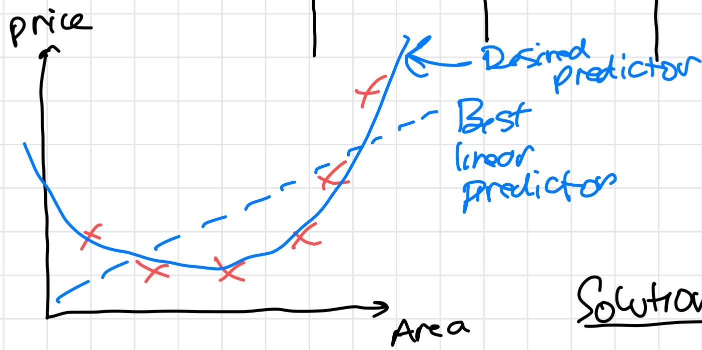
1/16/2024: Linear Regression II

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

(y) Sale price	Area (Real number)	#bed (Integer)	house type (categorical)	$Area^2$ (New feature)	$Area^3$ (New feature)
500k	1200	2	Condo	144,000	~

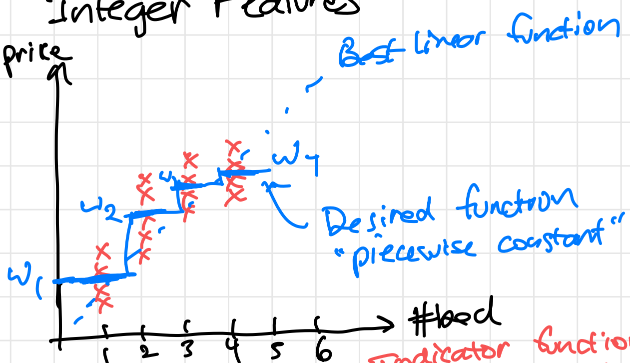


prediction $= w_1 \cdot Area + w_2 \cdot Area^2 + w_3 \cdot Area^3$

Solution: Add more features

Linear regression is linear in the input features
(which we control)

Integer Features



Solution:
Add indicator features!
some boolean function

y	#bed = 1?	#bed = 2?	#bed = 3 ...
500k	1	0	0
700k	0	1	0
⋮	0	1	0
	0	0	1

Indicator function: $\mathbb{1}[true] = 1$, $\mathbb{1}[False] = 0$

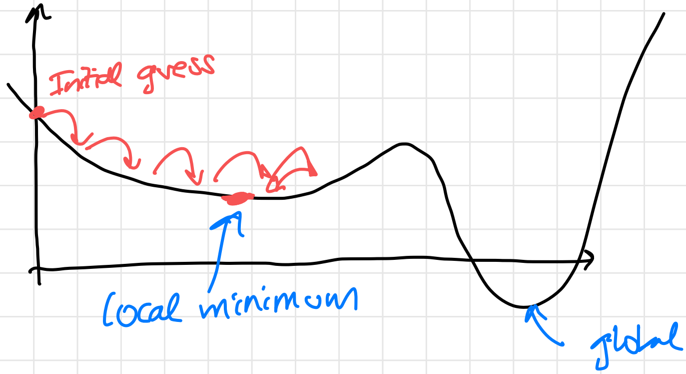
Now:
prediction $= w_1 \cdot \mathbb{1}[\#bed = 1] + w_2 \cdot \mathbb{1}[\#bed = 2] + \ldots$

For categorical Features, use same idea, e.g. $\mathbb{1}[\text{house type} = \text{"condo"}]$

"Feature Engineering": Process of choosing features to use

Why does gradient descent work?

Answer: Linear regression is **convex**



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

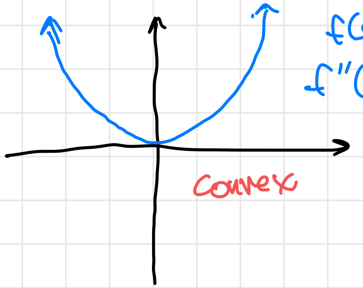
② For all convex functions, every local minimum is a global minimum

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

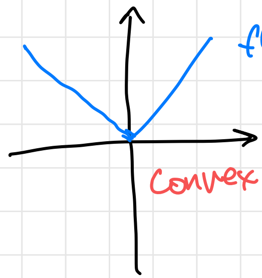
But this only holds when f'' exists everywhere

