CS 331, Fall 2024

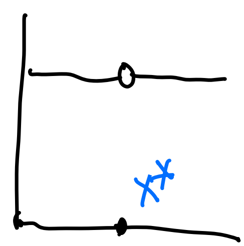
Lecture 16 (10/23)

Today:
- Convexity & Smoothness
- Gradient descent

Convexity & Smoothness (Part VI, Section 3.1)

Last time: Unstructured $\min_{x \in \mathcal{X}} f(x)$

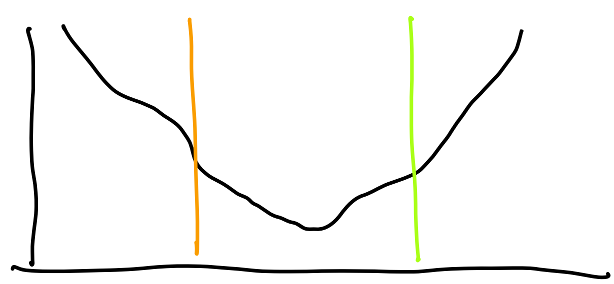
... is possible!



With structure, Unimodal $\min_{x \in \mathcal{X}} f(x)$

$\exists$ algorithms: is possible when $\mathcal{X} \subseteq \mathbb{R}$

(ternary search)



What about high dimensions? $X \in \mathbb{R}^d$

Can't quite ternary search...

Under stronger condition <u>convexity</u> ($\Rightarrow$ unimodal)

we can use

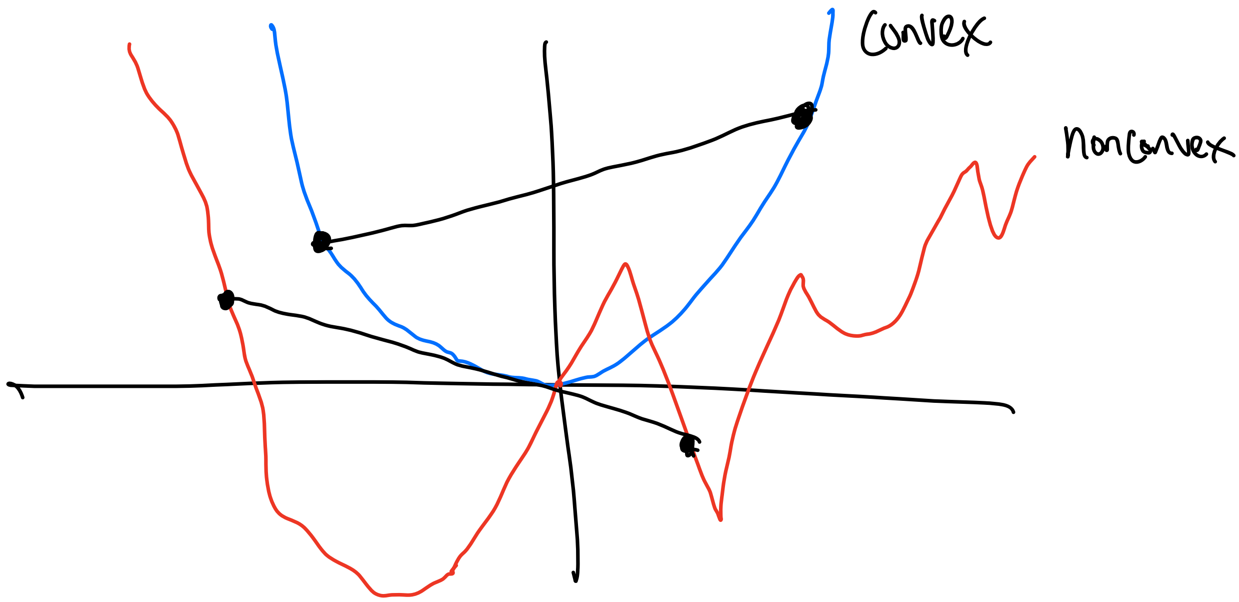
Gradient descent

This is the most important algo in modern ML. It's how ChatGPT learned.

