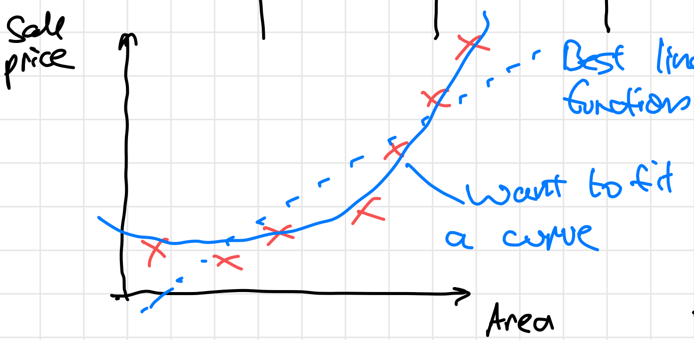
Linear Regression II (8/29/2023)

1. ~~①~~ How do we learn more complex functions?
2. ~~②~~ Why does gradient descent work for linear reg?
3. ③ Why use squared error?

real number → Area; integer → #bed; categorical → house type

(y) Sale price	Area	#bed	house type	$Area^2$	$Area^3$	...
500k	1200	2	Condo	1440000	~	



"predicted"

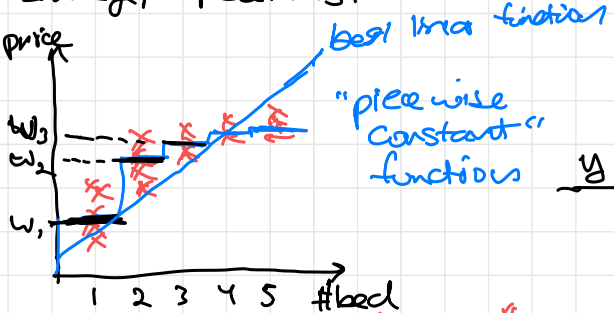
want to learn:

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

Solution:
Just add more features!

Linear regression is linear in the input features (that we can choose)

Integer Features?



"indicator function"
$\mathbb{1}[true] = 1$, $\mathbb{1}[false] = 0$

Add "indicator" features
some boolean function

y	#bed=1	#bed=2	#bed3
	1	0	0
	0	1	0
	0	1	0
	0	0	0
	0	0	1

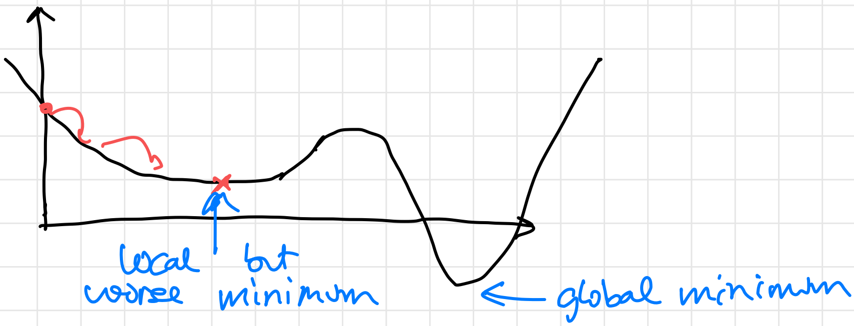
Now: $\hat{y} = w_1 \cdot \boxed{\mathbb{1}}[\#bed=1] + w_2 \cdot \mathbb{1}[\#bed=2] + \cdots$

parameters

"Feature Engineering": Process of choosing what features to use

Categorical features: add $\mathbb{1}[\text{is\_condo}]$, $\mathbb{1}[\text{is\_townhouse}]$

Convexity (why does gradient descent work?)

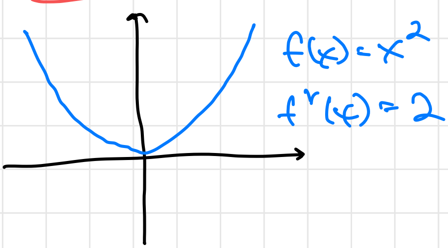


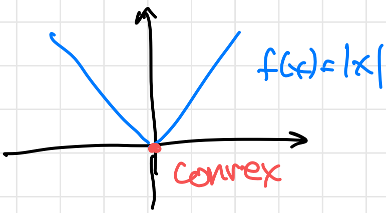
① Linear regression loss function $L(w)$ is convex

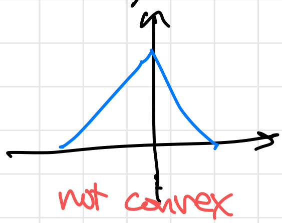
② For any convex function, all local minima are global minima

