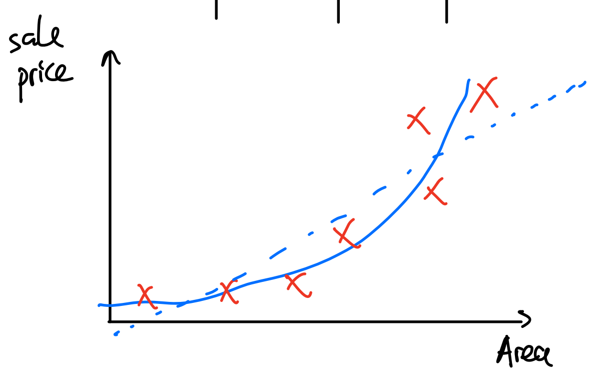
1. Features
2. Convexity
3. Closed-form solution

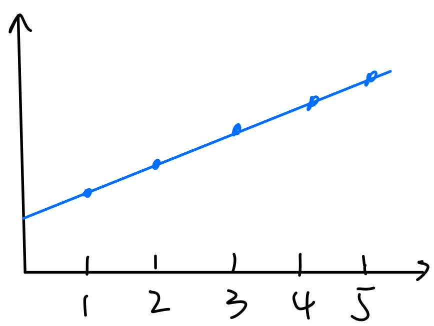
Featurization

sale price (Y)	Area	#bed	house type	Area^2	Area^3
	500	1	condo	250000	
	1000	2	town house	1000000	



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

Linear regression is linear in the features.



indicator features

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

$$\boxed{\begin{aligned} I\{\text{true}\} &= 1 \\ I\{\text{false}\} &= 0 \end{aligned}}$$

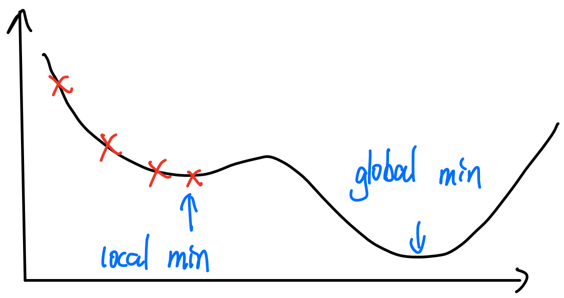
$$\hat{y} = w_1 \, I\{\text{bed} = 1\} + w_2 \, I\{\text{bed} = 2\} + \dots$$

zip code : naively 10^5 features

"Feature engineering": art of choosing features to use

$I\{$zip code is in LA$\}$
allows sharing information between nearby zip codes

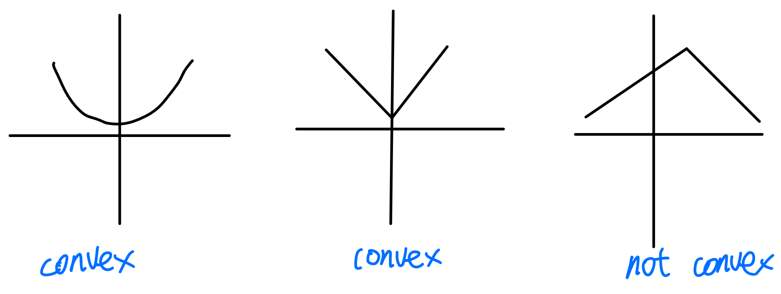
Convexity Why does gradient descent work?



① Linear regression is a convex problem, $L(w)$ is a convex function.

② For a convex function, any local minimum is a global minimum.

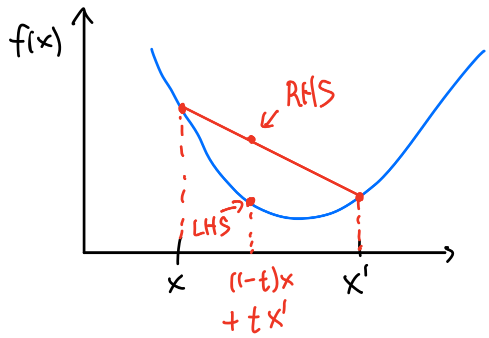
Def 1: $f(x)$ is convex $\Leftrightarrow$ $f''(x) \geq 0$



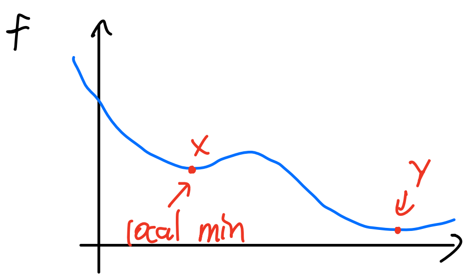
"Def 2": convex function holds water

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

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



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



Given that x is a local minimum of f, assume $\exists y$, $f(y) < f(y)$, show f cannot be convex. (by contradiction)

x is a local min of f iff $\exists \varepsilon > 0$ s.t. $\forall z \in \underbrace{B_\varepsilon(x)}$, $f(z) \ge f(x)$

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

Choose some $t > 0$ s.t.

$$(1-t)x + tx' \in B_\varepsilon(x)$$

① Because it's a local min, $f((1-t)x + tx') \ge f(x)$

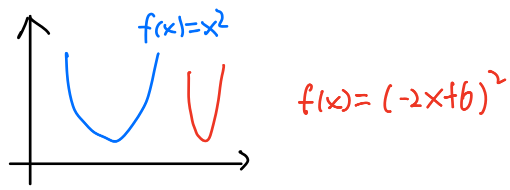
② Because of convexity,

$$\begin{aligned} f((1-t)x + tx') &\le (1-t) f(x) + t f(x') \\ &< (1-t) f(x) + t f(x) \\ &= f(x) \end{aligned}$$

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

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

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 $L(w)$ is convex (by ③)

Closed-form for Linear Regression ("Normal Equations")

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

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

$X^T X w$ $\quad X = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix}$ $\qquad y =$ vector of $y^{(i)}$s

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

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

i-jth entry

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

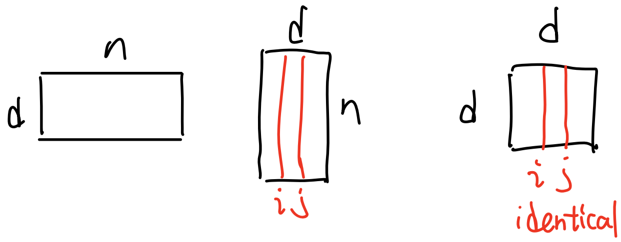
$$X^T X = \sum_{k=1}^{n} x_i^{(k)} x_j^{(k)}$$

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

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

Question: When would $X^T X$ is not invertible?

What if



suppose $x_i^{(k)} = x_j^{(k)}$ $\forall k$

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

$w_i - 100$ $w_j + 100$

Result: answer is not unique anymore!

If i & j^{th} column almost identical $\Rightarrow$ leads to instability

In practice,

① Pseudoinverse A^+

- $A^+ = A^{-1}$ when A^{-1} exists
- For normal equations, A^+ always gives you an optimal solution

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

② Avoid highly-correlated features