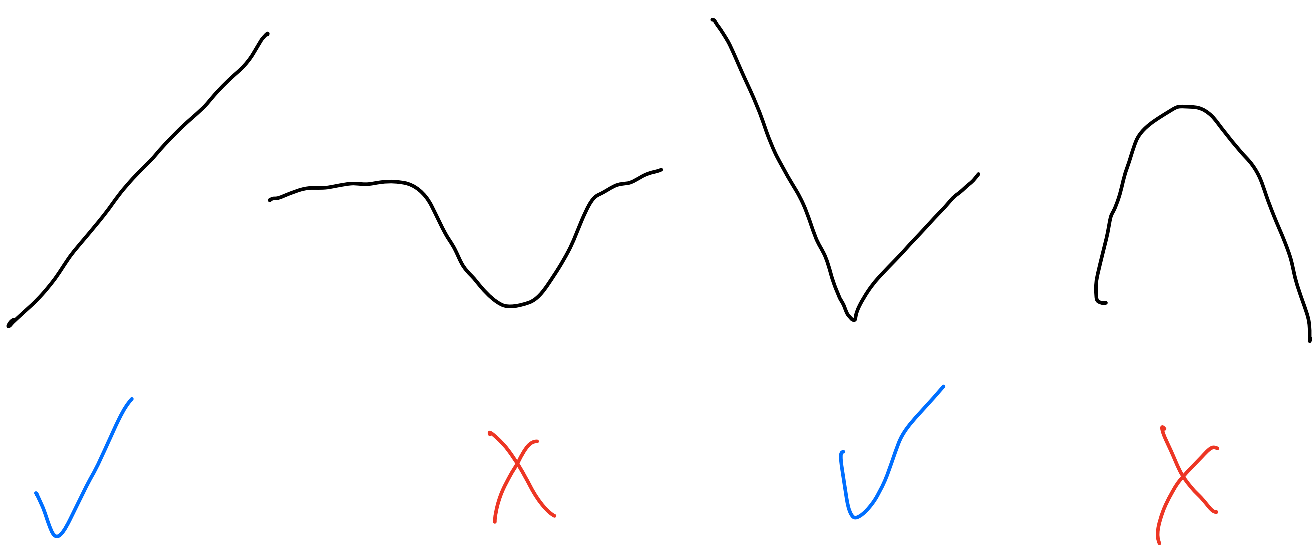
Today: GD in 1-d, $\mathcal{X} = \mathbb{R}$

Convexity

f is convex if at/below the line.



Other convex / nonconvex:

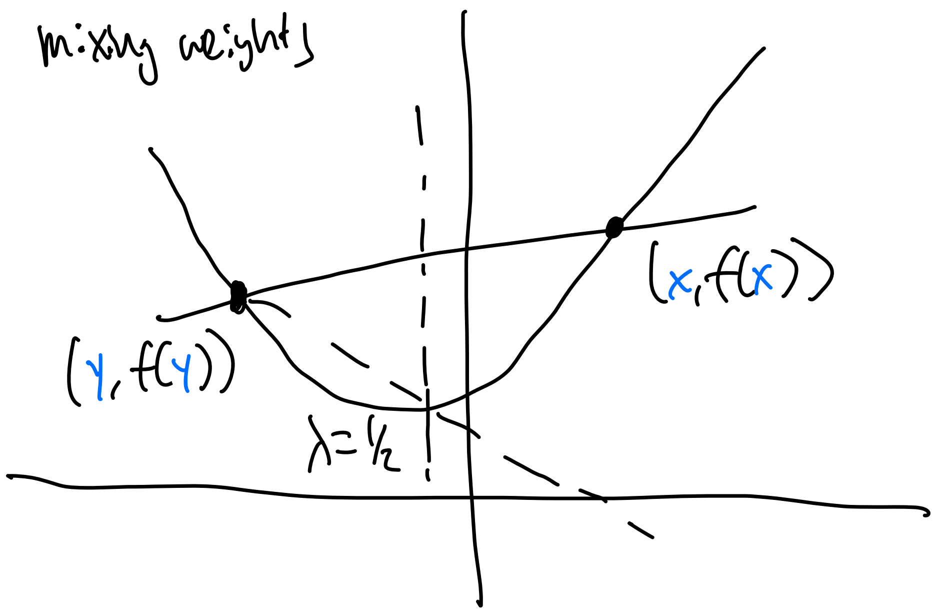


("at the line")

