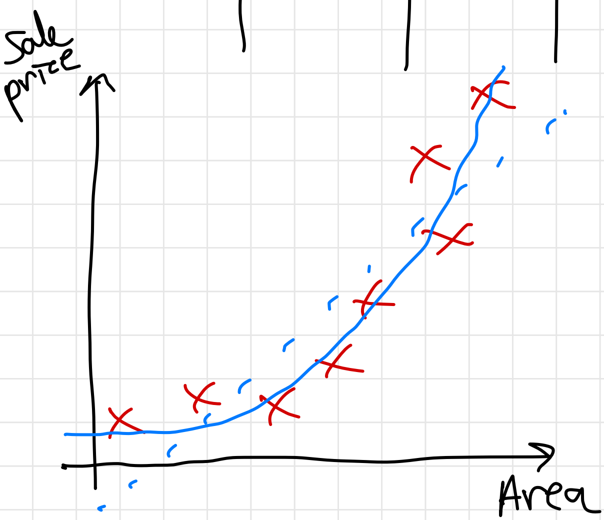
Linear Regression II (1/17/2023)

1. Features
2. Convexity
3. Closed-form solution

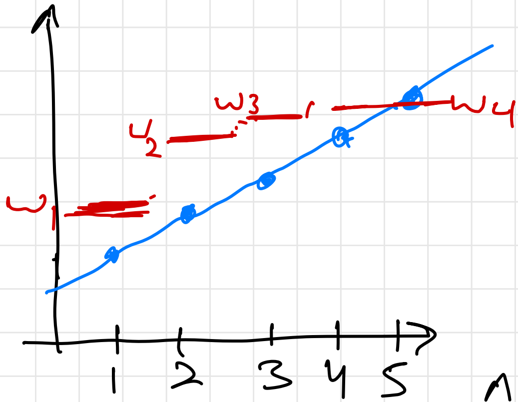
Featurization

sale price (y)	Area	#bed	house type	$Area^2$	$Area^3$
	500	1	condo	250000	
	1000	2	townhouse	1,000,000	
				⋮	



$$\hat{y} = w_1 \text{Area} + w_2 \text{Area}^2 + w_3 \text{Area}^3 + \cdots$$

Linear regression is linear in the features



indicator features

y	#bed=1	bed=2	=3	≥4
	1	0	0	0
	0	1	0	0

$$\boxed{\begin{aligned}\mathbb{1}(\text{true}) &= 1\\ \mathbb{1}(\text{false}) &= 0\end{aligned}}$$

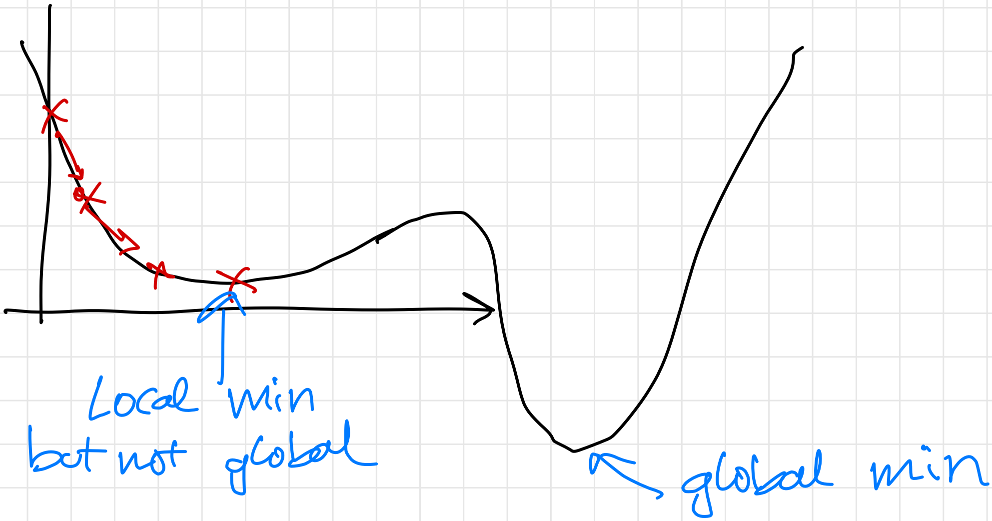
$$\begin{aligned}\hat{y} = \; & w_1 \mathbb{1}[\#\text{bed}=1] \\ & + w_2 \mathbb{1}[\#\text{bed}=2] \\ & + w_3 \mathbb{1}[\#\text{bed}=3] \\ & + w_4 \mathbb{1}[\#\text{bed} \geq 4]\end{aligned}$$

