Linear Regression

$\mathbf{x} \longrightarrow y$

$\mathbf{x}$ is a vector

$\mathbf{x} \in \mathbb{R}^d \rightarrow$ $\mathbf{x}$ is a vector of length d

y is a scalar

$y \in \mathbb{R}$

Prediction or Regression

Predict future income

Current GPA , # AI courses taken

$y = 7000$

$y = 300000$

$\boxed{3.8 \mid 4}$ y ?

Training data

GPA , # AI

Income

10000

3.7, 2	100000
3.9, 6	500000
2.4, 1	1,000,000

3.5, 2	?

Functional models.

$$y = f(\mathbf{x})$$

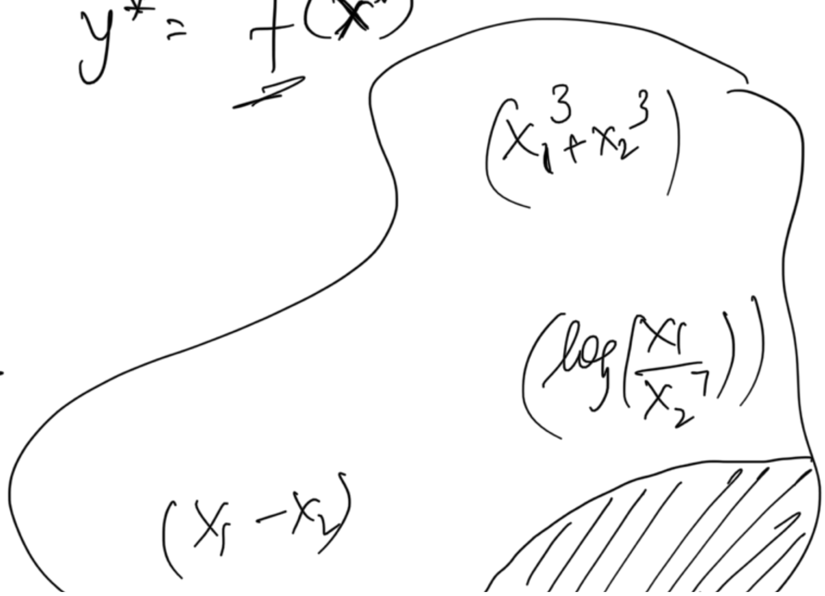
some function.

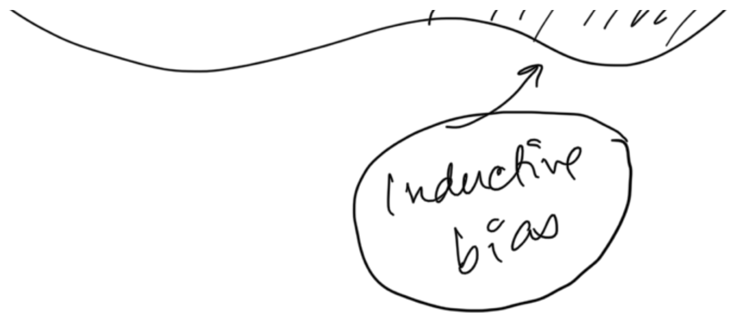
If we learn $f()$,

then for a new $\mathbf{x}^*$ $\mathbf{x} = \boxed{x_1, x_2}$

$$y^* = \underline{f}(\mathbf{x}^*)$$

a big bag of functions





Monday Feb 8

$$\mathbf{X} \longrightarrow Y$$

Functional models	Probabilistic Models.
$Y = f(\mathbf{X})$	$p(\mathbf{X}, Y)$
	$p(Y \mid \mathbf{X}) \leftarrow$ Bayes Rule

$$\mathbf{X} \rightarrow \begin{array}{c} X_1 \\ \boxed{x_1 \mid x_2 \mid \; \mid \; \mid \; \mid \; \mid \; \mid} \\ - - - - - x_D \end{array}$$

$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$

$\mathbf{X}_1 \quad \mathbf{X}_2$ $\qquad \boxed{X_{11} \quad X_{12} \; - - -}$

D

$$\mathbf{x}\cdot\mathbf{y} = \sum_{i=1} x_i y_i$$

$$= x_1 y_1 + x_2 y_2 + \dots + x_D y_D$$

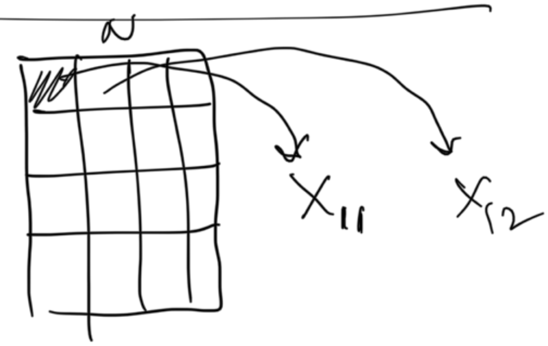
$$|\mathbf{x}| = \sum_{i=1}^{D} |x_i|$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{D} x_i^2}$$ l_2 norm

