CS 331, Fall 2025

Lecture 15 (10/20)

Today:
- Continuous optimization
- Convexity & Smoothness
- Gradient descent

Continuous optimization (Part VI, Section 1)

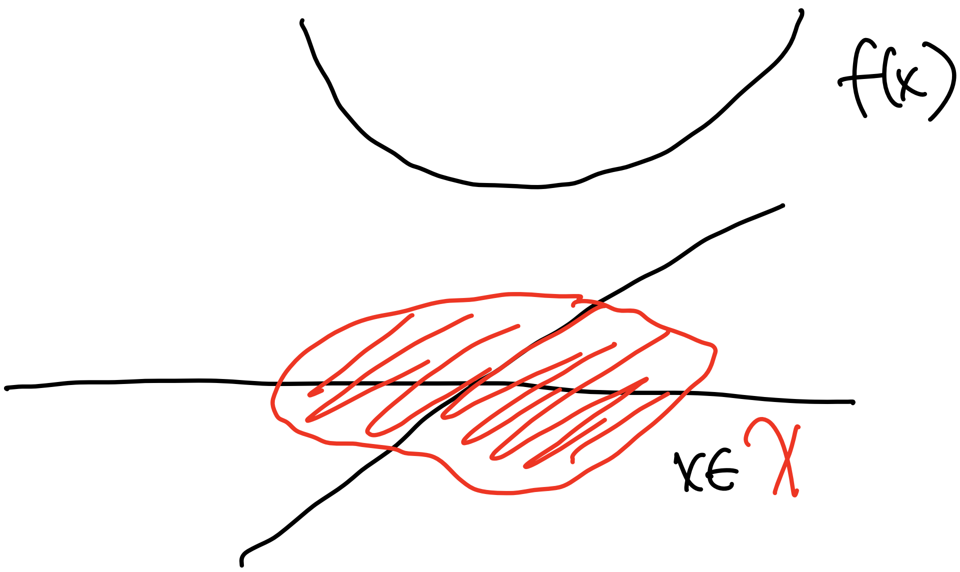
This unit:

"decision variable"

$$\min_{x \in \mathcal{X}} f(x) \quad \text{or} \quad \max_{x \in \mathcal{X}} f(x)$$

e.g.	f	$\mathcal{X}$
Scheduling	size	non-overlapping intervals
MST	total weight	spanning trees
s-t shortest path	total weight	s-t paths

Key difference: now $\underbrace{\mathcal{X}}_{\text{continuous}} \subseteq \mathbb{R}^d$



- infinite sets
- typically cannot hope for <u>exact</u> only high-accuracy (some exceptions)

e.g. s-t maxflow

$$\max_{x \in \mathcal{X}} f(x)$$

$$f(x) = \sum_{e=(s,v) \in E} x_e$$

$$\mathcal{X} = \left\{ x \in \mathbb{R}^E \;\middle|\; \begin{array}{l} 0 \le x_e \le c_e \\ \displaystyle\sum_{e=(v,u)} x_e - \sum_{e=(u,v)} x_e = 0 \\ \forall v \in V \setminus \{s,t\} \end{array} \right\}$$

What are the rules of the game?

