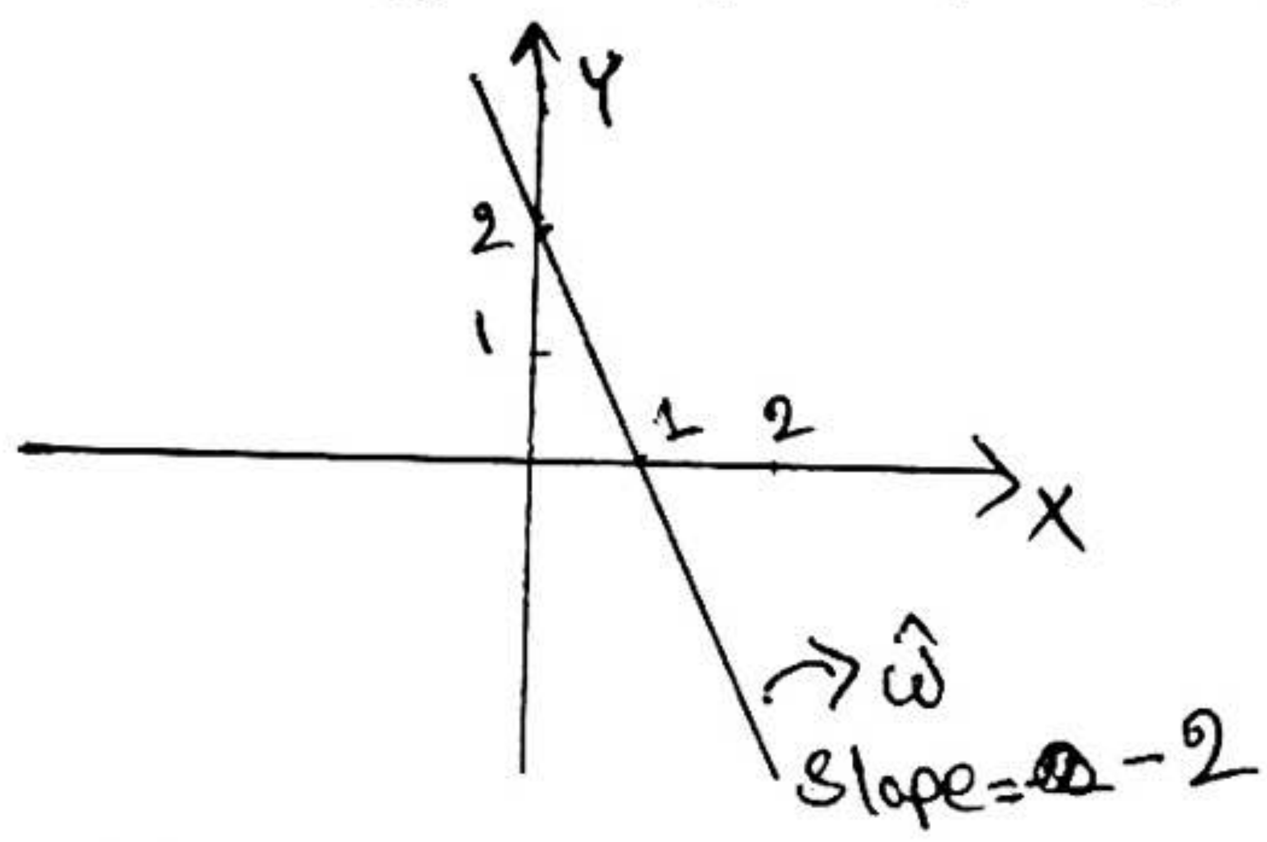
So, $[\omega_1\ \omega_2] \begin{bmatrix} 2 \\ 1 \end{bmatrix} = y_1 = 2$