$\|\mathbf{x}\|$

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{D} x_i^p\right)^{1/p}$$

Matrix

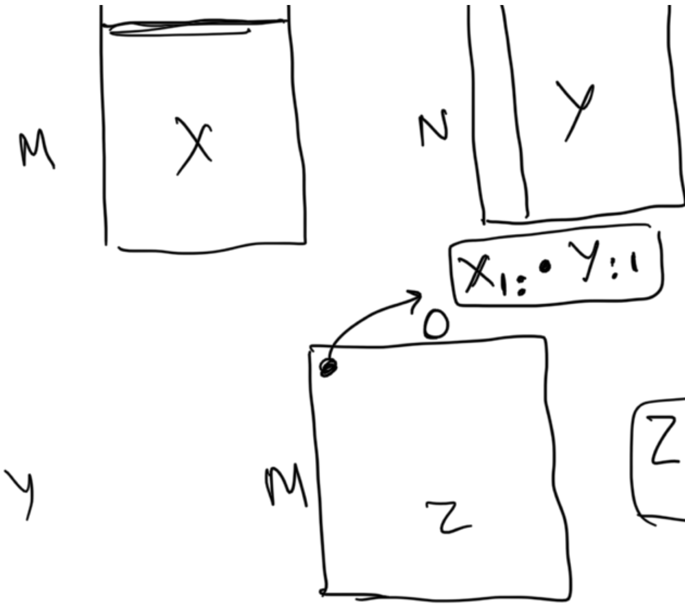


$$\boxed{\mathbf{X} \in R^{M\times N}}$$

$$Y = X^T \quad (N\times M)$$

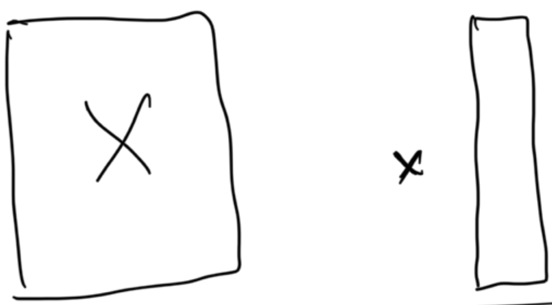
$$c\mathbf{X} = \begin{bmatrix} cx_{11} & cx_{12} & \cdots \\ \vdots & \ddots & \\ & & \end{bmatrix}$$

