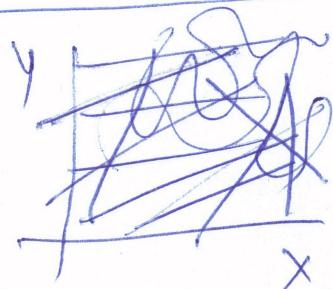


$$x_1 = 10 \quad y_1 = 7$$

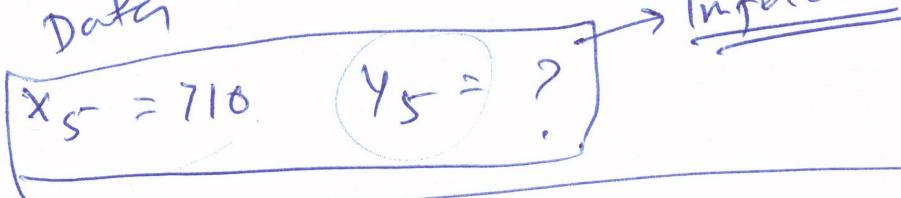
$$x_2 = 49 \quad y_2 = 41$$

$$x_3 = 3 \quad y_3 = 1$$

$$x_4 = 650 \quad y_4 = 304$$



Training Data



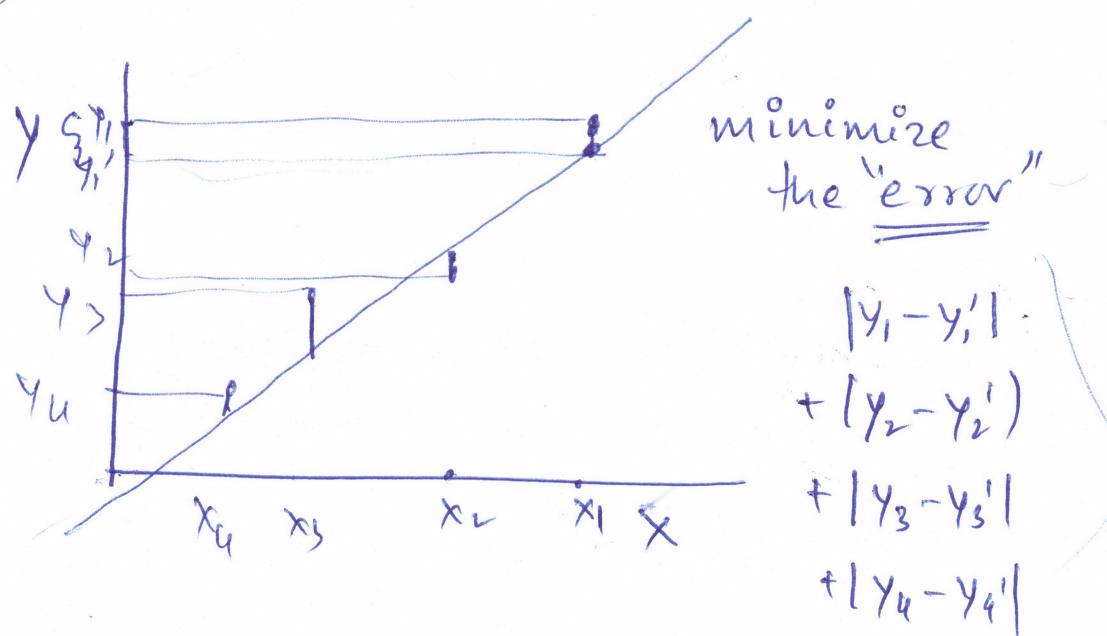
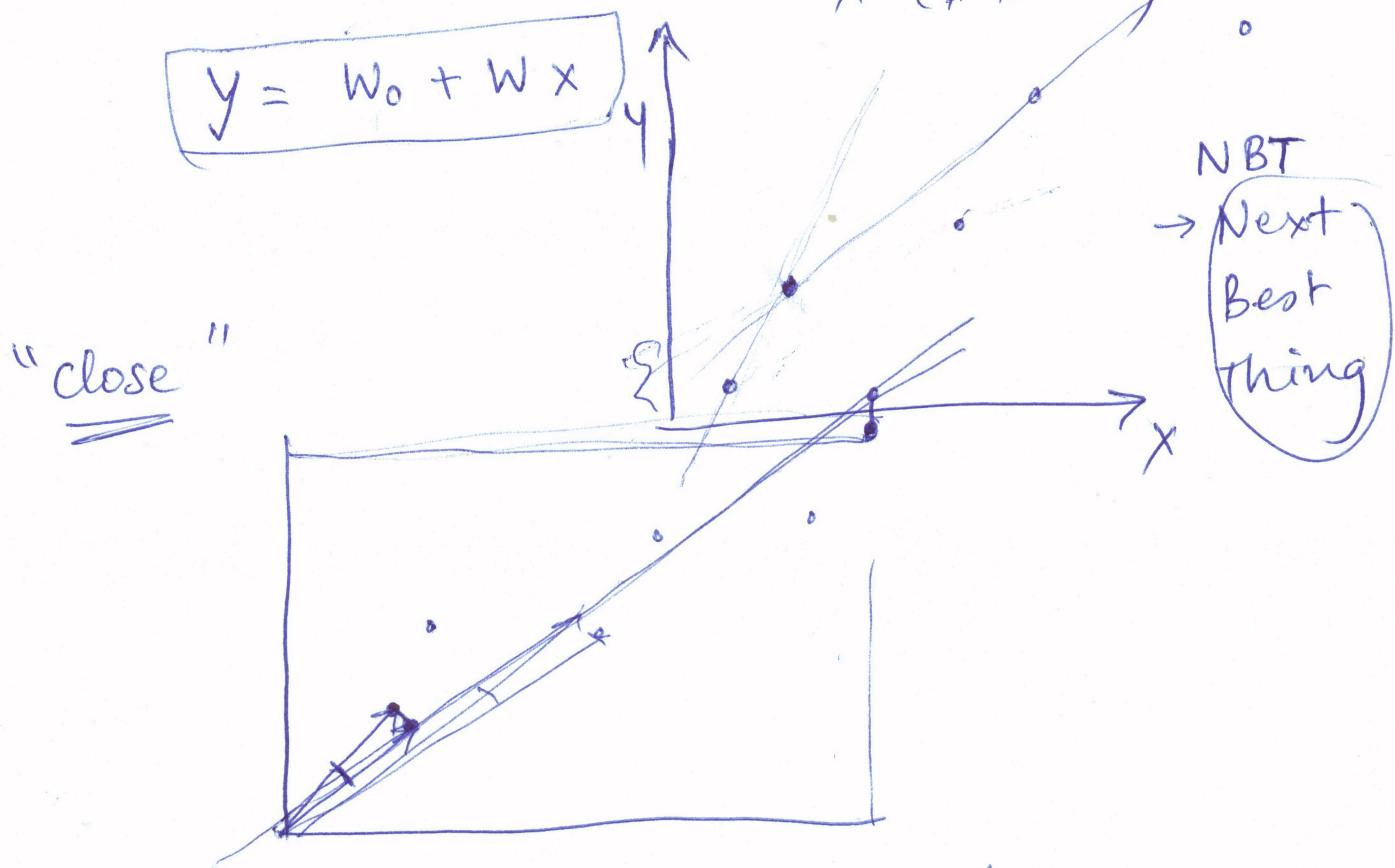
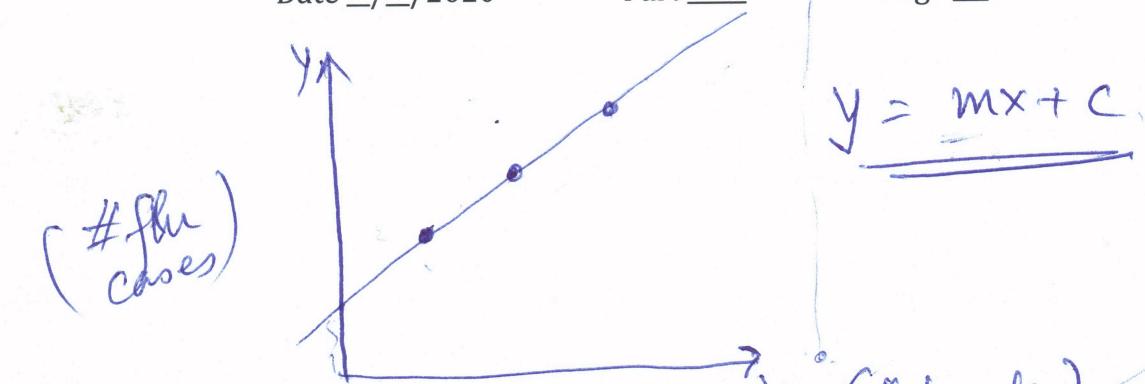
INDUCTIVE BIAS — Reduce the hypothesis space