Formally, $\forall \lambda \in [0,1]$

$$f((1-\lambda)x + \lambda y) \le (1-\lambda) f(x) + \lambda f(y)$$

mixing weights

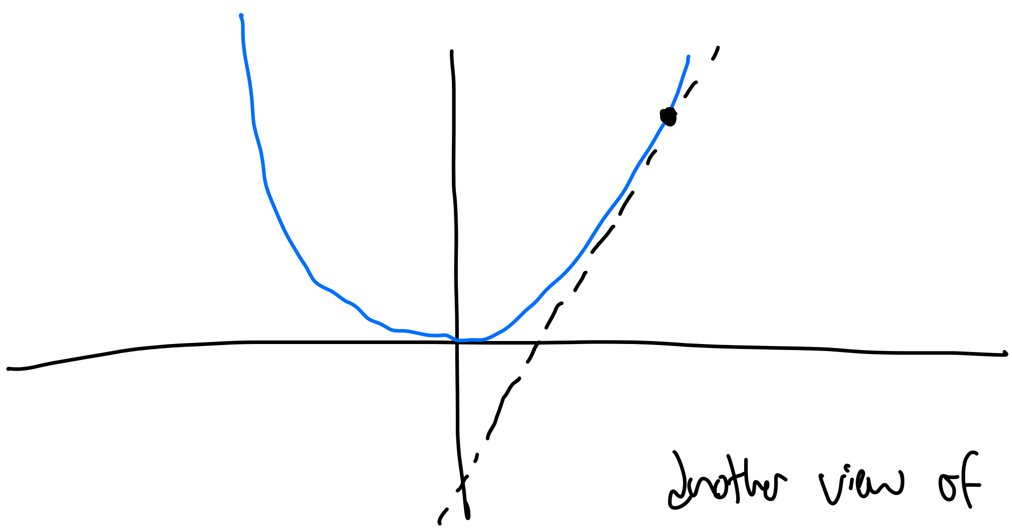


Another fact:

$$f(y) \ge f(x) + \underbrace{\frac{f(x+\lambda(y-x)) - f(x)}{\lambda}}_{\lim_{\lambda \to 0} \; = f'(x)(y-x)}$$

1st order summary $\quad \forall x, y$

$$\underbrace{f(y) \geq f(x) + f'(x)(y-x)}_{\text{"Euler approximation"}}$$



another view of convexity...
above the <u>tangent</u> line

Why good? Binary search!

$$f(x^*) \geq f(x) + f'(x)(x^* - x)$$

$f'(x) > 0$: go left $\quad$ else: go right

Another equivalent condition: $f'' \geq 0$.

e.g.

"The derivative (gradient) increases"

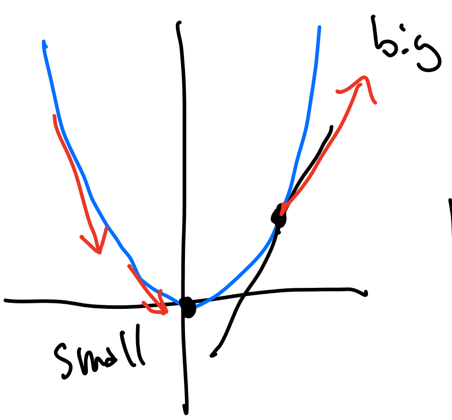
Issue 1: What if f' doesn't exist?

Issue 2: What about $\mathbb{R}^d$???

Idea: gradient descent! (ignore 1)

in high-d, $\nabla f(x)$

$$x \leftarrow x - \underbrace{\eta}_{\text{step size}} f'(x)$$



Why scale w $f'(x)$?