$\mathbf{X}\mathbf{Y}$



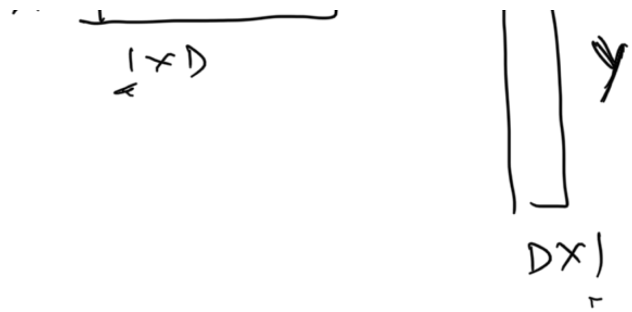
$Z = XY$

$$\boxed{Z_{ij} = \underline{X_{i:} \cdot Y_{:j}}}$$

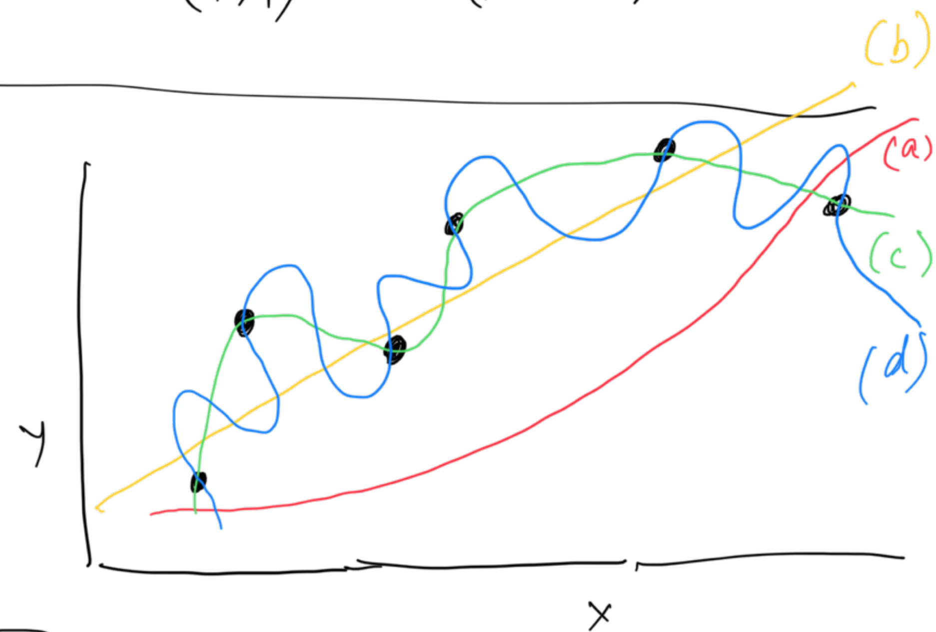


A D-length vector is a (D×1) matrix

$$x \cdot y = x^T y$$



$$z = x^T y = (1\times 1) = (x^T) \cdot y$$



error

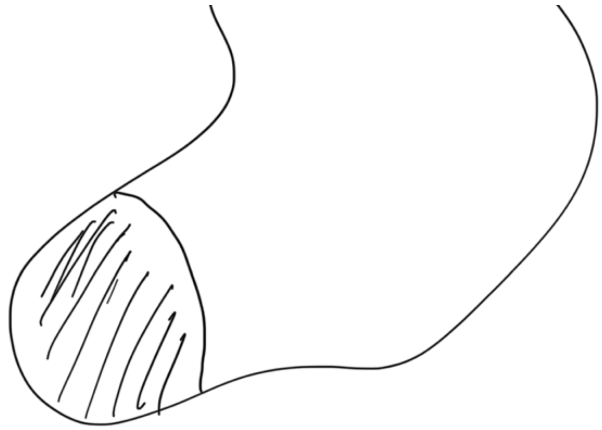
$(b) \rightarrow \quad y = \underline{m}x + \underline{c}$

$(c) \rightarrow \quad y = \underline{a}x^3 + \underline{b}x^2 + \underline{c}x + \underline{d}$

$(d) \rightarrow \quad y = \underline{a}x^{10} + \underline{b}x^9 + \text{---} + \text{—}$

<u>Linear Regression</u>

$y = f(\mathbf{x})$



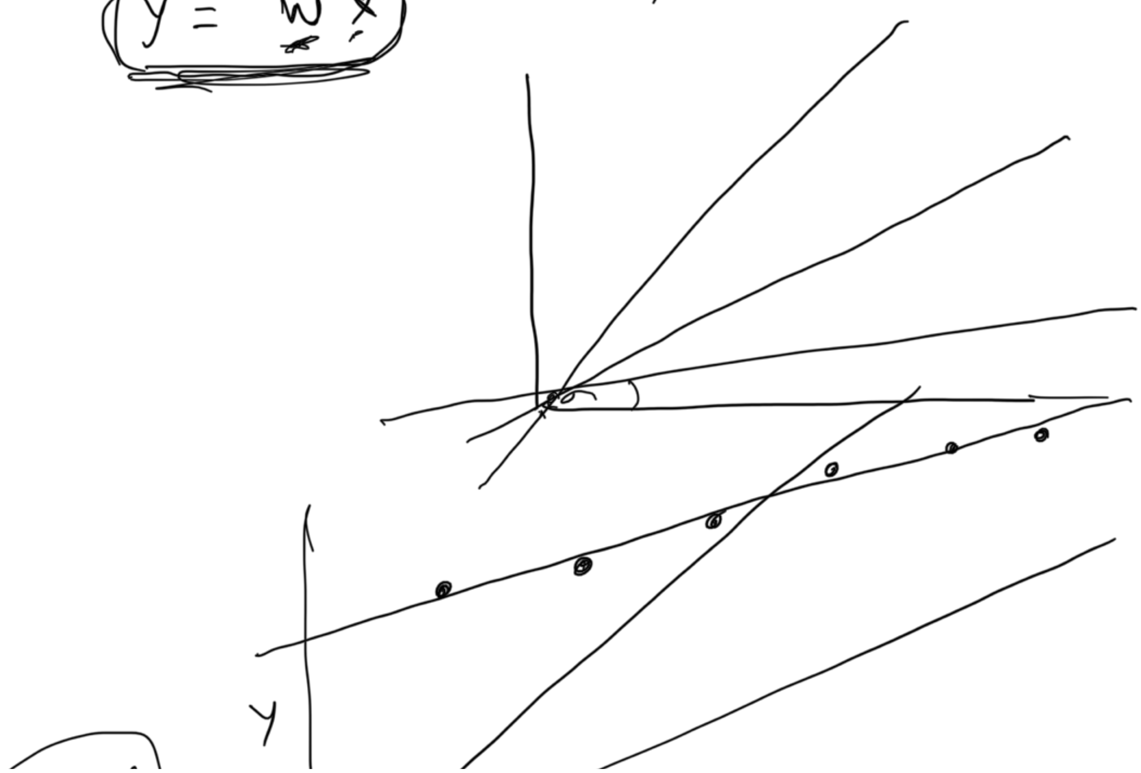
$y = \underline{\mathbf{w}^T \mathbf{x}}$ → $D \times 1$