zip code:
naively 10^5 new features

"Feature engineering": art of choosing features to use

→ $\mathbb{1}[\text{zip code is in CA}]$
allows sharing info between nearby zipcodes

Convexity why does gradient descent work?



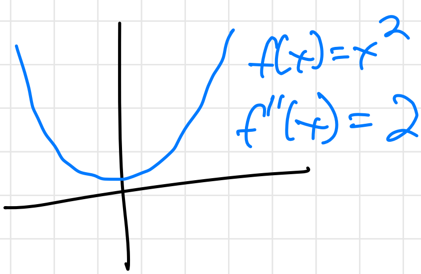
① Linear regression is a convex problem

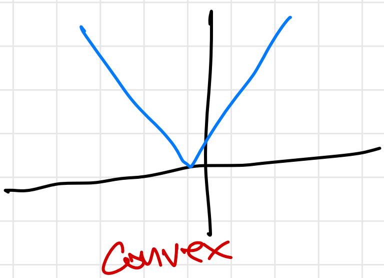
$\boxed{L(w) \text{ is a}}$ $\boxed{\text{convex function}}$ ✓

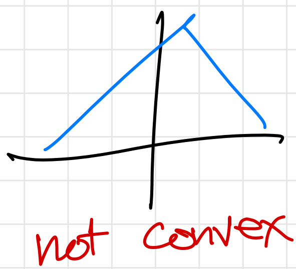
② For a convex function,

☆ any local minimum is a global minimum

Def 1: $f(x)$ is convex $\Longleftrightarrow f''(x) \geq 0$



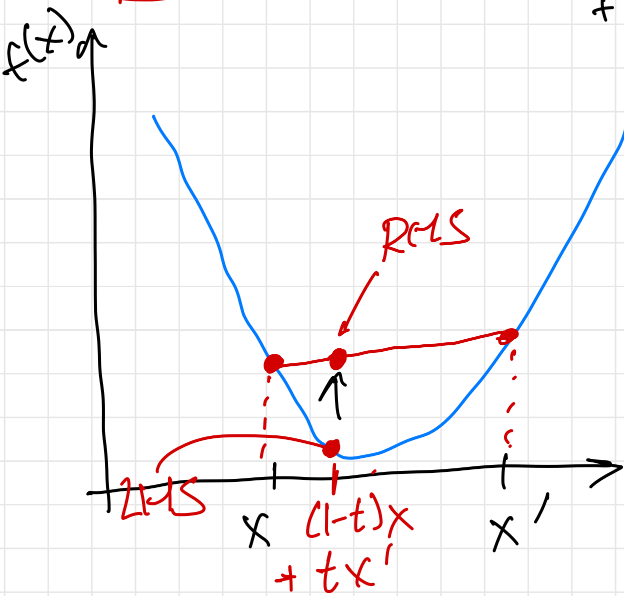




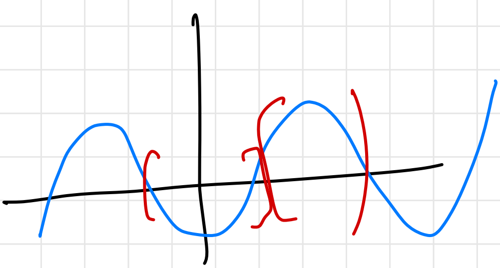
"Def" 2: Convex function "holds water"

