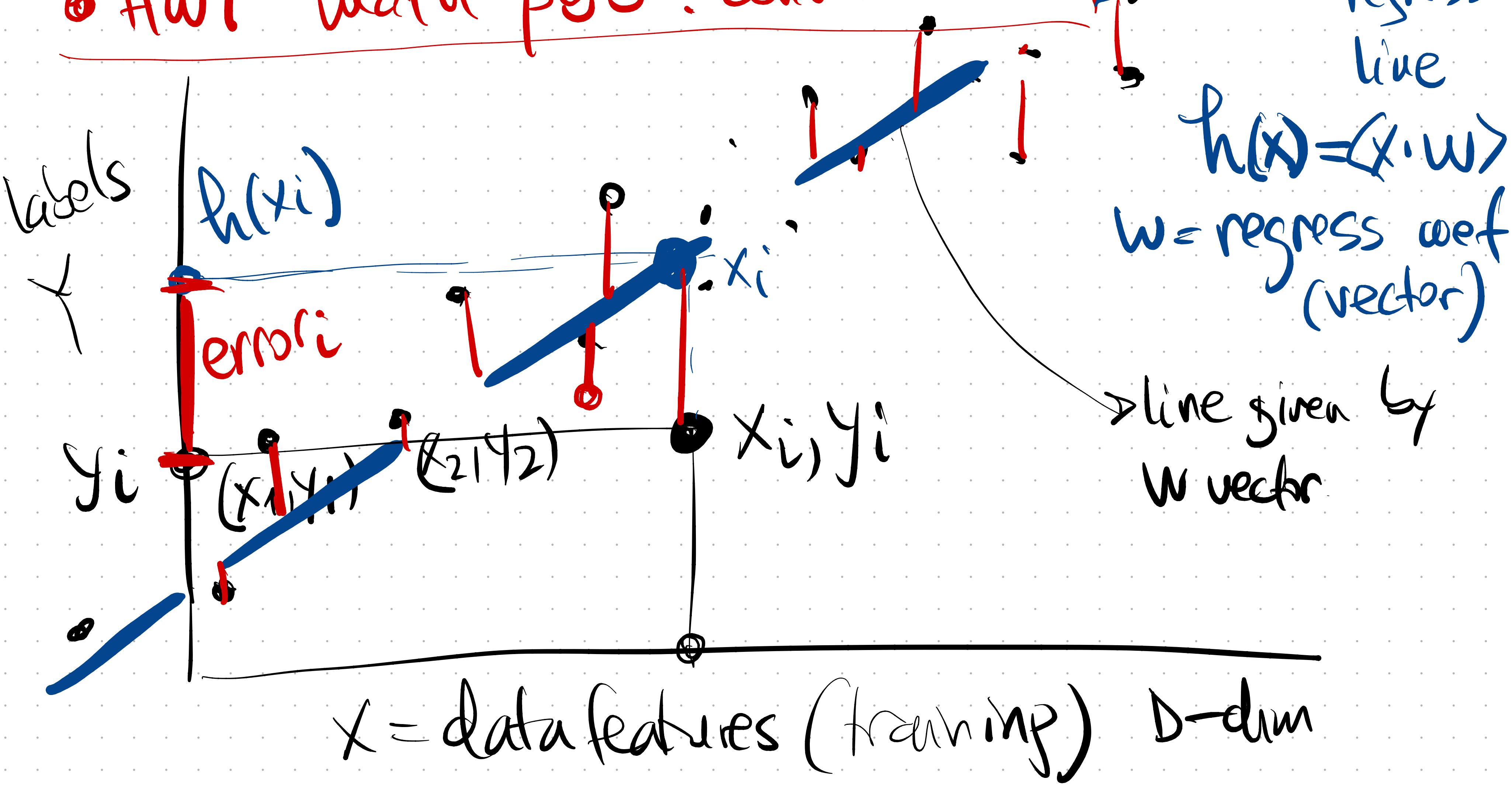
Lecture 5/14/25

- Linear regression
- HW1 (demo brief) 7:20 - Caleb
- HW1 math pb5 : convex hulls