→ weight vector ($D \times 1$)

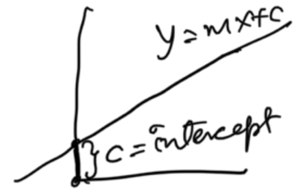
$= w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$

For the flu example

$\boxed{y = w x}$



$\underline{y = wx}$



$$\boxed{y = \underset{\text{bias-term}}{w_0} + w_1 x}$$

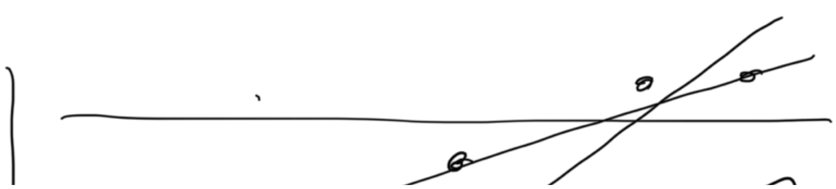
Given some data:

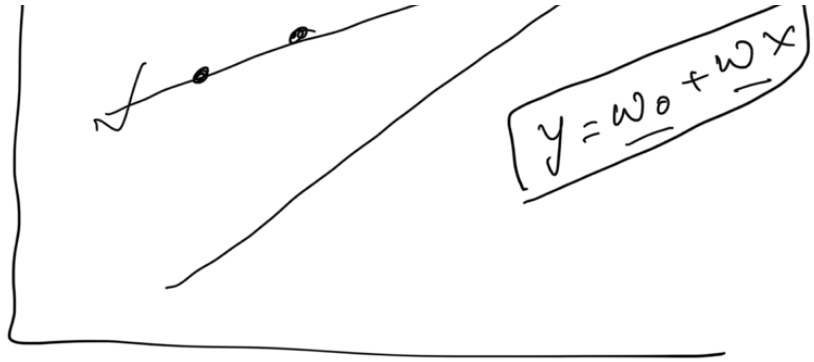
x	y
12	9
28	15
15	11
48	21
56	22

Find the best w_0, w_1 which "fits" the data best.

Wed, Feb 10

Resources on Piazza





Find w_0, w that does what?

X	Y	Pred. $\bar{y}$
x_1	y_1	$\bar{y}_1$
x_2	y_2	$\bar{y}_2$
x_3	y_3	$\bar{y}_3$
$\vdots$	$\vdots$	$\vdots$
x_N	y_N	$\bar{y}_N$

For a given w, w_0

$$\bar{y}_1 = w_0 + w x_1$$

$$\vdots$$

$$\bar{y}_i = w_0 + w x_i$$

Error: $e_i = y_i - \bar{y}_i$

$$J = 1 \sum^{N} e_i^2$$

$2 \quad i=1$

$$\boxed{J(w_0, w) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - (w_0 + w x_i)\right)^2}$$