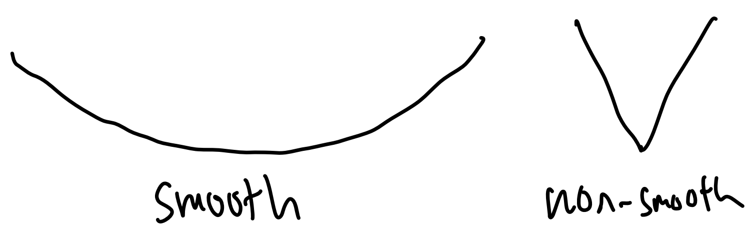
further from x^* $\Rightarrow$ $|f'|$ bigger

We would love to just simulate a rolling ball ($\eta \to 0$). Unfortunately, we must discretize ☹

Important to pick η carefully.

To make it work, we need...

$\boxed{\text{Smoothness}}$



f is L-Lipschitz if

$$|f(x) - f(y)| \leq L|x - y|$$

nearby points have close values.

f is L-smooth if "no corners"

$$|f'(x)-f'(y)| \le L|x-y|$$

$\boxed{\text{Gradient queries}}$ Why ok?

$$\left|f'(x) - \underbrace{\frac{f(x+\delta)-f(x)}{\delta}}_{\text{sim. with function queries}}\right| = \underset{\to 0}{O(\delta)} \text{ if smooth.}$$

GD Take 1: Quadratics (Part VI, Section 3.2)

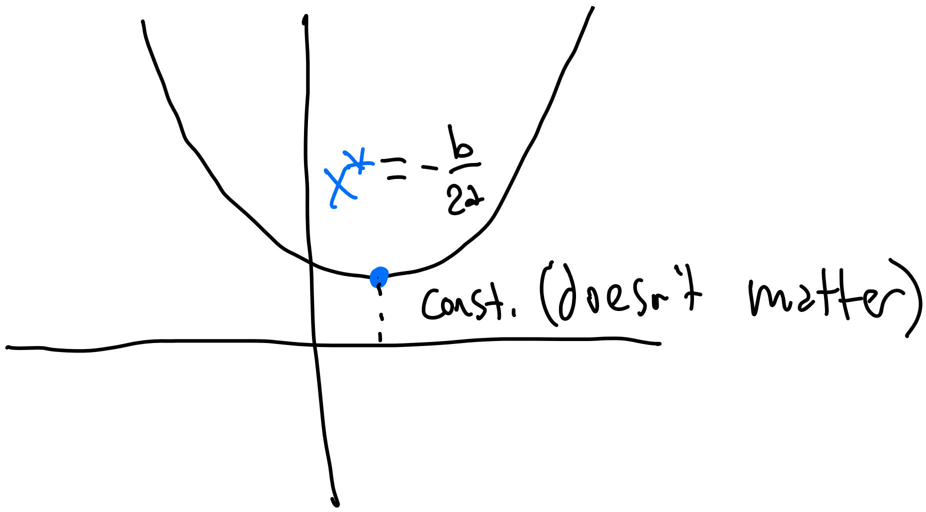
In this section, $f = q$ a quadratic

$$q(x) = ax^2 + bx + c, \quad a > 0$$

What is x^*?

Complete the square:

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + \text{const.}$$



Suboptimality @ x:

$$a\left(x + \frac{b}{2a}\right)^2 - a\left(x^* + \frac{b}{2a}\right)^2$$

$$= a\left(x - x^*\right)^2$$

What is g'?