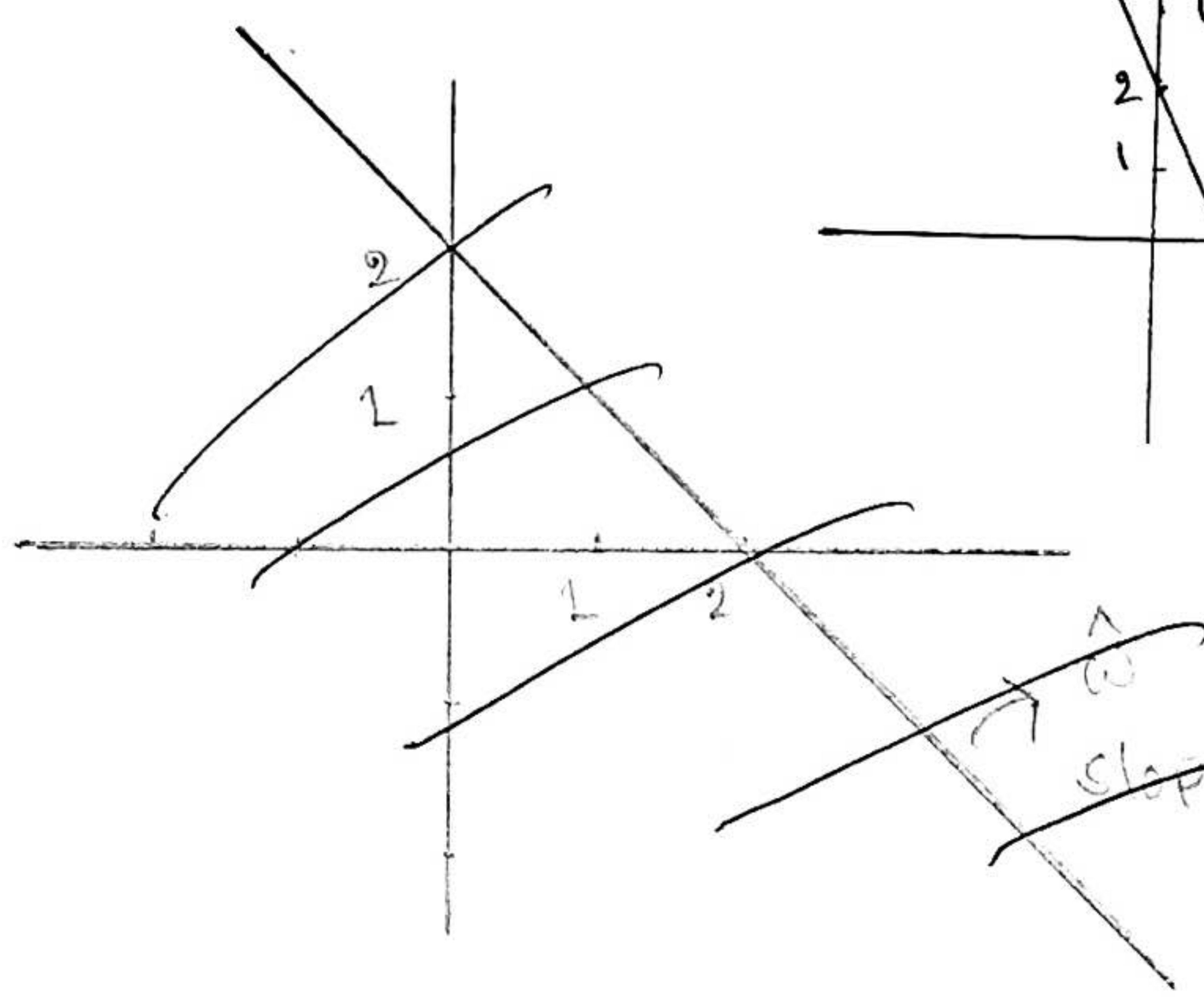
$\Rightarrow 2\omega_1 + \omega_2 = 2$

$\Rightarrow \omega_2 = -2\omega_1 + 2 \rightarrow$ equation of line

So, there exists infinitely many $\hat{\omega}$, satisfying $\hat{\omega}^T x_1 = y_1$

as ~~there's no~~ the solution can be anywhere on the line.