Squared loss function

$\frac{1}{2} \rightarrow$ just for mathematical convenience

---

what if x is not 1-D

$\Rightarrow \quad x \in \mathbb{R}^d$ where $d > 1$

$y = w_0 + \mathbf{w}^T x$

$\mathbf{w} \in \mathbb{R}^d$

Squared loss function

$$\boxed{J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T x_i\right)^2}$$

absorbed

added a

w_0 into $\mathbf{w}$ 1 to x_i

Find $\mathbf{w}$ that minimizes $J(\mathbf{w})$

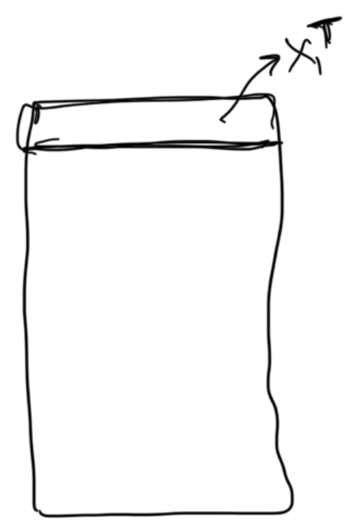
$$\mathbf{w}^T \mathbf{x}_i = \sum_{j=0}^{d} (w_j \, x_{ij})$$

$\mathbf{w} \cdot \mathbf{x}_i$

$\mathbf{w}^T \mathbf{x}_i$

$\mathbf{x}_1$	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$
x_N	y_N

$\mathbf{y}$ is a vector $(N \times 1)$



$\mathbf{X}$ is a matrix $(N \times (d+1))$ $\mathbf{X}$

Training data: $\mathbf{X}, \mathbf{y}$ — target vector

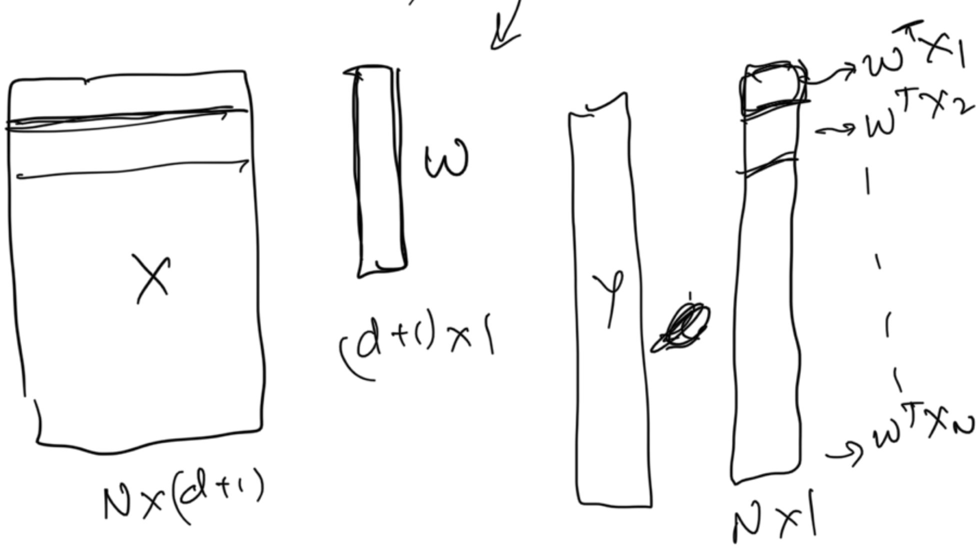
$(N \times (d+1))$ $(N \times 1)$

datamatrix

$$J(w) = \frac{1}{2}\sum_{i=1}^{N} (y_i - w^T x_i)^2$$

$$\begin{array}{cc} x_1 & y_1 - w^T x_1 \\ x_2 & y_2 - w^T x_2 \\ \vdots & \vdots \\ x_N & y_N - w^T x_N \end{array} \qquad e = \begin{bmatrix} y_1 - w^T x_1 \\ y_2 - w^T x_2 \\ \vdots \\ y_N - w^T x_N \end{bmatrix}$$