$$g'(x) = \frac{\partial}{\partial x}\left(ax^2 + bx + c\right)$$

$$= 2ax + b = 2a\left(x - x^*\right)$$

Sanity check: $g'(x^*) = 0$

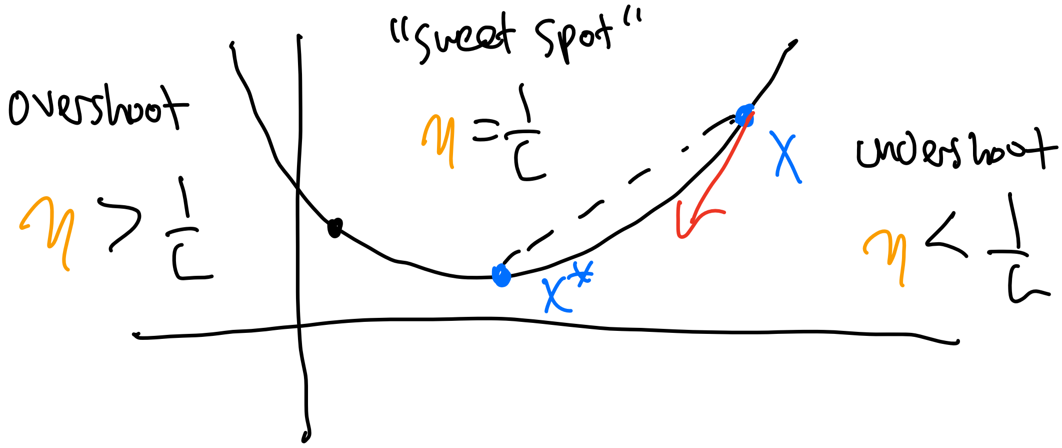
Hence, $L = 2a$-smooth:

$$|g'(x) - g'(y)| = 2a|x - y|$$

What step size?

$$x - \eta f'(x) = x - \eta \cdot 2a(x - x^*)$$

$$\eta = \frac{1}{2a} = \frac{1}{L} \Rightarrow \text{step to } x^*$$



full characterization of quadratics

GD Take 2: Smoothness (Part VI, Section 3.3)

$$|f'(x) - f'(y)| \leq L|x - y|$$

Why is it good? Upper bounded by q

Intuition: $q'' = L$, $f'' \leq L$

$$\lim_{y \to x} \frac{f'(x) - f'(y)}{x - y}$$

Just as best linear fit

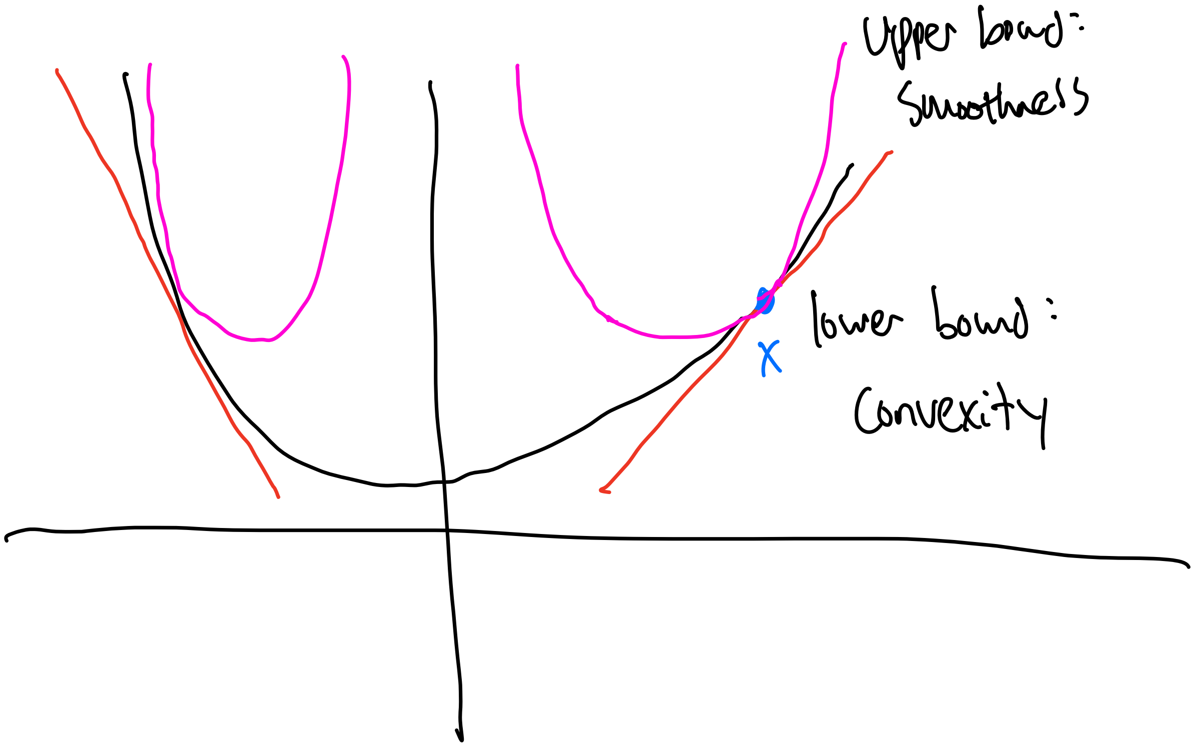
$$f(y) \approx f(x) + f'(x)(y-x)$$

Best quadratic fit

$$f(y) \approx f(x) + f'(x)(y-x) + \frac{1}{2}\underbrace{f''(x)}_{\leq L}(y-x)^2$$

