

CS 331, Fall 2026

Lecture 16 (10/21)

Today: - Convexity & Smoothness

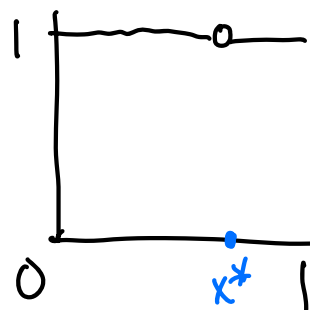
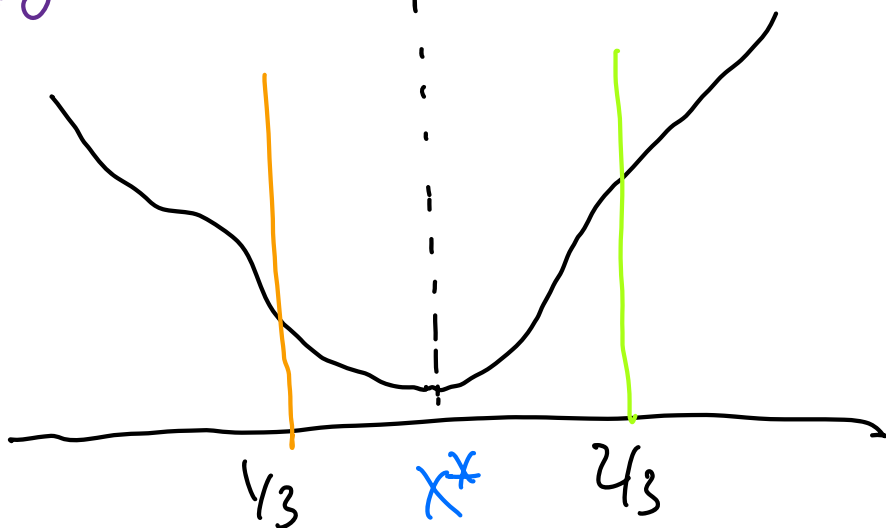
- Gradient descent

Gradient descent (Part VI, Section 3)

Key lesson last time: unstructured opt impossible

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



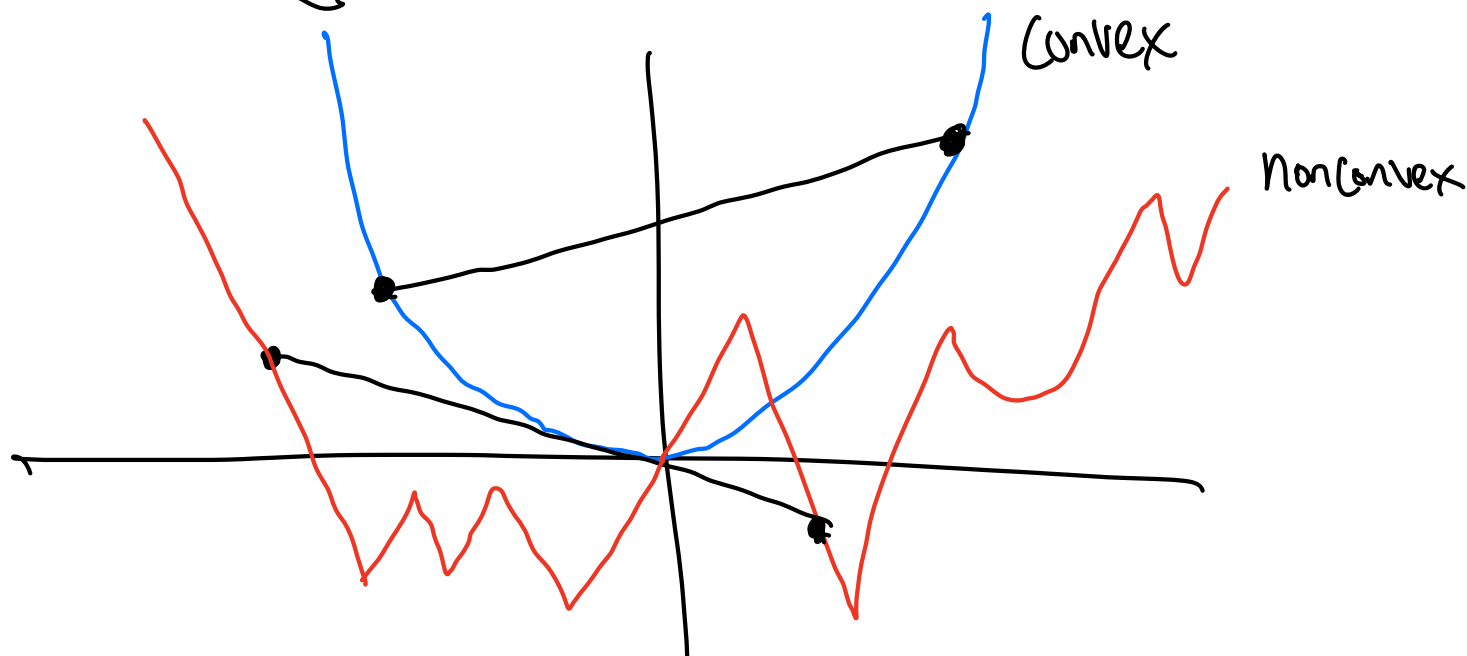
This is the most important algo in modern ML. It's how ChatGPT learned.