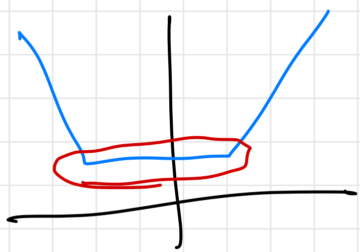
Def 3 (Actual): A function f is convex iff for all x, x' in domain of f and all $t \in [0, 1]$:

$$f((1-t)x + tx') \le (1-t)f(x) + t f(x')$$



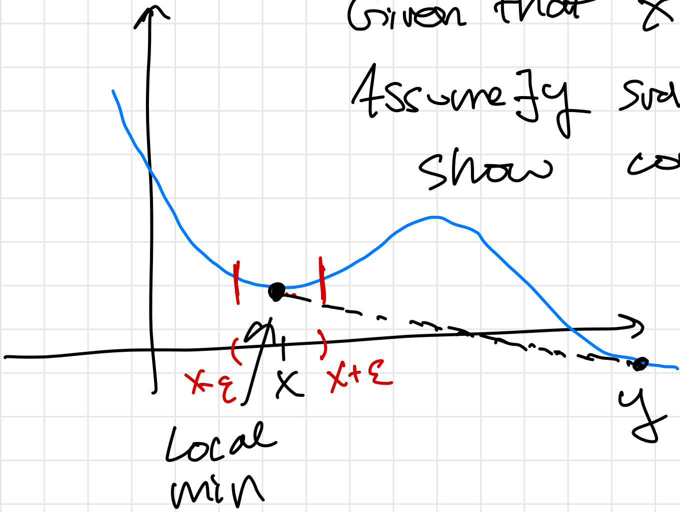
TLDR:
If you draw a line connecting $(x, f(x))$ and $(x', f(x'))$, it must lie above the function itself





Given that x is local min of f,
Assume $\exists y$ such that $f(y) < f(x)$
show contradiction

f and convex



y has lower function value

x is a local min of f iff
$\exists\, \varepsilon > 0$ such that $\forall\, z \in \underbrace{B_\varepsilon(x)}$
$f(z) \geq f(x)$

set of points z
where $\|x - z\| \leq \varepsilon$

Choose $t > 0$ such that
$(1-t)x + ty \in B_\varepsilon(x)$

① Because local min, $f((1-t)x+ty) \geq f(x)$

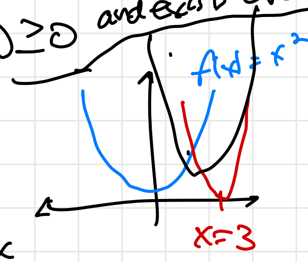
② Because convexity:

$$f((1-t)x+ty) \leq (1-t)f(x) + t f(y)$$
$$< (1-t)f(x) + t f(x)$$
$$= f(x)$$

contradiction

$$C(w) = \frac{1}{n}\sum_{i=1}^{n}\left(w^T x^{(i)} - y^{(i)}\right)^2$$

① If $f:\mathbb{R}\to\mathbb{R}$ and $f''(x)\geq 0$ and exists everywhere
then f is convex



$f(x) = (-2x+6)^2$

② If f is convex, then
$g(x) = f(Ax+b)$ is convex

③ If $f(x)$ and $g(x)$ are convex,
So is $f(x)+g(x)$

① $f(x)=x^2$ is convex (by ①)

② $(w^T x^{(i)} - y^{(i)})^2$ is convex function of w (by ②)