$f(x) = x^2$

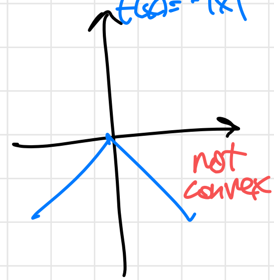
$f''(x) = 2$



$f(x) = |x|$



$f(x) = -|x|$



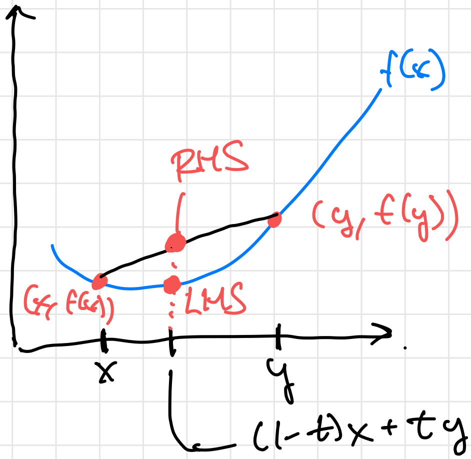
Def 2 (informally): convex functions "hold water"

Def 3 (formal): A function f is convex iff

For every x, y in its domain and every $t \in [0,1]$,

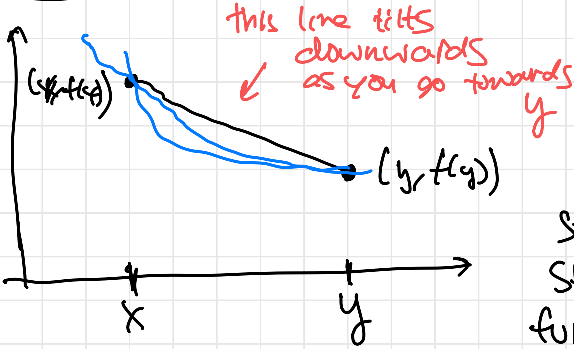
$$f((1-t)x + ty) \leq (1-t) f(x) + t f(y)$$

"go t% of the way from x to y" — "go t% of way from f(x) to f(y)"



TLDR:
If you draw line between $(x, f(x))$ and $(y, f(y))$, the line must be above the function

All local minima of convex functions are global minima



Let y be global min
x be any other point where $f(x) > f(y)$

Starting from x, if you step in direction of y, function must go down.

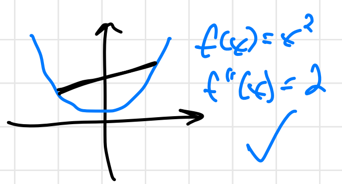
Therefore, x is not local minimum

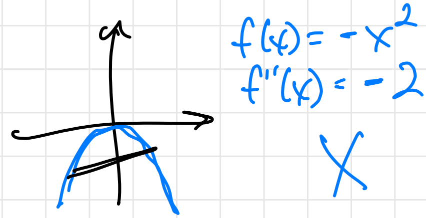
Linear regression is convex

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

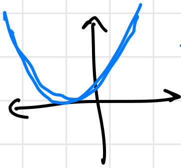
Rules:

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





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



$f(x) = (x+1)^2$

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

(4) If f is convex and
c is a constant ≥ 0
then $c \cdot f(x)$ is convex

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

- $f(x) = x^2$ is convex by (1)
- $(w^T x^{(i)} - y^{(i)})^2$ is convex by (2)

parameter constants

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

Maximum Likelihood Estimation (MLE)

- Posit probabilistic process that generated the data
- Choose parameters to make observed data most likely

E.g. coin flips

Observe data = [H, T, H, H, H] ← observed data

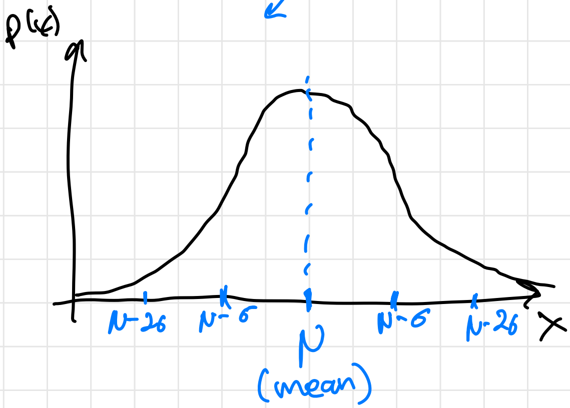
unknown p = probability of heads ← unknown parameter

Goal: choose p that makes data most likely] ← "learning"

Linear Regression: Assume $y^{(i)}$ is drawn from Gaussian distribution w/ mean $\boxed{w^T x^{(i)}}$, variance $\boxed{\sigma^2}$

Constant

determined by "true" value of w



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

Likelihood of data:

"parameterized by w"

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

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

Trick: Take the log (increasing function)

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

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

maximizing $\log \mathcal{L}(w)$ is same as minimizing original $L(w)$ from linear regression