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

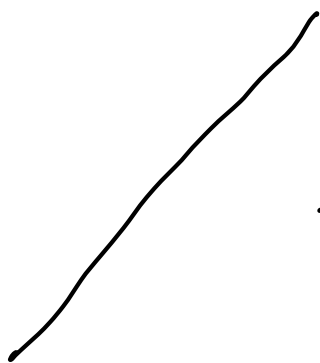
Structural assumptions (Part VI, Section 2.1)

Convexity

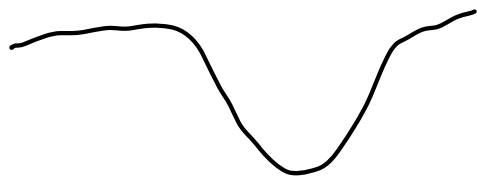
f is convex if at/below the line.



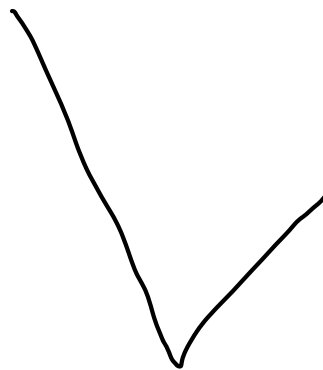
Other convex / nonconvex:



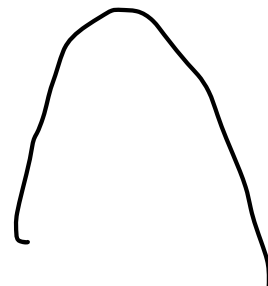
✓



✗



✓



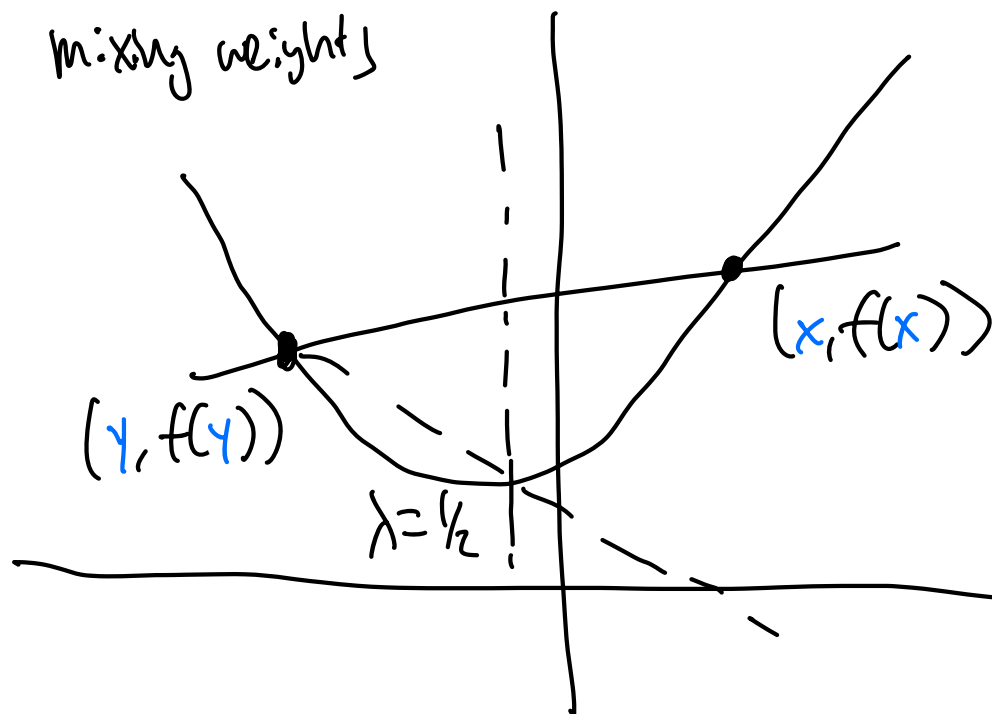
✗