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

assumes f'' exists everywhere



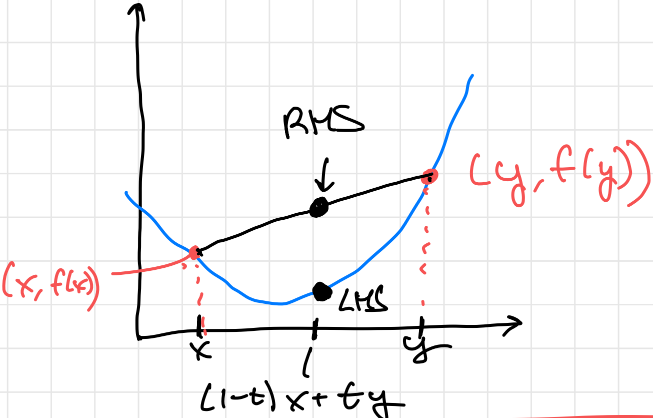




Def 2 (informal): Convex function "holds water"

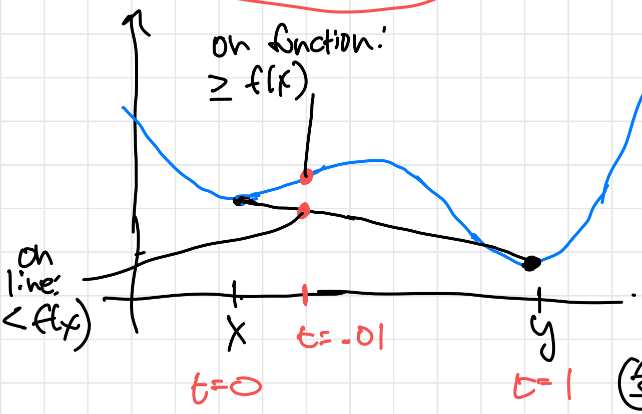
Def 3 (formal): A function f is convex iff for every x, y in its domain and every $t \in [0,1]$

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



TL;DR:
If you draw line connecting $(x, f(x))$ & $(y, f(y))$,
it must be above the function

① All local minima of convex function are global minima



① because x is local min, there's some small t such that

$$f((1-t)x+ty) \geq f(x)$$

② Line from $(x, f(x))$ slopes downward
③ So $f((1-t)x+ty))$ must be $\boxed{\text{above}}$ the line between x & y
④ Hence, f is not convex.

② linear regression is convex

$$\rightarrow L(w) = \frac{1}{n}\sum_{i=1}^{n}(w^T x^{(i)} - y^{(i)})^2$$

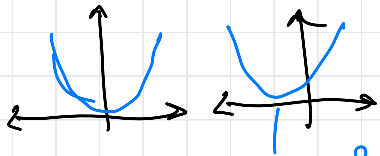
Rules for convexity:

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

then f is convex

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

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



$y = (x+1)^2$

④ If $f(x)$ is convex, and $c > 0$
$c\,f(x)$ is convex

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^{\boxed{2}}$$

← why not 4?
absolute value?

① $f(x) = x^2$ is convex by ①

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

parameter (w)   constants ($x^{(i)}$, $y^{(i)}$)

③ $\frac{1}{n}\sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^2$ is convex by ③ & ④

Why Square? <u>Maximum Likelihood Estimation</u>

- → posit probabilistic process that generated data
- → choose parameters make observed data most likely?

E.g. coin flips
observe $[H, T, H, H, H]$ ← observed data
unknown p = prob. of heads ← parameter
Goal: Choose p that makes data most likely ← "learning"

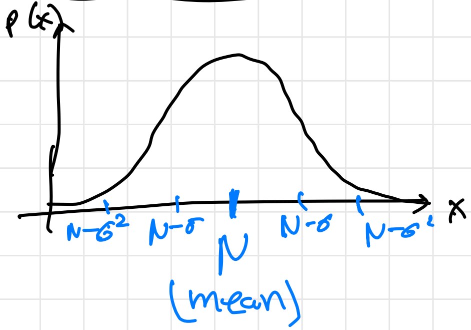
Linear Regression: Assume $y^{(i)}$ drawn from Gaussian w/ mean $w^T x^{(i)}$ & variance σ^2

determined by "true" value of parameter w

independently

constant

Recall Gaussian



$$p(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^{\boxed{2}}}{2\sigma^2}\right)$$

Likelihood of data (probability of data as a function of w)

$$\mathcal{L}(w) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; w)$$

"parameterized by"

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}\right)$$

Trick: take the log (monotonically increasing)

$$\log \mathcal{L}(w) = \sum_{i=1}^{n} \underbrace{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}_{} + -\frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma^2}$$

$$= \text{Constant} + \boxed{-\frac{1}{2\sigma^2}} \sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)})^{\boxed{2}}$$

Maximizing $\log \mathcal{L}(w)$ equivalent to minimizing old $L(w)$