No INDUCTIVE BIAS



linear Models

We assume that f is a straight line



$$x_1 \quad y_1 \quad y'_1 = w_0 + w x_1$$

$y = w_0 + w X$

$$x_2 \quad y_2 \quad y'_2 = w_0 + w x_2$$

$$x_3 \quad y_3 \quad y'_3 = w_0 + w x_3$$

$$x_4 \quad y_4 \quad y'_4 = w_0 + w x_4$$

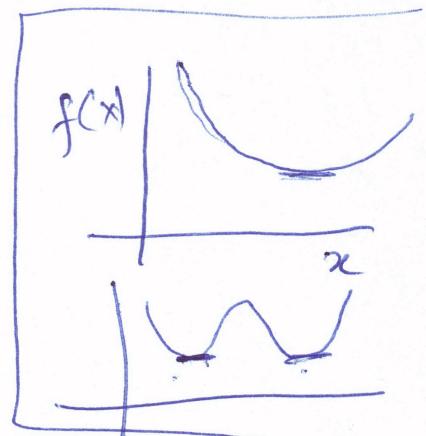
Parameters
of the model

$$|(w_0 + w x_i) - y_i| + |(w_0 + w x_2) - y_2|$$

$$J = \frac{1}{N} \sum_{i=1}^N |(w_0 + w x_i) - y_i|$$

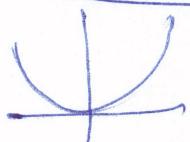
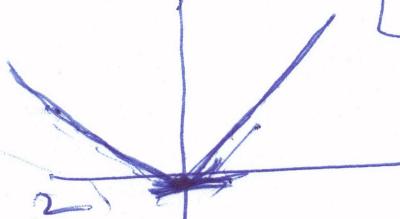
$$J(w_0, w)$$

Min.
Abs.
Dev.



Optimization Problem

$$f(x) = |x|$$



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

← Squared Loss

(linear regression Model)

$$J = \frac{1}{2} \sum (y_i - (w_0 + w x_i))^2 \quad \leftarrow \text{Objective function}$$

$$\begin{array}{c} f(u) \\ \hline \frac{d}{du} f(u) \\ f'(u) \end{array}$$

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

$$\frac{\partial J}{\partial w_0} = 0$$

$$\frac{\partial}{\partial w} \left[\frac{1}{2} \sum (y_i - (w_0 + w x_i))^2 \right]$$

$$\begin{aligned} \frac{\partial}{\partial w} z^2 &= \frac{\partial z^2}{\partial z} \frac{\partial z}{\partial w} \\ &= 2z \cdot 1 \end{aligned}$$

$$= \frac{1}{2} \sum \left[\frac{\partial}{\partial w} (y_i - (w_0 + w x_i))^2 \right]$$

$$= \frac{1}{2} \sum \left[2(y_i - (w_0 + w x_i)) \cdot (-x_i) \right]$$

$$= - \sum (y_i - (w_0 + w x_i)) x_i$$

$$\frac{\partial}{\partial w_0} \left[\frac{1}{2} \sum (y_i - (w_0 + w x_i))^2 \right]$$

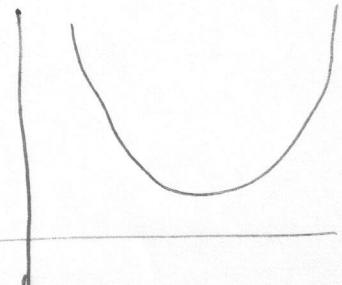
$$= - \sum (y_i - (w_0 + w x_i))$$

for w, w_0 to be the minima points.

$$\left. \begin{array}{l} \frac{\partial J}{\partial w} = 0 \\ \frac{\partial J}{\partial w_0} = 0 \end{array} \right\} \begin{array}{l} -\sum (y_i - (w_0 + w x_i)) x_i = 0 \\ -\sum (y_i - (w_0 + w x_i)) = 0 \end{array}$$

Solve these to get w and w_0

Do at home



What if we have multiple input features.