3 3 b

$\hat{\omega}(0) = 0$

$$\frac{dL}{d\hat{\omega}} = \frac{2}{n} X^T (X\hat{\omega} - t)$$

when, $\hat{\omega}(0) = 0$ and ~~$X = X_1$~~, $X = X_1$, $t = t_1$, $n = 1$, $d = 2$:

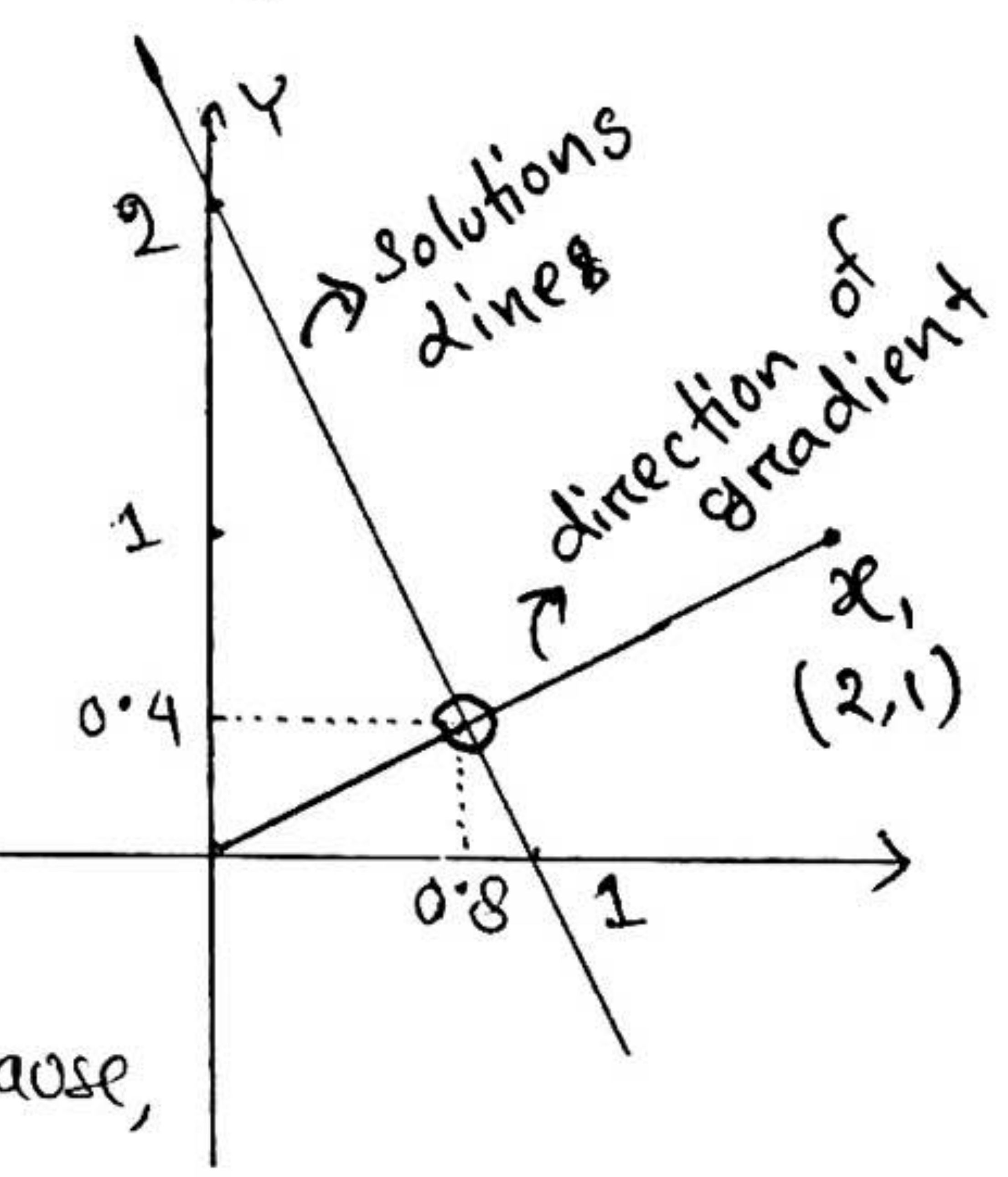
$$\frac{dL}{d\hat{\omega}} = \frac{2}{n}\left(-X^T t\right) = \left(-2X_1^T t_1\right)$$

As, $t_1 = 2$ a constant, the direction of the gradient is along X_1. And, it doesn't change along the trajectory as there is no $\hat{\omega}$ term in the derivative.