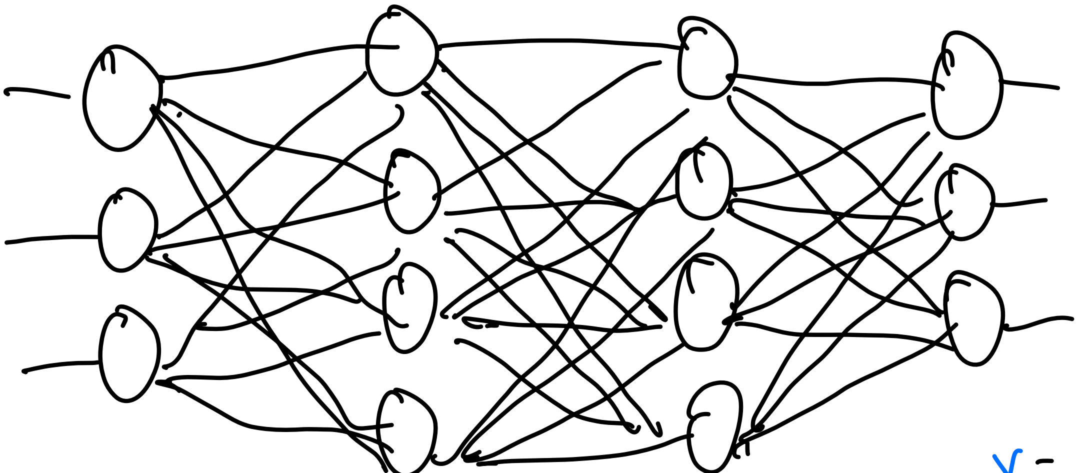
$$\min_{x \in \mathcal{X}} f(x)$$

- ability to evaluate f @ x
- ability to evaluate f' @ x reduces to

$$f'(x) \approx \frac{f(x+\delta) - f(x)}{\delta}$$

Sometimes, this is all that's reasonable.

$f(x)$ = average of neural network loss @ 1000000 pictures???



x = neural net weights

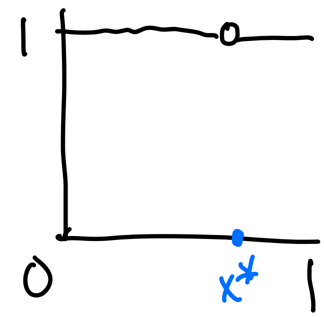
However, this is not enough. Need structure

e.g. Can we compute $\hat{x} \in \mathcal{X}$

s.t. $f(\hat{x}) \le \min_{x \in \mathcal{X}} f(x) + 0.99$,

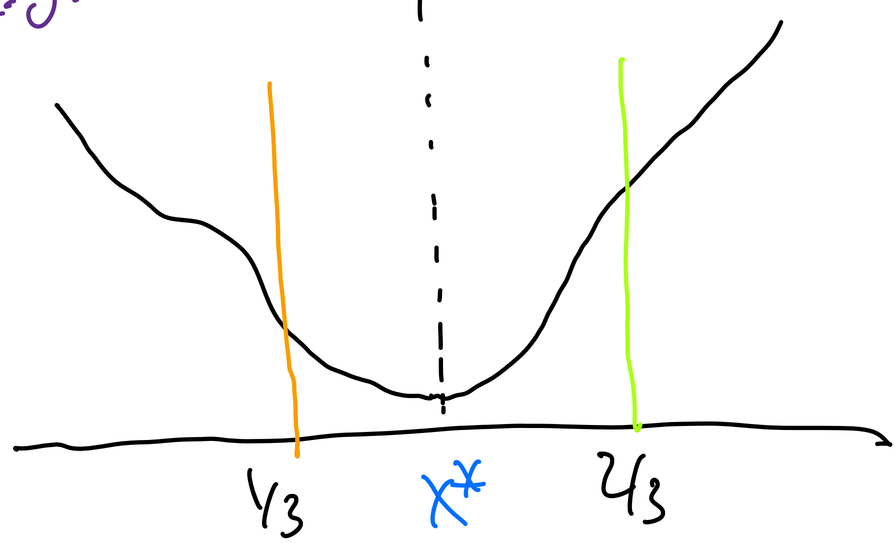
$\mathcal{X} = [0,1]$, $f : [0,1] \to [0,1]$?

No! Let $f(x) = \begin{cases} 0 & x = x^* \text{ (secret)} \\ 1 & \text{else} \end{cases}$



What structure helps?

e.g. Unimodality



"ternary search"
repeatedly throw out 1/3 using only f queries

Key difference: "hints" about x^*

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

Can't quite ternary search...