Regression line (classifier) by coef $w = (w^1\ w^2 \cdots w^D)$

datapoint x_i

$$h(x_i) = \langle x_i \cdot w \rangle = \sum_{d=1}^{D} x_i^d \cdot w^d$$

want $\underbrace{h(x_i)}_{\text{predict}} \simeq \underbrace{y_i}_{\text{label}}$

$$\text{sq error}_i = (h(x_i) - y_i)^2 = (\langle x_i w \rangle - y_i)^2$$

$$x_i = (x_i^1\ x_i^2 \cdots x_i^D)$$

All datapoints $\Rightarrow$ matrix $X_{N \times D}$, Y

$$\begin{array}{c|cccc|}
 & 1 & 2 & & D \\
\hline
x_1 & x_1^1 & x_1^2 & \cdots & x_1^D \\
x_2 & x_2^1 & x_2^2 & \cdots & x_2^D \\
 & & & X & \\
x_N & x_N^1 & x_N^2 & \cdots & x_N^D \\
\hline
\end{array}
\qquad
\begin{array}{|c|}
\hline
y_1 \\
y_2 \\
\vdots \\
y_N \\
\hline
\end{array}$$

$$\begin{array}{c} \\ x_i \\ \\ N \end{array} \begin{bmatrix} & & & \\ x_i^1 & \cdots & \cdots & x_i^D \\ & & & \end{bmatrix}_{D} \cdot \overset{w}{\begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix}}_{D\times 1} = i \underbrace{\begin{bmatrix} x_1 w \\ x_2 w \\ \vdots \\ x_i w \\ \vdots \\ x_N w \end{bmatrix}}_{X \cdot w} \text{ VS labels (diff) } \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

def

$$\boxed{\underset{N\times D}{X} \cdot \underset{D\times 1}{W} - \underset{N\times 1}{Y} = \underset{N\times 1}{E}} = \begin{bmatrix} x_1 w - y_1 \\ x_2 w - y_2 \\ x_N w - y_N \end{bmatrix}; \quad E^T = [x_1 w - y_1, x_2 w - y_2, \dots, x_N w - y_N]$$

want sq error: $J(w) = \frac{1}{2}\sum_{i=1}^{N} (h(x_i) - y_i)^2 = \frac{1}{2} \underset{1\times N}{E^T} \cdot \underset{N\times 1}{E} = \frac{1}{2}\underbrace{(Xw-Y)'}_{1\times N \; E^T}\underbrace{(Xw-Y)}_{E}$

$$\langle E^T \cdot E \rangle = \sum_{i=1}^{N} e_i \cdot e_i = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (x_i w - y_i)^2$$

<u>want</u>: to find w vector $w = (w^1 w^2 \dots w^D)$ to minimise the sq error $\frac{1}{2}(Xw-Y)^T \cdot (Xw-Y)$

CONVEX: $\frac{\partial \partial}{\partial w} > 0$ positive intuition $J(w) = \frac{1}{2}(Xw - Y)^T(Xw - Y)$ convex in w

$\Rightarrow \frac{\partial J}{\partial w} = 0$ when $J(w)$ is minimum.

Q?

1×1 dim $(1\times D)(D\times N)(N\times1)$

same transposed

$(AB)^T = B^T A^T$

$$\frac{\partial J}{\partial w} = \frac{1}{2}\frac{\partial}{\partial w}(Xw - Y)^T(Xw - Y) + \frac{1}{2}\frac{\partial}{\partial w} w^T w \cdot \lambda$$

$$\frac{\partial J}{\partial w} = 0$$

$$= \frac{1}{2}\partial/\partial w\,(w^T X^T X w - w^T X^T Y - Y^T X w - Y^T Y) + \frac{1}{2}\frac{\partial}{\partial w} w^T w \cdot \lambda$$

$$= \frac{1}{2}\partial/\partial w\,(w^T X^T X w - 2 w^T X^T Y - Y^T Y) + \frac{1}{2}\frac{\partial}{\partial w} w^T w \cdot \lambda$$

? $\partial = 0$

pretend 1-dim

$$= \frac{1}{2}\partial/\partial w\; w^T X^T X w - \partial/\partial w\; w^T X^T Y + \lambda I \cdot w$$

pretend is 1-dim(w) pretend 1-dim

id. matrix

$$= \frac{1}{2} 2 X^T X w - X^T Y + \lambda I \cdot w$$

$$= X^T X w - X^T Y + \lambda I w$$

want 0

sq $D\times D$, pos definite (pos eig val)

$$\boxed{X^T X} w = X^T Y \longrightarrow (X^T X + \lambda I) w = X^T Y$$

invert-left by $(X^T X)^{-1}$

$$\boxed{w = (X^T X + \lambda I)^{-1} \cdot X^T Y}$$

L_2-reg

$$\boxed{w = (X^T X)^{-1} X^T Y}$$

NORMAL EQ (closed form)

Best (Th) w for min sq error of regr. line

1-dim w differential

$$\frac{\partial}{\partial w}\left(\underline{w}^T \cdot C \cdot \underline{w}\right) = \frac{\partial}{\partial w}\left(w^2 \cdot c\right) = \underline{2wc}$$

$\frac{\partial}{\partial \underline{w}} = ?$ ⇒ let math people worry about this!