The corresponding unit norm vector of X_1: $\frac{1}{\sqrt{5}}\begin{bmatrix}2\\1\end{bmatrix}$

Using the above, we get

$$\hat{\omega} = 2 \cdot \frac{1}{(\sqrt{5})^2}\begin{bmatrix}2\\1\end{bmatrix} = \begin{bmatrix}0.8\\0.4\end{bmatrix}$$ (using squared-norm)



Here,

the gradient descent finds the closest solution from w_0 (0,0).

It is the closest distance from O because, the direction of gradient and solutions line are orthogonal to each other (i.e. slope of the solutions line, $m_1 = -2$

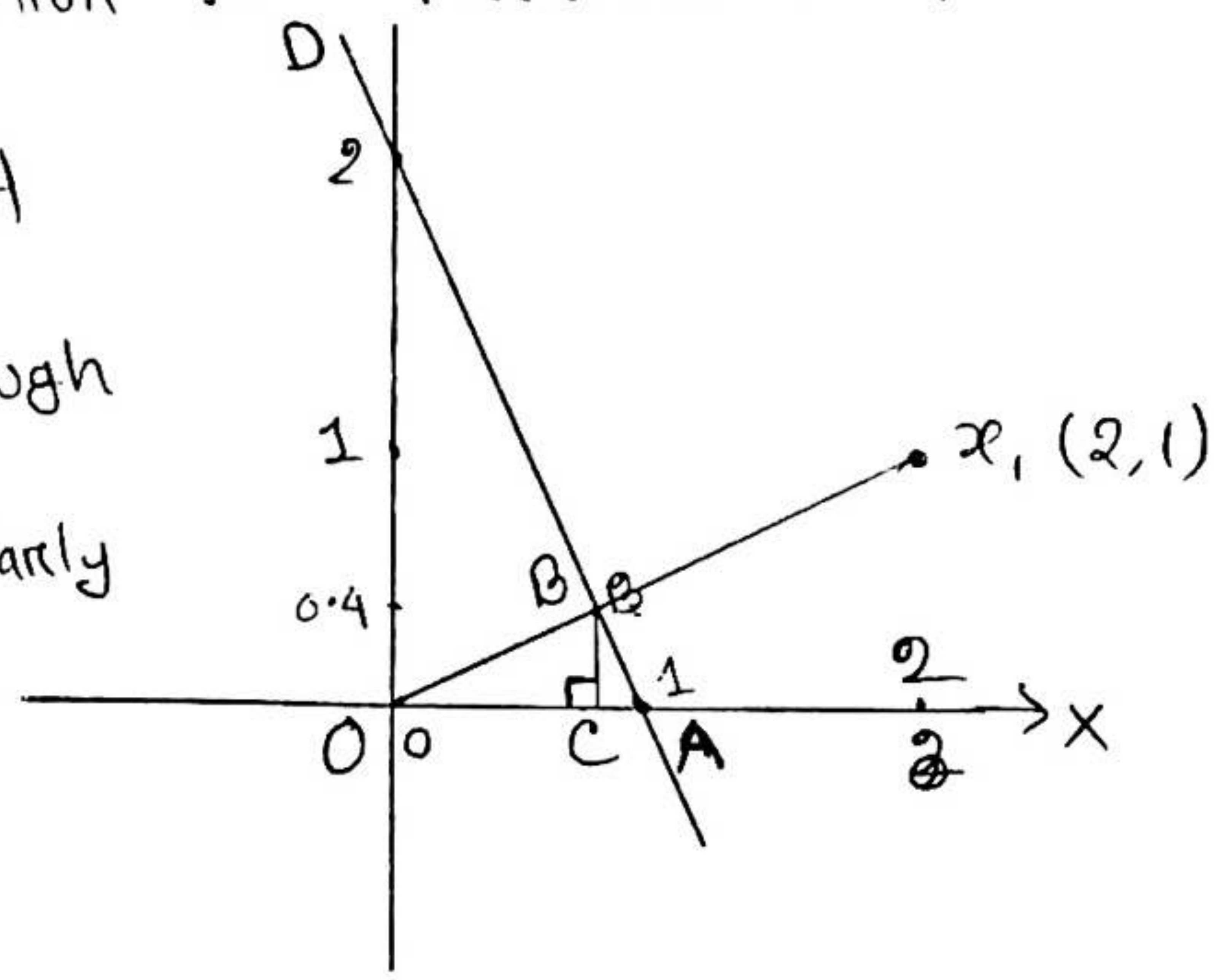
" " " gradient des. ", $m_2 = \frac{1}{2}$

$m_1 m_2 = -1 \Rightarrow$ perpendicular)

(3)(3)(c)

Using Pythagoream Theorem:

B is the gradient descent solution from previous part.