$$e = (y - Xw)$$



e

$$\sum (y_i - w^T x_i)^2 = (y - Xw)^T (y - Xw)$$

$$e_i = y_i - w^T x_i$$

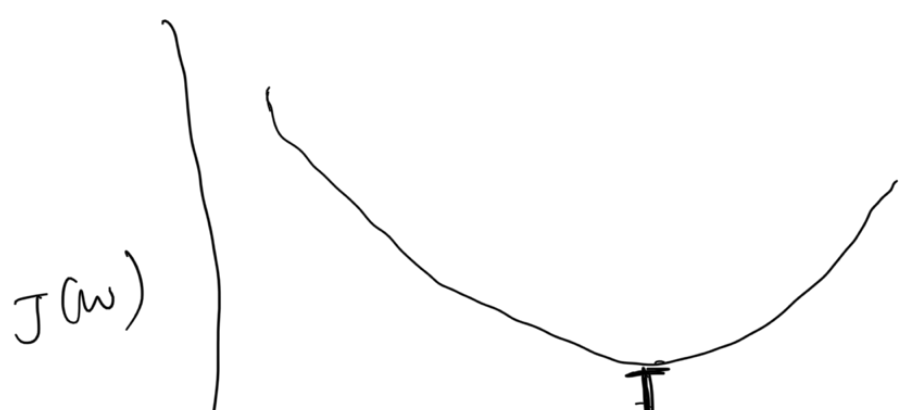
$$\mathbf{e} = \begin{bmatrix} y_1 - w^T x_1 \\ \vdots \\ y_N - w^T x_N \end{bmatrix}$$

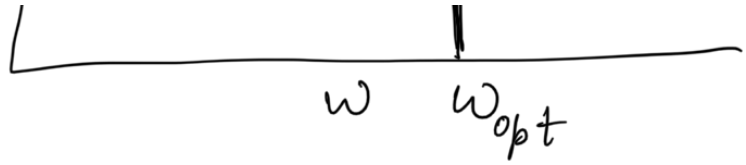
$$\boxed{\begin{array}{l} \text{any vector } \mathbf{p} \\ \mathbf{p}^T \mathbf{p} = \sum p_i^2 \end{array}}$$

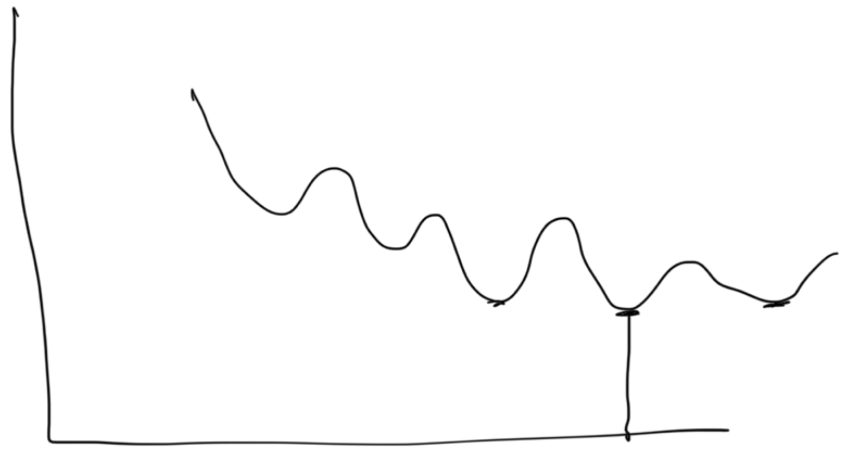
$$\sum (y_i - w^T x_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

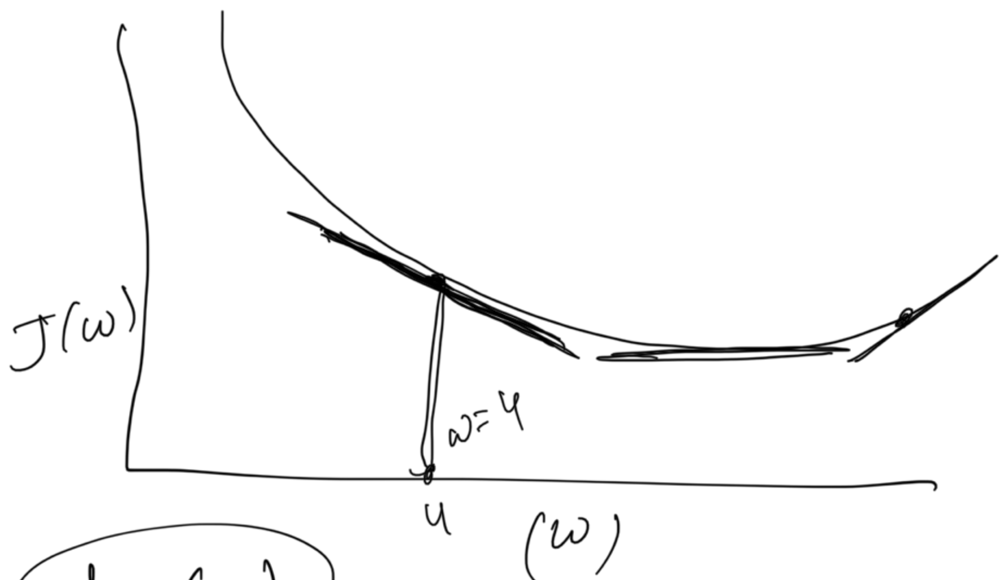
$$\boxed{J(w) = \frac{1}{2} (y - Xw)^T (y - Xw)}$$