↘ vector D-dim

⇒ if interested, talk about at OH (not required for course)

Naive way: we want $X \cdot W \simeq Y$ (NxD, Dx1, Nx1)

why not (if $X \simeq$ sq matrix) make it square?

$W = X^{-1} Y$?

unstable, W not min sq err.

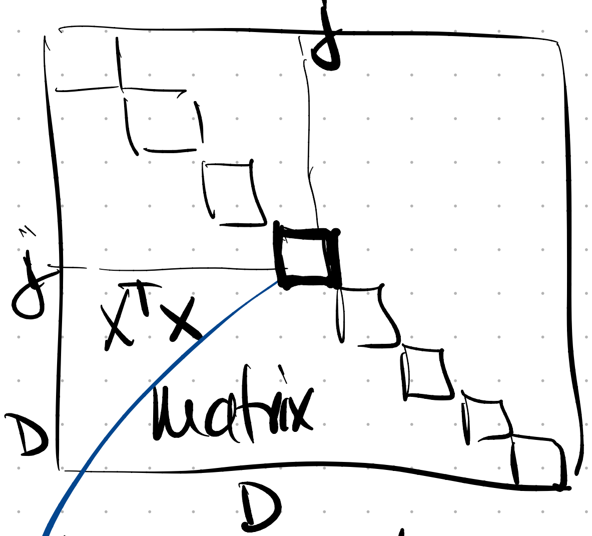
X^{-1} terrible to invert

Correct way

$X^T X$ stable matrix, pos def → (DxD)

⇒ invert this matrix

X = centered (all column mean = 0) ⇒ $X^T X = N \cdot \mathrm{COVAR}(x)$



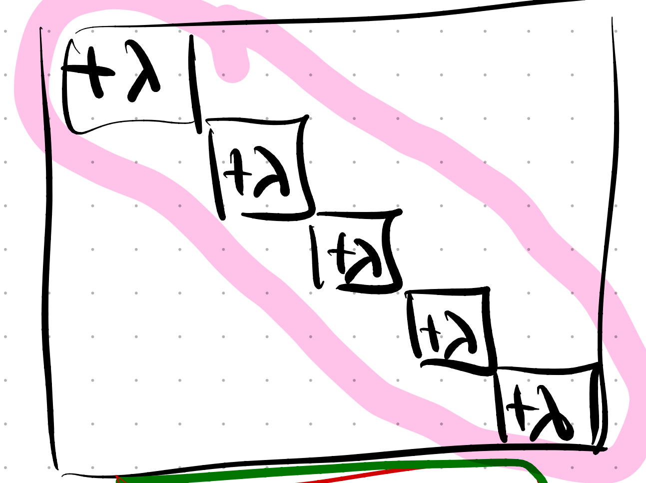
diagonal vals (pos j)

$\langle \text{column } j \text{ of } X, \text{column } j \text{ of } X \rangle$

var(j feature) = var(X^j)

add stability $\Longrightarrow$

(for inversion $(X^TX)^{-1}$)