Can be made rigorous: $\forall\, x, y$

$$\left.\begin{aligned} f(y) \leq f(x) + f'(x)(y-x) \\ + \frac{L}{2}(y-x)^2 \end{aligned}\right\} q(y)$$



Note: @ x, $q' = f' = \ell'$

Also, q is L-smooth.

Hence, $x \leftarrow x - \frac{1}{L} f'(x)$

$= x - \frac{1}{L} q'(x)$

minimizes q @ each step.

Critical points

L-smooth, not convex

Algo: assume $f(x_0) - f(x^*) \le \Delta$

iterate $\forall t \in [T]$

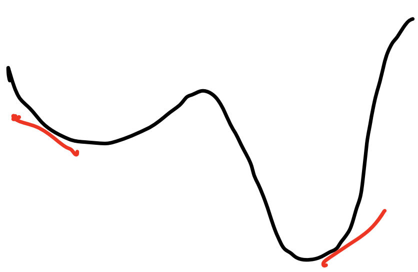
$$x_t \leftarrow x_{t-1} - \frac{1}{L} f'(x_{t-1})$$

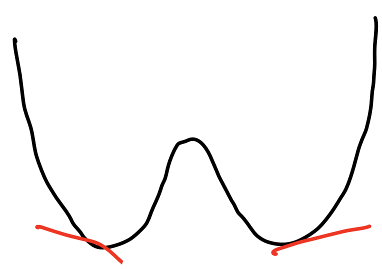
Claim: $T \ge \frac{2L\Delta}{\varepsilon^2} \Rightarrow \min_t \underbrace{|f'(x_t)| \le \varepsilon}_{\approx \text{ local minimum}}$

Proof: Suppose otherwise. Progress / iter

$$\ge \frac{1}{2L} |f'(x_t)|^2 \ge \frac{\varepsilon^2}{2L} \quad \star$$

After T iters, more than Δ progress ■





modern neural nets?

$\boxed{\text{Global optimality}}$

Algo: assume $|x_0 - x^*| \leq 1$,

f is L-smooth & convex

iterate $\forall\, t \in [T]$

$$x_t \leftarrow x_{t-1} - \frac{1}{L} f'(x_{t-1})$$

Helper claim: $|x_t - x^*| \leq |x_0 - x^*|$

(see notes) never overshoots.

Claim: After $T \geq \frac{L^2}{\varepsilon^2}$ iters,

$$f(x_t) \leq \underbrace{f(x^*)}_{\text{global opt}} + \varepsilon$$

Proof: $\Delta \leq \frac{L}{2}$,

$$f(x_0) \leq f(x^*) + \frac{L}{2}\overbrace{(x_0 - x^*)}^{\leq 1}{}^2$$

Use critical points, $T \geq \frac{L^2}{\varepsilon^2} = \frac{2L\Delta}{\varepsilon^2}$

$$f(x) - f(x^*) \leq f'(x)(x - x^*)$$
(convexity)
$$\leq |f'(x)| \leq \varepsilon$$

$\boxed{\text{... and beyond}}$

We proved $T \geq \frac{L^2}{\varepsilon^2}$ suffices today.

1) Same algo but smarter analysis $\Rightarrow \frac{L}{\varepsilon}$

2) Optimal algo: $\sqrt{\frac{L}{\varepsilon}}$ (Nesterov)

Key idea: momentum

for chat GPT, $d = 10^{12}$

Why is this important?

Generalizes verbatim to $\mathbb{R}^d$

(critical points & global optima)

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_d}(x) \right)$$

"gradient"

"hint" of where to find x^*