How do we find $\mathbf{w}$ that minimizes $J(\mathbf{w})$









$$\frac{d}{dw} J(w)$$

e.g. $J(w) = 3w^3 + 4w$

$$\frac{dJ(w)}{dw} = 9w^2 + 4$$

at $w = 4$

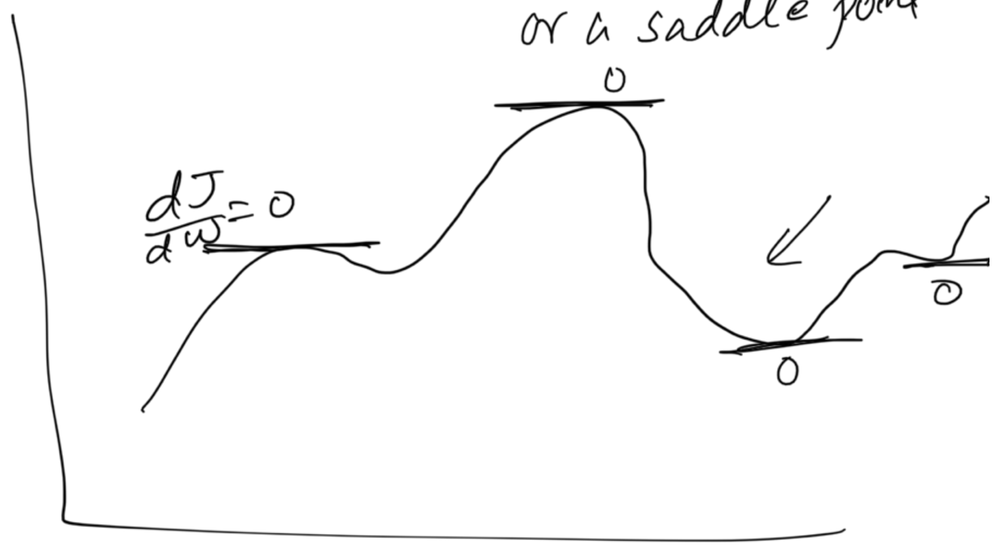
$$\frac{dJ(w)}{dw} = \underline{148}$$

If $\frac{d}{dw} J(w) = 0$ at a given w

that means $w \to$ point of minima

or maxima

or a saddle point



If $J(w)$ is convex

then $\boxed{\frac{d}{dw} J(w) = 0}$

solution for this will give us w for which $J(w)$ is minima

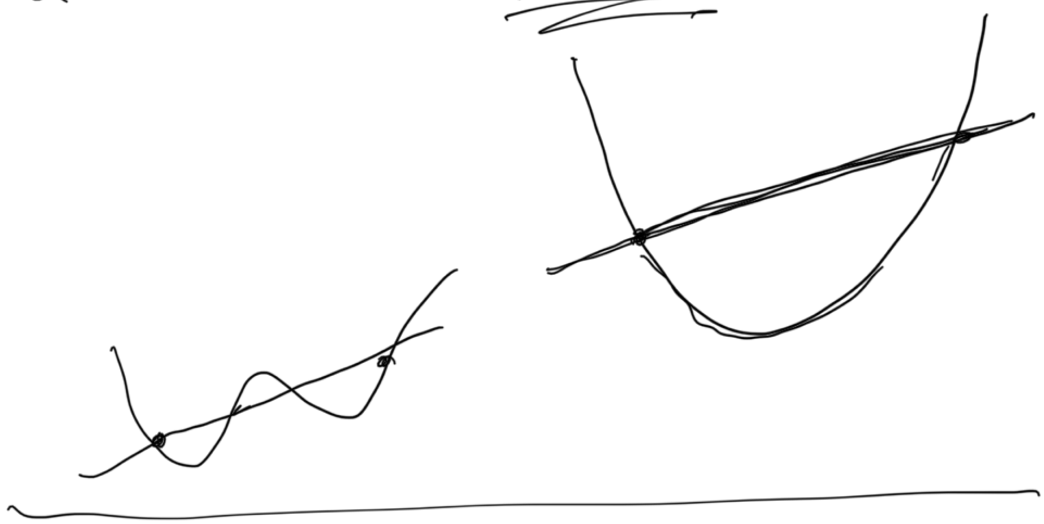
Feb 12

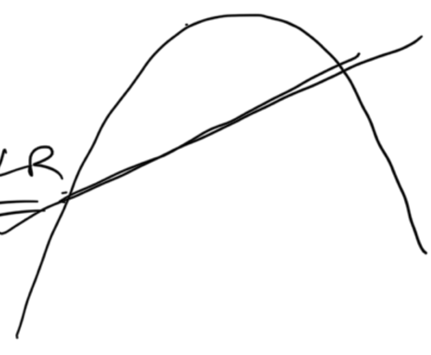
$$f(w) = 7w^3 - 4w^2 + 8$$

scalar

$$\frac{d}{dw} f(w) = f'(w) = 21w^2 - 8w$$

$$\frac{d^2}{dw^2} f(w) = \frac{d}{dw}\left(\frac{df(w)}{dw}\right)$$





Squared loss function for LR is convex

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

Output

$1 \div D$	$1 \div D$
$1 \div D$	$n \div D$
$(d+1) \div D$	$(d+1) \div D$
$1 \div D$	$n \div D$

Input ($\mathbf{w}$)

$$\frac{1}{2}\sum (y_i - w^T x_i)^2$$

$$\frac{d}{d\mathbf{w}} J(\mathbf{w})$$

$$= \nabla J(w)$$

gradient

$$f(\mathbf{w}) = 3w_0^2 + 17w_1 + 8w_0 + 9$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\frac{\partial}{\partial w_0} f(\mathbf{w}) = 6w_0 + 8$$

$$\frac{\partial}{\partial w_1} f(w) = 17$$

$$\frac{d}{d\mathbf{w}} f(w) = \begin{bmatrix} \frac{\partial}{\partial w_0} f(\mathbf{w}) \\ \frac{\partial}{\partial w_1} f(w) \end{bmatrix}$$

Gradient

$$\nabla f(\mathbf{w})$$

$$\frac{d^2}{dw^2} f(w)$$

$$\frac{d}{dw} \nabla f(w)$$

$$f(\mathbf{w}) = \begin{bmatrix} 7w_0^2 + 8w_1 \\ w_0 w_1 + 9w_1^3 \end{bmatrix}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$= H$$