We can use

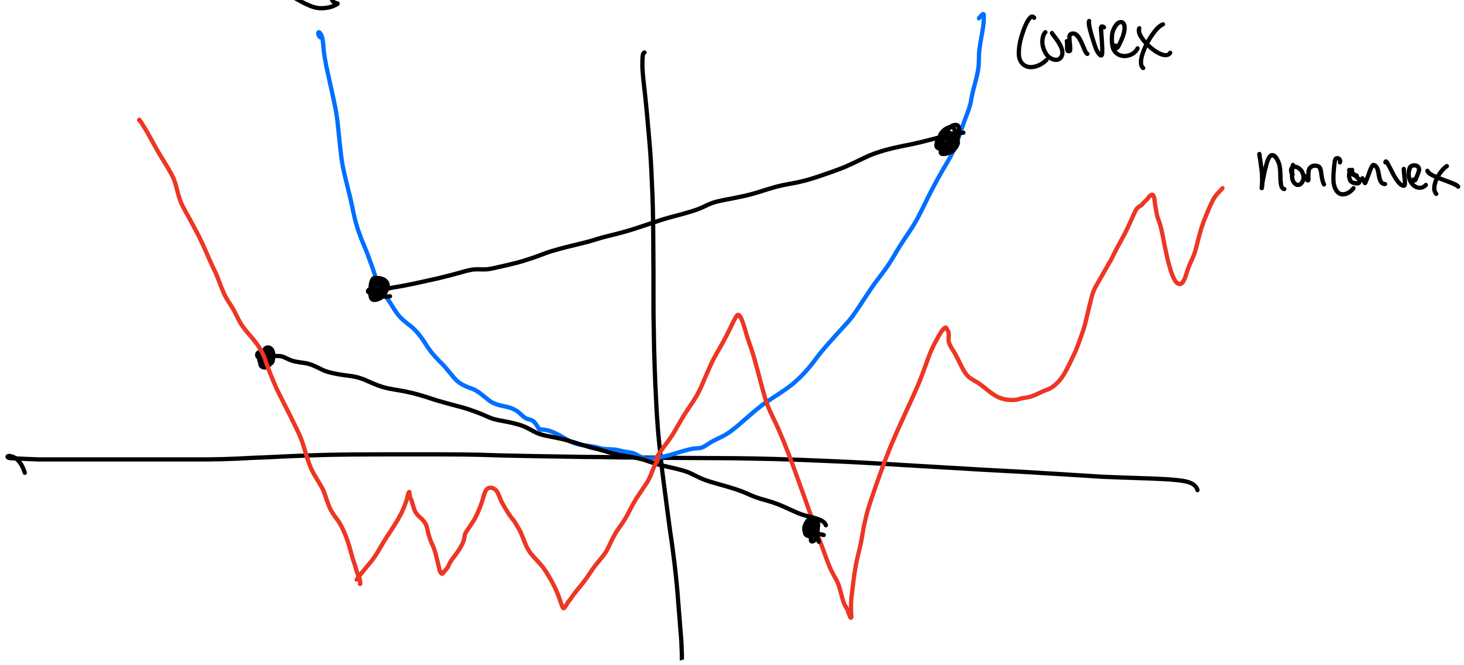
Gradient descent

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

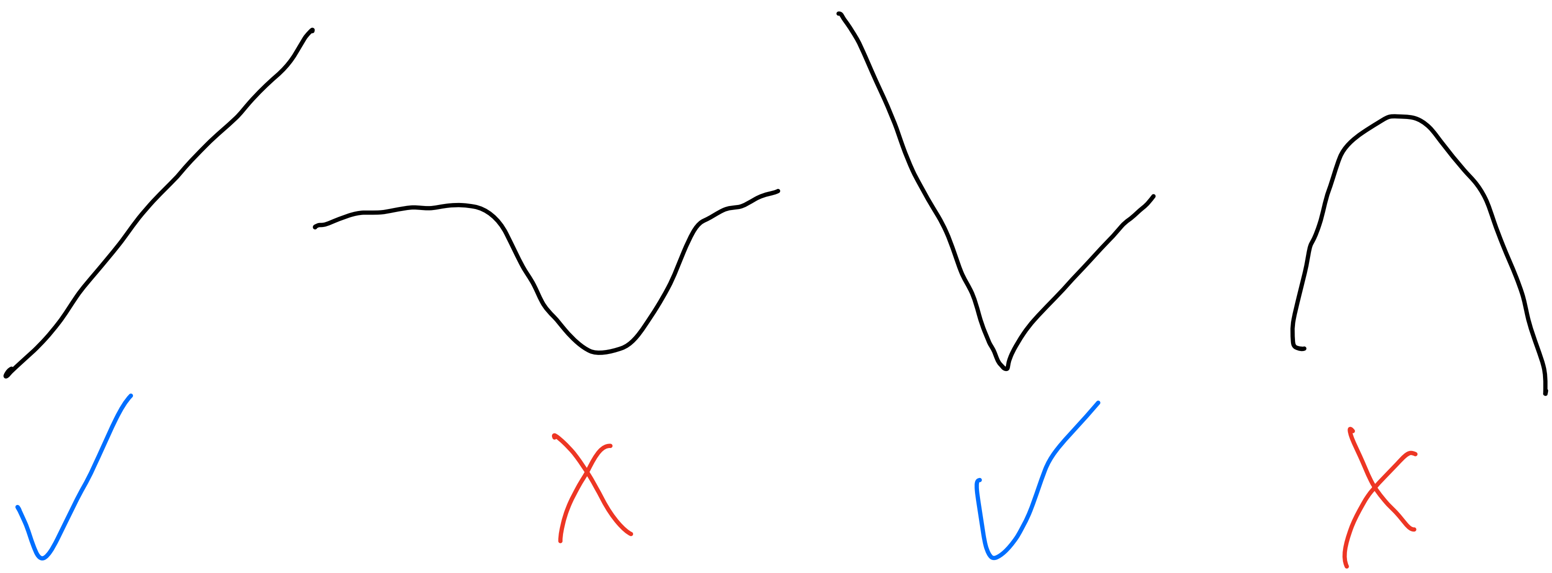
Today: GD in 1-d, $x = \mathbb{R}$