("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:

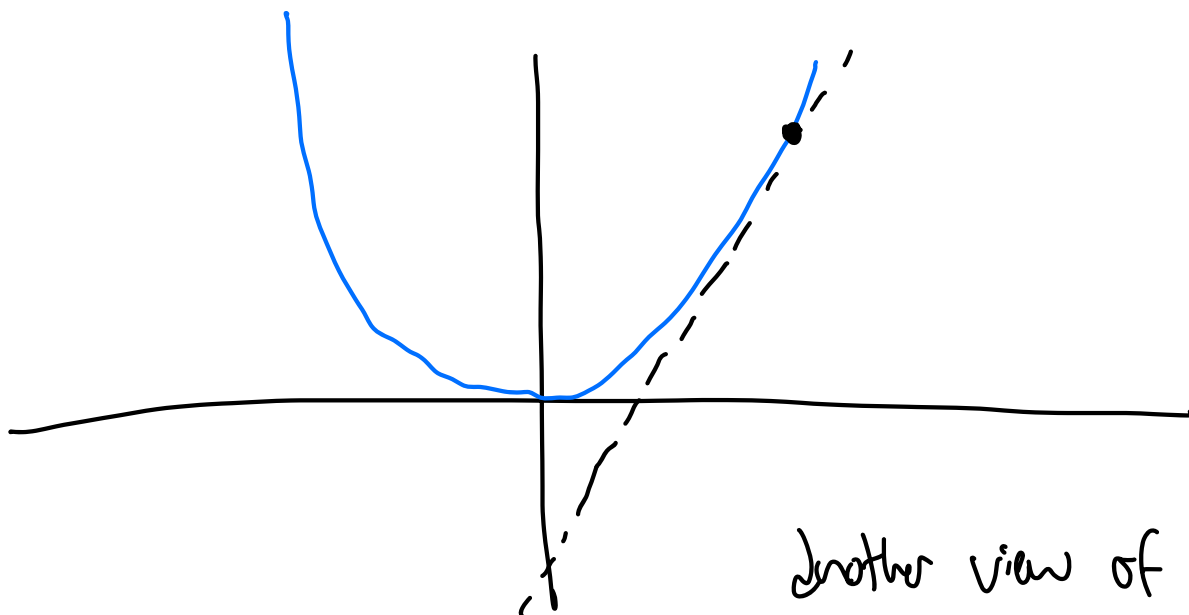
$$f(y) \geq f(x) + \underbrace{\frac{f(x + \lambda(y-x)) - f(x)}{\lambda}}_{\lim_{\lambda \rightarrow 0} = f'(x)(y-x)}$$

In summary $\forall x, y$

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



"Euler approximation"



Another view of convexity...
above the tangent line

Why good? Binary search!