To show, B has the smallest Euclidean norm, it is enough to prove that OB perpendecularly intersects AD ($OB \perp AD$)

Here, $OA = 1$

$OC = 0.8$

$BC = 0.4$ [$BC \perp OA$]

So, $OB = \sqrt{OC^2 + BC^2} = \sqrt{\frac{2}{5}}$

$AB = \sqrt{CA^2 + BC^2} = \sqrt{(OA - OC)^2 + BC^2}$

$= \sqrt{(0.2)^2 + (0.4)^2} = \sqrt{\frac{1}{5}}$

In $\cancel{\Delta OAB}$ ΔOBA,

$OB^2 + AB^2 = \frac{4}{5} + \frac{1}{5} = 1 = OA^2$

So, $\cancel{\Delta OAB}$ ΔOBA is a right-angled Δ. $\Rightarrow OB \perp AB$

$\Rightarrow OB \perp AD$

[Proved]

—o—

Alternative Proof:

(*) It can also be shown ~~also~~ using slopes:

$m_{AD} = -2$ and $m_{OB} = 1/2$

So, $m_{AD} \cdot m_{OB} = -1 \Rightarrow OB \perp AD \Rightarrow$ OB has the smallest norm.

③ ④ ⓐ With, $\hat{\omega}(0) = 0$, we get gradient vector in the span of X. [previous part suggests]

As the gradient vector is always spanned by the rows of X, we can get $\hat{\omega}$ as a linear combination of X and some other matrix. Let, P is that matrix.

So, $\hat{\omega} = X^T P$

Then, $\frac{2}{n} X^T (X\hat{\omega} - t) = 0$

$$\Rightarrow X^T(XX^TP - t) = 0$$

$$\Rightarrow XX^T(XX^TP - t) = X.0 = 0$$

$$\Rightarrow XX^TP - t = (XX^T)^{-1}.0 = 0 \quad [\text{for } d > n,\ XX^T \text{ is invertible}]$$

$$\Rightarrow XX^TP = t$$

$$\Rightarrow P = (XX^T)^{-1} t$$

Thereby, $\hat{\omega} = X^T P$

$$= X^T (XX^T)^{-1} t$$

The solution is unique as it's just a linear transformation of t.

③ ④ ⓑ

Zero-loss solution with $\hat{\omega}_1$.

So, $\hat{\omega}_1^T X - t = 0 \Rightarrow \hat{\omega}_1^T X = t \; \ldots ①$

$$(\hat{\omega} - \hat{\omega}_1)^T \hat{\omega} = \left(X^T (XX^T)^{-1} t - \hat{\omega}_1\right)^T \hat{\omega}$$

$$= \left(t^T (XX^T)^{-1} X - \hat{\omega}_1\right) \cancel{\hat{\omega}} \left(X^T (XX^T)^{-1} t\right)$$

$$= \left(t^T (XX^T)^{-1} X X^T (XX^T)^{-1} t - \hat{\omega}_1 X^T (XX^T)^{-1} t\right)$$

$$= t^T (XX^T)^{-1} t - t^T (XX^T)^{-1} t \quad [\text{using } ①]$$

$$= 0$$

So, $(\hat{\omega} - \hat{\omega}_1)$ and $\hat{\omega}$ are perpendicular to each other.

$\Rightarrow \hat{\omega}_1$ and $\hat{\omega}$ are perpendicular to each other.

~~To~~ So, like $\hat{\omega}_1$, all other solutions are perpendicular to $\hat{\omega}$. And, this gradient descent solution, $\hat{\omega}$ has the smallest Euclidean norm similarwise we proved before (using Pythagoream theorem).