Structural assumptions (Part VI, Section 2.1)

Convexity f is convex if at/below the line.



Other convex/nonconvex:

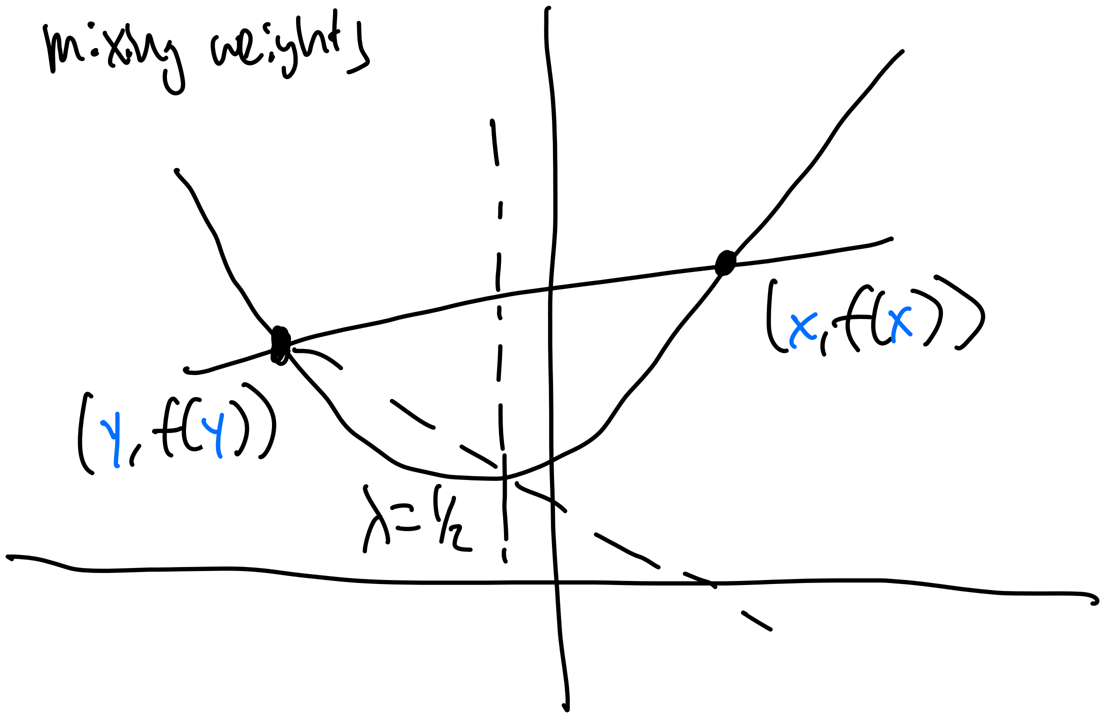


("at the line")