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

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

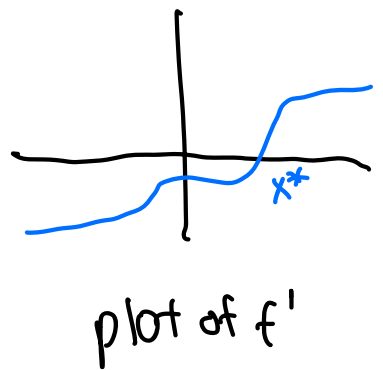
Another equivalent condition: $f'' \geq 0$.

e.g.

"The derivative (gradient) increases!"

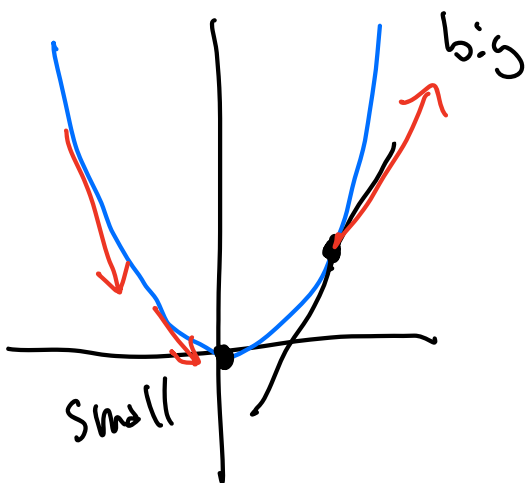
(Can binary search for $f'(x) = 0$)

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 $f'(x)$?

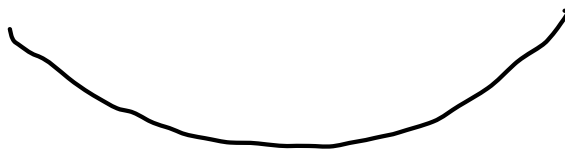
further from $x^* \Rightarrow |f'|$ bigger

We would love to just simulate a rolling ball ($n \rightarrow 0$). Unfortunately, we must discretize :-

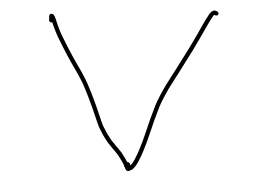
Important to pick M carefully.

To make it work, we need...

Smoothness



Smooth



non-smooth

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"

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

Gradient queries

Why ok?