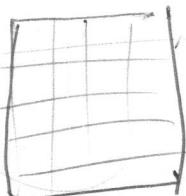
	#tweets	#Commuters in public station		#flu cases
x_1	x_{11}	x_{12}	\dots	x_{1D}
x_L	x_{21}	x_{22}	\dots	x_{2D}
\vdots	\vdots	\vdots	\vdots	\vdots
x_N	x_{N1}	x_{N2}	\dots	x_{ND}
w_0	w_1	w_2	\dots	w_D

Prediction = $(w_0 + w_1 x_{11} + w_2 x_{12} + \dots + w_D x_{1D})$
at x_1

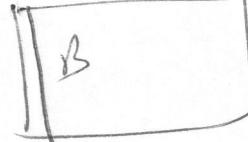
or Prediction at
any x_i

$$= w_0 + \sum_{j=1}^D w_j x_{ij}$$

$$J(w_0, w_1, \dots, w_D) = \frac{1}{2} \sum_{i=1}^N (y_i - (w_0 + \sum_{j=1}^D w_j x_{ij}))^2$$

Matrix $A_{m \times n}$ 
 $\begin{matrix} A & B \\ m \times n & n \times k \end{matrix}$

Transpose

 A^T $n \times m$ Vector m
 a
(m)

dot-product

or inner-product

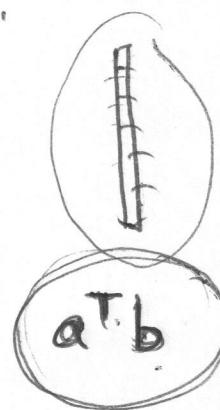
$$\sum_{i=1}^m a_i b_i$$

length of vectors
 a & b

Vector is a 1-column matrix.

 a is a $(m \times 1)$ matrix

A dot-product between

 a and b $(m \times 1)$ b
($m \times 1$) a^T b
 w
→ weight
vector

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_D \end{bmatrix}$$

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix}$$

I can now write the prediction

as

$$\hat{y}_i = (w_0 + w^T x_i)$$

$$w_0 + w^T x_i \equiv w_0 + \sum_{j=1}^D w_j x_{ij}$$

what if $x_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix}$ and $w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$

$$w_0 + w^T x_i \equiv \underline{w^T x_i}$$

Now w is a ~~$(D+1) \times 1$~~ is a $\underline{(D+1) \times 1}$ vector.

Each x_i is also a $\underline{(D+1) \times 1}$ vector.

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

$$f(x) \sim x^2$$

Training data is is a $\underline{(D+1 \times 1)}$ matrix

x_1

x_2

x_3

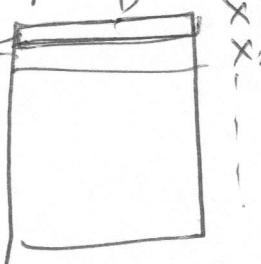
x_4

\vdots

x_N

X

D



→ data matrix.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$\frac{d}{dx} f(x) \sim 2x$$