Hessian of $f(w)$

$$\frac{d}{dw} f(w) = \begin{bmatrix} 14w_0 & 8 \\ w_1 & w_0 + 27w_1^2 \end{bmatrix}$$

Jacobian

$J(w)$

Calculate $\nabla J(w) \equiv \frac{d}{dw} J(w)$

and then solve for w: $\boxed{\nabla J(w) = 0}$

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

$$J(w) = \frac{1}{2}\left[\left(y^T - (Xw)^T\right)(y - Xw)\right]$$

$$= \frac{1}{2}\left[\left(y^T - w^T X^T\right)(y - Xw)\right]$$

$$= \frac{1}{2}\left[(y^T y) - y^T X w - w^T X^T y + w^T X^T X w\right]$$

$$= \frac{1}{2}\left[y^T y - 2 w^T X^T y \right.$$

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

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

$a^T b = b^T a$

$y^T X w$

$$+ w^T X^T X w \Big]$$

$y^T z$

$\underline{z^T y}$

$$\frac{d}{dw} J(w) = \frac{1}{2}\Big[\frac{d}{dw} \underline{y^T y} - 2 \frac{d}{dw} w^T X^T y$$

$$+ \frac{d}{dw} \underline{w^T X^T X w} \Big]$$

$$= \frac{1}{\cancel{2}}\Big[-\cancel{2} X^T y + 2 X^T X w\Big]$$

$$= -X^T y + X^T X w$$

Set $\frac{d}{dw} J(w) = 0$ $\quad \Big[\nabla f(x) = 0\Big]$

$$-X^T y + X^T X w = 0$$

$$X^T X w = X^T y$$

$$(X^T X)^{-1} (X^T X) w = (X^T X)^{-1} X^T y$$

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

$A^{-1} A = I$

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$X - N \times (d+1)$

$y - N \times 1$

$X^T X = (d+1) \times (d+1)$

Scalability $\quad A \quad A^{-1} \quad O(d^3)$

$d \times d$

$\boxed{\frac{1}{x} \to x = 0}$

$\boxed{A^{-1}} \to$ is not easy to calculate

$W = (X^T X)^{-1} X^T Y$