③ $\sum_{i=1}^{n}(w^T x^{(i)} - y^{(i)})^2$ is convex by ③

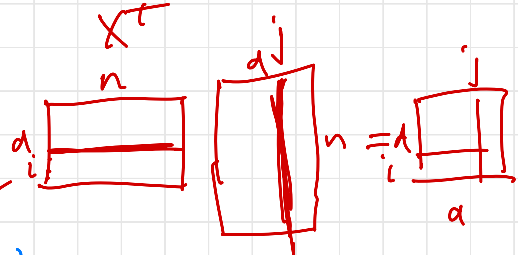
Closed form for Linear Regression ("Normal Equations")

$$\nabla_w L(w) = \cancel{\frac{1}{n}} \sum_{i=1}^{n} \cancel{2}\,(w^T x^{(i)} - y^{(i)}) \cdot x^{(i)} = 0$$

$$\boxed{\sum_{i=1}^{n} (w^T x^{(i)})\, x^{(i)}} = \boxed{\sum_{i=1}^{n} x^{(i)} y^{(i)}} = X^T y$$

X = matrix where each row is $x^{(i)}$

$$\begin{array}{cc} & \\ n & \begin{bmatrix} \longmapsto x^{(1)} \longrightarrow \\ \vdots \\ \longmapsto x^{(n)} \longrightarrow \end{bmatrix} \\ & \times \quad d \end{array}$$



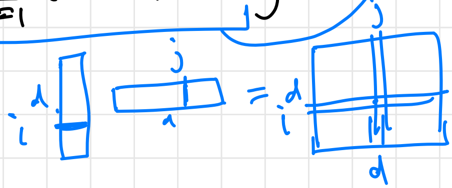
y = vector of $y^{(i)}$'s

claim: this equals $\boxed{X^T X}\, w$

$$= \sum_{k=1}^{n} x^{(k)} (x^{(k)T} w)$$

$$\sum_{k=1}^{n} (x^{(k)} x^{(k)T})\, w$$

$$= \left(\sum_{k=1}^{n} (x^{(k)} x^{(k)T}) \right) w$$



i-jth entry

$$\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)}$$

$||$

i-jth entry?

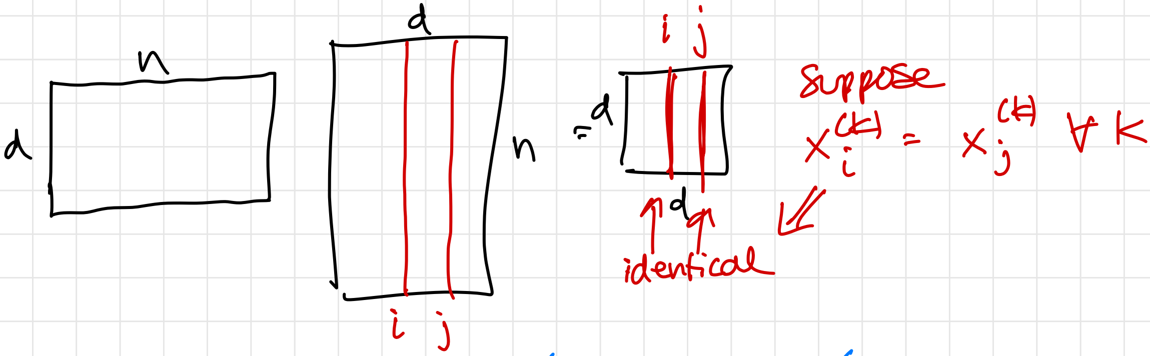
$$\sum_{k=1}^{n} x_i^{(k)} \cdot x_j^{(k)}$$

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

"Normal Equation"

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

Question: What if X^TX is not invertible?
"When would"



$$\hat{y} = w_1 x_1 + \cdots \underset{w_i - 100}{\cancel{w_i}} x_i + \cdots + \underset{w_j + 100}{\cancel{w_j}} x_j + \cdots$$

Result: answer is not unique anymore!

If i & j th cols almost identical:
leads to instability

In practice

① Pseudoinverse A^+
- $A^+ = A^{-1}$ when inverse exists
- For Normal Equations, A^+ always gives you an optimal solution

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

② Avoid highly correlated features

---

① Feature engineering: Learn non-linear functions of original features w/ linear regression

② Convexity $\Rightarrow$ local min is sufficient

③ Normal Equations - Closed Form