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

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

mixing weights

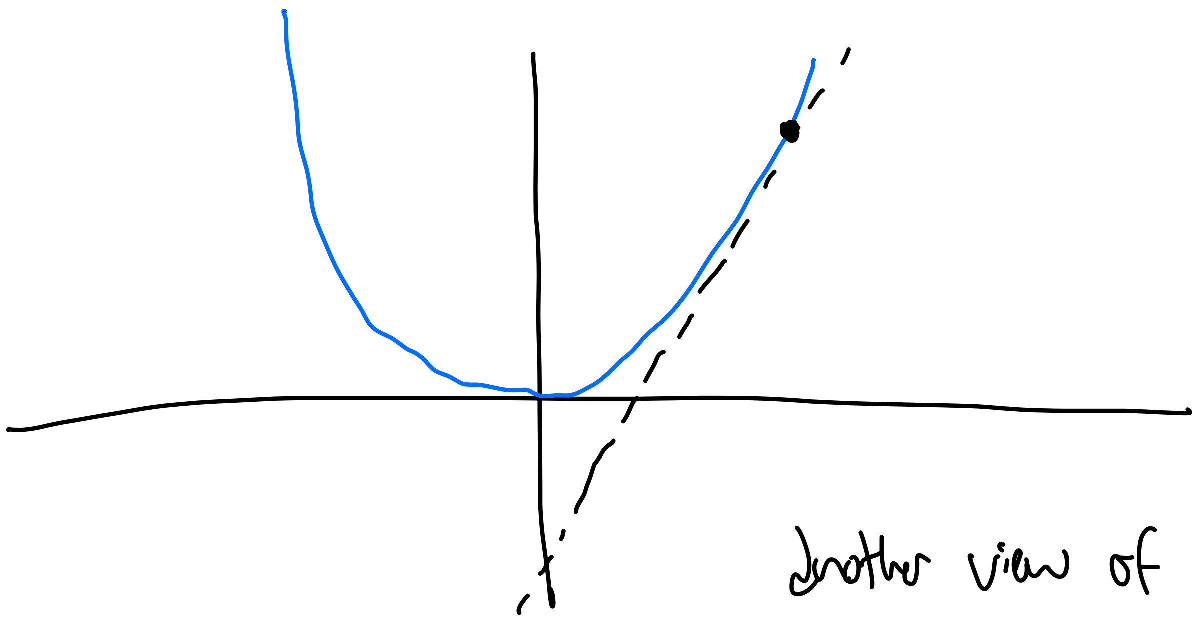


Another fact:

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

lu summary $\quad \forall x, y$

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



another view of convexity...

above the <u>tangent</u> line

Why good? Binary search!

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

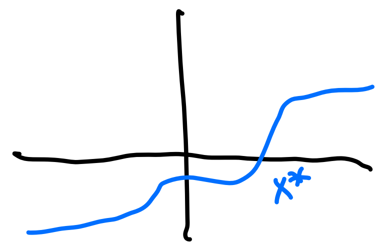
$f'(x) > 0$: go left $\quad$ else: go right

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

e.g.

"The derivative (gradient) increases"

Can binary search for $f'(x) = 0$



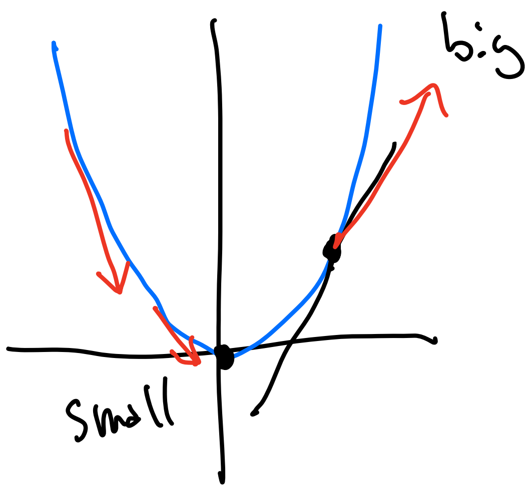
plot of f'

Issue: What does this imply for high dimensions ($d \gg 1$)?

Idea: gradient descent!

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

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



Why scale w $f'(x)$?