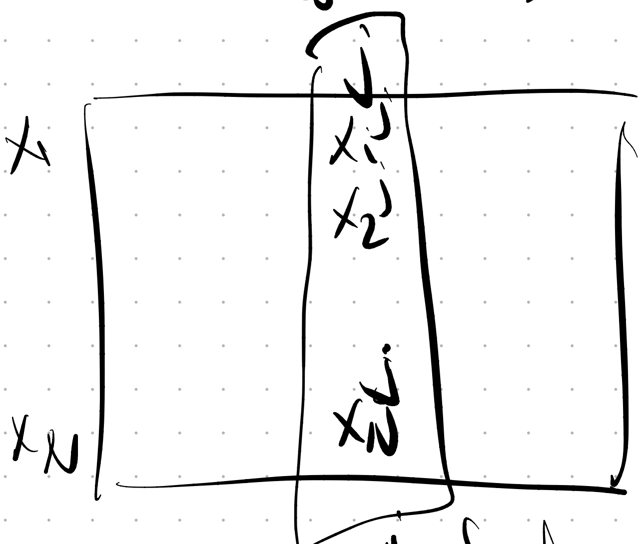


Increase main diagonal

$\boxed{X^TX + \lambda I}$ more stable $\neq X^TX$

identity

$$I = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$



X^j = feature j

$\boxed{\lambda = \text{hyperparam of } L_2 \text{ reg}}$

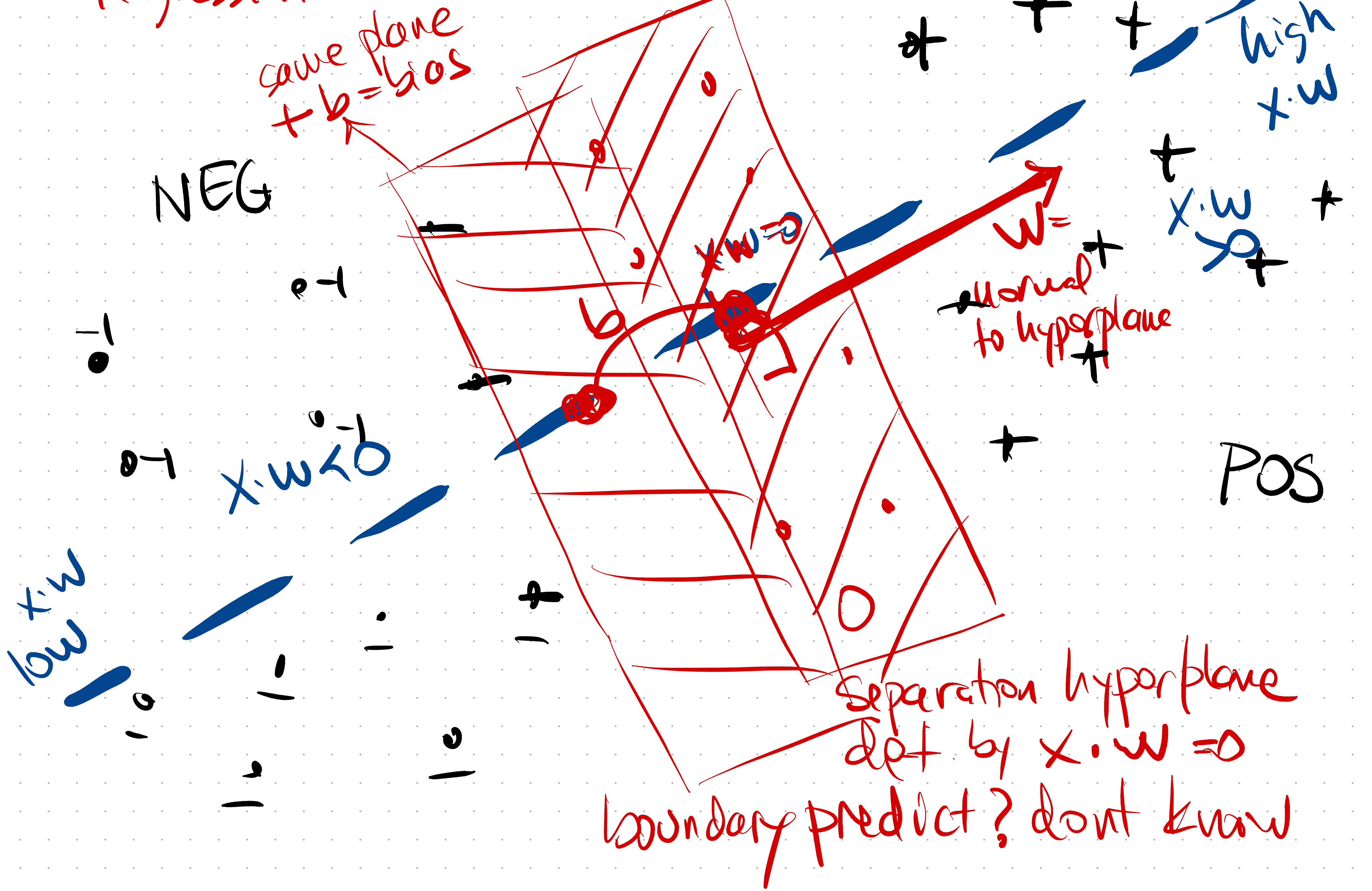
this op $X^TX + \lambda I$ actually corresponds to regularized L_2 error

$$\underbrace{J(w)}_{\text{sq error}} = \underbrace{\tfrac{1}{2}(Xw-y)^T(Xw-y)}_{OBJ} + \tfrac{1}{2}\lambda \sum_{d=1}^{D} (w^d)^2$$

$$= \frac{1}{2}(Xw-y)^T(Xw-y) + \frac{1}{2}\lambda w^T w$$

prevents large w values $\Longleftarrow$ L_2 regularizat

Regression Visual intuition



HW1 PB3: 1-dim $w = a$ bias $= b$

Reg $h(x_i) = x_i \cdot a + b \underset{\sim}{\overset{\text{want}}{\sim}} y_i \quad \forall i = 1:n$

$$\begin{array}{cc} \text{1-dim} & \text{"1" vec} \\ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \end{array} \cdot \begin{bmatrix} a & b \end{bmatrix} = \begin{bmatrix} x_1 \cdot a + b \\ x_2 \cdot a + b \\ \vdots \\ x_N a + b \end{bmatrix} \text{ vs } \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$