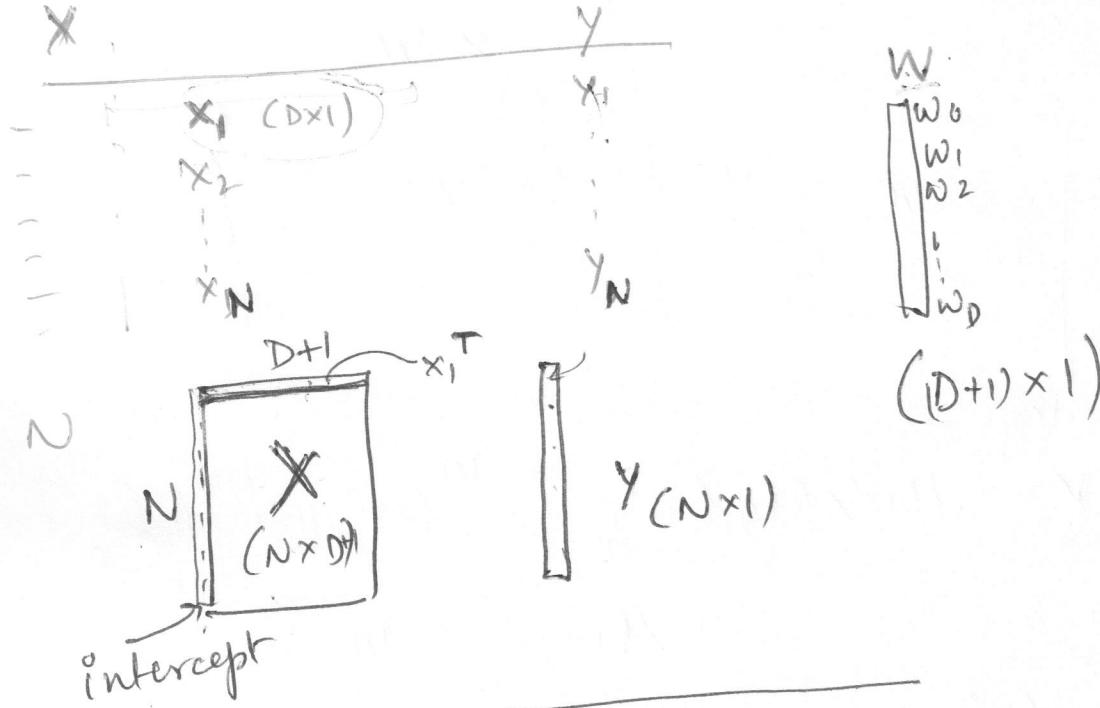
$$f(x) \sim \frac{1}{2}x^2$$

$$\frac{d}{dx} f(x) \sim x$$

$$Y_{(N \times 1)}$$

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

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



$$\begin{aligned} a^T b &= b^T a \\ a^T a &= \sum_{j=1}^D a_j^2 \\ (AB)^T &= B^T A^T \end{aligned}$$

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

$$Xw =$$

$$\begin{bmatrix} w^T x_1 \\ w^T x_2 \\ \vdots \\ w^T x_N \end{bmatrix} \xrightarrow{x_i^T w} \text{Prediction at } x_i$$

$$y - Xw =$$

$$\begin{bmatrix} y_1 - w^T x_1 \\ y_2 - w^T x_2 \\ \vdots \\ y_N - w^T x_N \end{bmatrix}$$

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

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

Find w that minimizes $J(w)$

$$\begin{aligned} (A+B)^T &= A^T + B^T \\ &\Rightarrow \end{aligned}$$

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

$$= \frac{1}{2} [y^T y - \cancel{y^T Xw} - \cancel{(Xw)^T y} + (Xw)^T Xw]$$

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

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

$$\nabla J = \frac{\partial}{\partial w} \frac{1}{2} [y^T y - 2w^T (y^T X) + w^T X^T X w]$$

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

$$\nabla J = 0$$

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

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

$$(X^T X)w = X^T y$$

Multiplying both sides by $(X^T X)^{-1}$

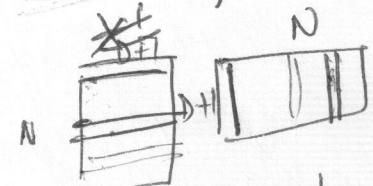
$$w = (X^T X)^{-1} X^T y$$

$$\begin{aligned} f(u, v) &= 7u^2 + 3v^3 \\ \frac{\partial f(u, v)}{\partial u} &+ 8u + 9 \\ \frac{\partial f(u, v)}{\partial v} & \end{aligned}$$

gradient with respect to (u, v)

$$\begin{aligned} \frac{d}{da}(a^T b) &= b \\ \frac{d}{da}(a^T M a) & \\ &= (M + M^T) a \\ &= 2Ma \quad [\text{if } M \text{ is symmetric}] \end{aligned}$$

$X^T X$ is symmetric



$$\alpha \approx \frac{1}{a}$$

$$\begin{aligned} x^2 - 7x \\ x(x-7) \end{aligned}$$

What is the complexity of this algorithm.

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

~~$D \times 1 \times D$~~

$(D \times 1)$

$O(D^3)$ \rightarrow # features.

$(X^T X)^{-1}$

$O(N) A^{-1} u$

A^{-1}

~~b~~

$\frac{1}{b}$

$$J(w) = (y - Xw)^T (y - Xw)$$

$$\nabla J(w) = -X^T y + X^T X w$$

Start at some w

while not converged:

Calculate $\nabla J(w)$

$$w = \underline{w} - \eta \nabla J(w)$$

~~check~~

$J(w) \rightarrow$ calculate

~~subtract~~ from $J(w_{\text{previous}})$

$$w = \underline{(X^T X)^{-1} X^T y} \rightarrow O(D^3)$$

~~If~~ $D^3 > N$