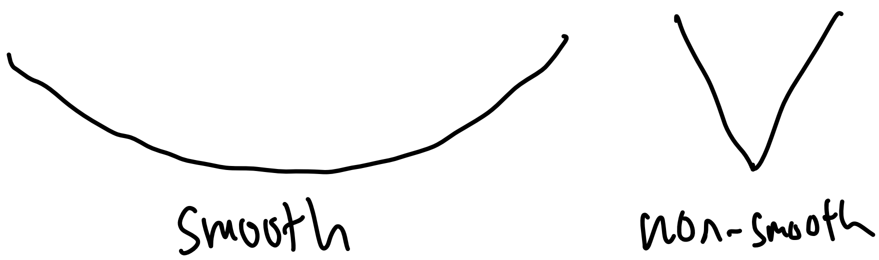
further from $x^* \Rightarrow |f'|$ bigger

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"

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

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

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

GD Take 1: Quadratics (Part VI, Section 2.2)

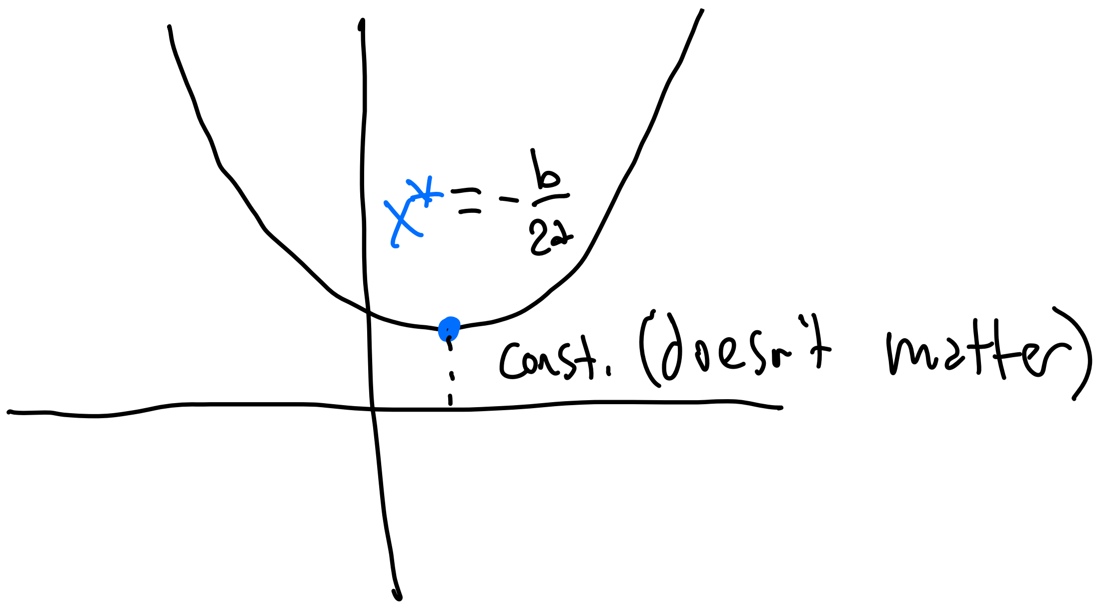
In this section, $f = q$ a quadratic

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

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 q'?

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

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

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

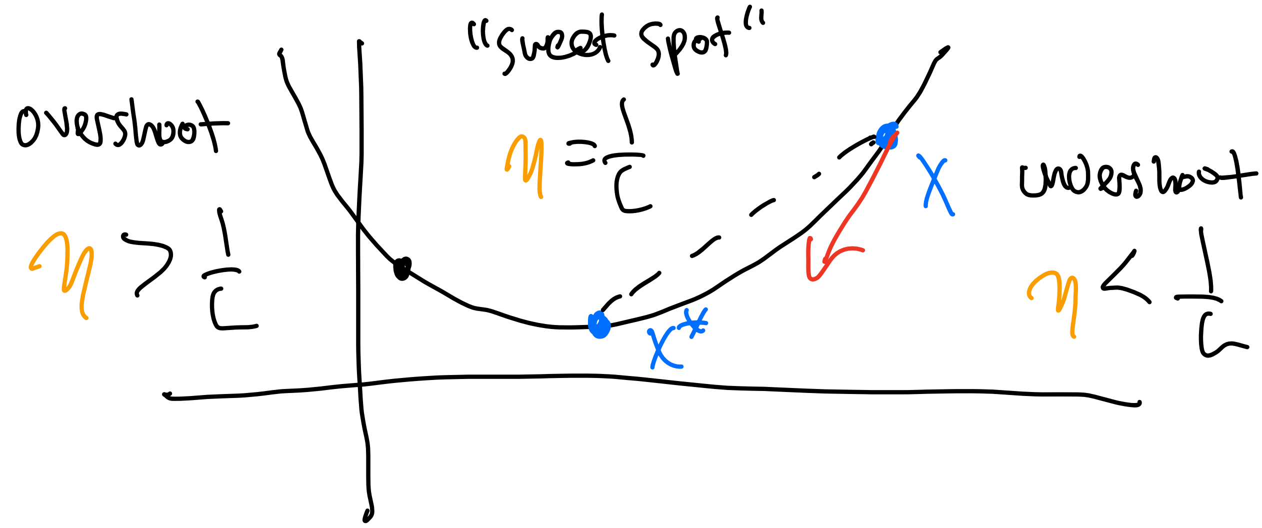
Hence, $L = 2a$-smooth:

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

What step size?

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

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



full characterization of quadratics

GD Take 2: Smoothness (Part VI, Section 2.3)

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

Why is it good? Upper bounded by q

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

$\boxed{\text{Useful facts}}$

$|f''| \le L \iff f$ is L-smooth

$|f'| \le L \iff f$ is L-Lipschitz

Just as best linear fit

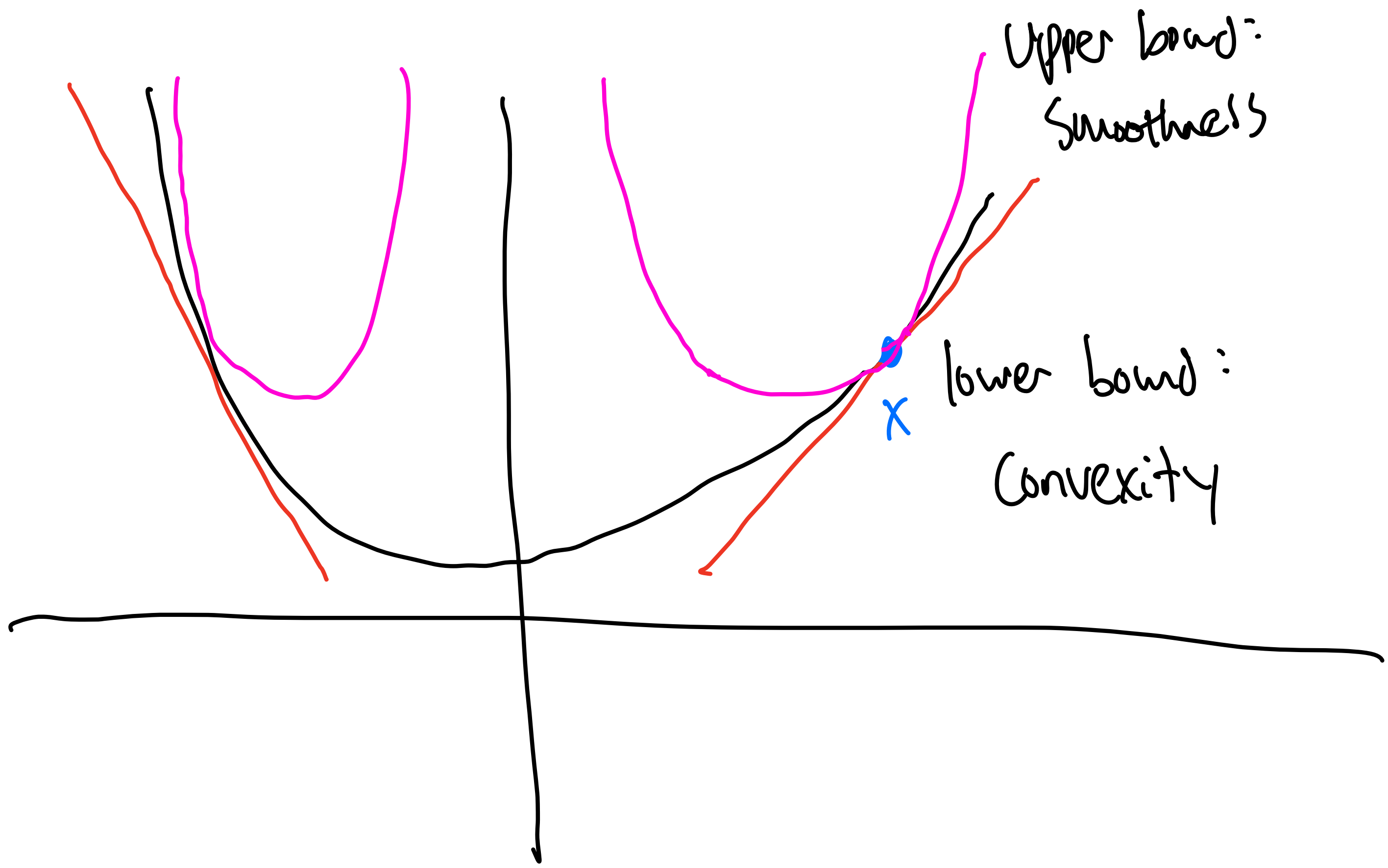
$$f(y) \approx f(x) + f'(x)(y-x)$$

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: $\forall\, x, y$

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



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

Also, q is L-smooth.

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

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

minimizes q @ each step.

Critical points

L-smooth, not convex

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

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

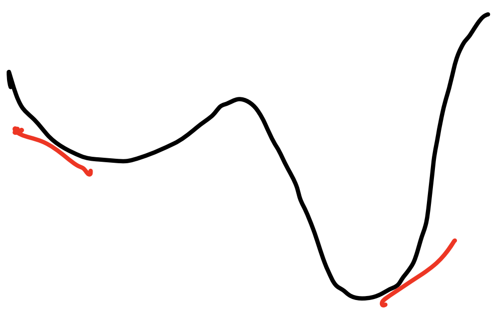
$$x_t \leftarrow x_{t-1} - \frac{1}{L} f'(x_{t-1})$$

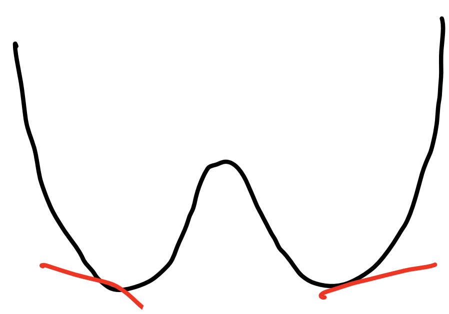
Claim: $T \geq \dfrac{2L\Delta}{\varepsilon^2} \Rightarrow \min_t \underbrace{|f'(x_t)| \leq \varepsilon}_{\approx \text{ local minimum}}$

Proof: Suppose otherwise. Progress / iter

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

After T iters, more than Δ progress ∎





modern neural nets?

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: $\Delta \le \frac{L}{2}$,

$$f(x_0) \le f(x^*) + \frac{L}{2}\big(\underbrace{x_0 - x^*}_{\le 1}\big)^2$$

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

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

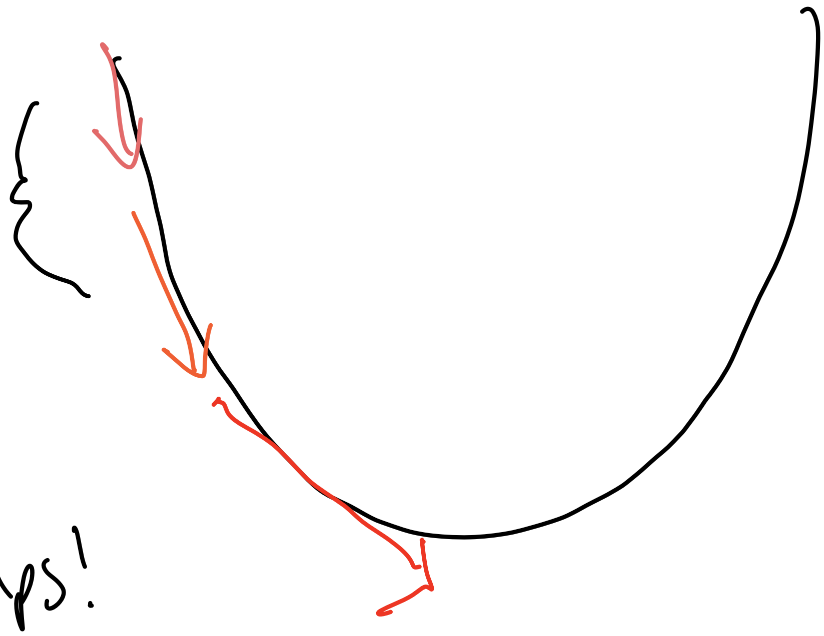
$\boxed{\text{... and beyond}}$

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

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

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

Key idea: momentum



Smoothness = history helps!

for ChatGPT, $d = 10^{12}$

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^*