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

sim. with function queries

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

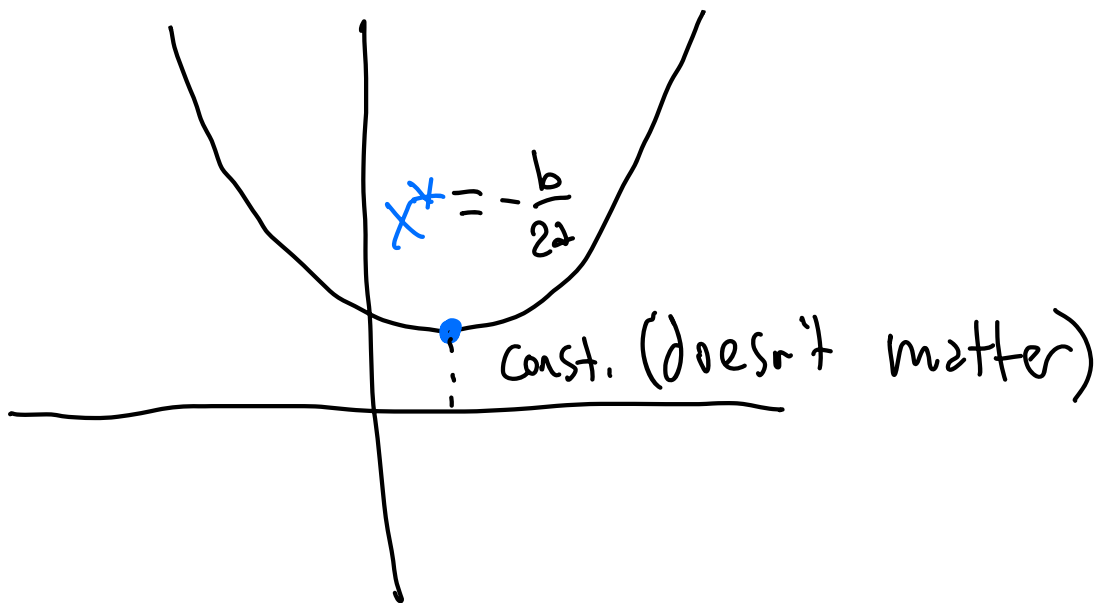
In this section, $f = q$ a quadratic

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

What is x^* ?

Complete the square:

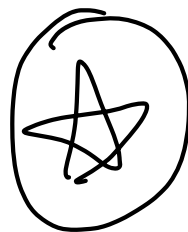
$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + \text{const.}$$



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' ?

$$q'(x) = \frac{d}{dx} (2x^2 + bx + c) \\ = 2a x + b = 2a (x - x^*)$$

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

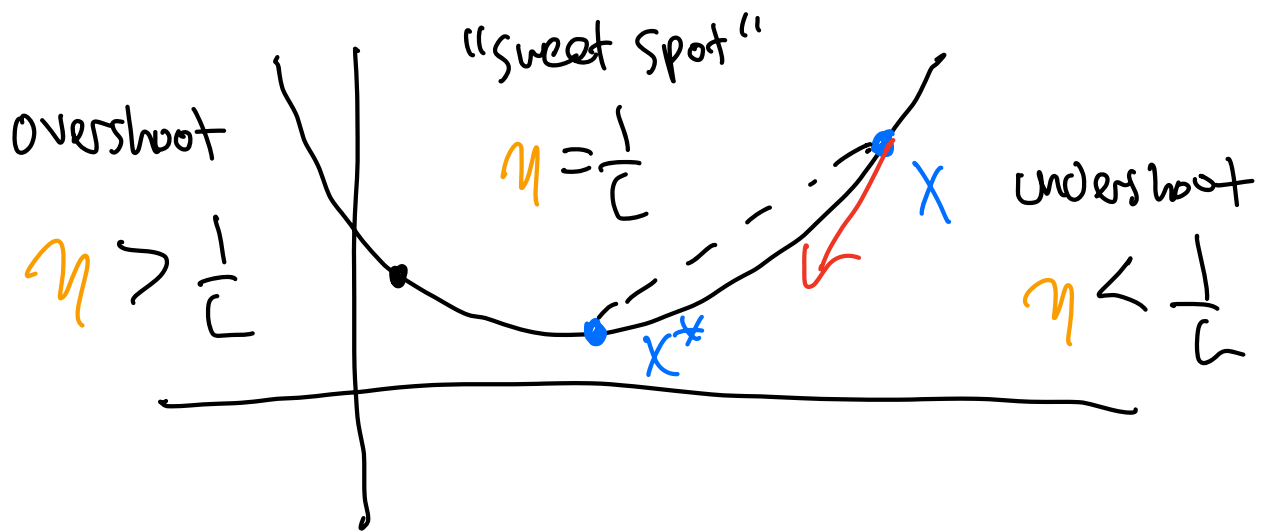
Hence, $L = 2a$ -smooth:

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

What step size?

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

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



full characterization of quadratics

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

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

Why is it good? Upper bounded by q

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

Useful facts

$$|f''| \leq L \Leftrightarrow f \text{ is } L\text{-smooth}$$

$$|f'| \leq L \Leftrightarrow f \text{ is } L\text{-Lipschitz}$$

Just as best linear fit

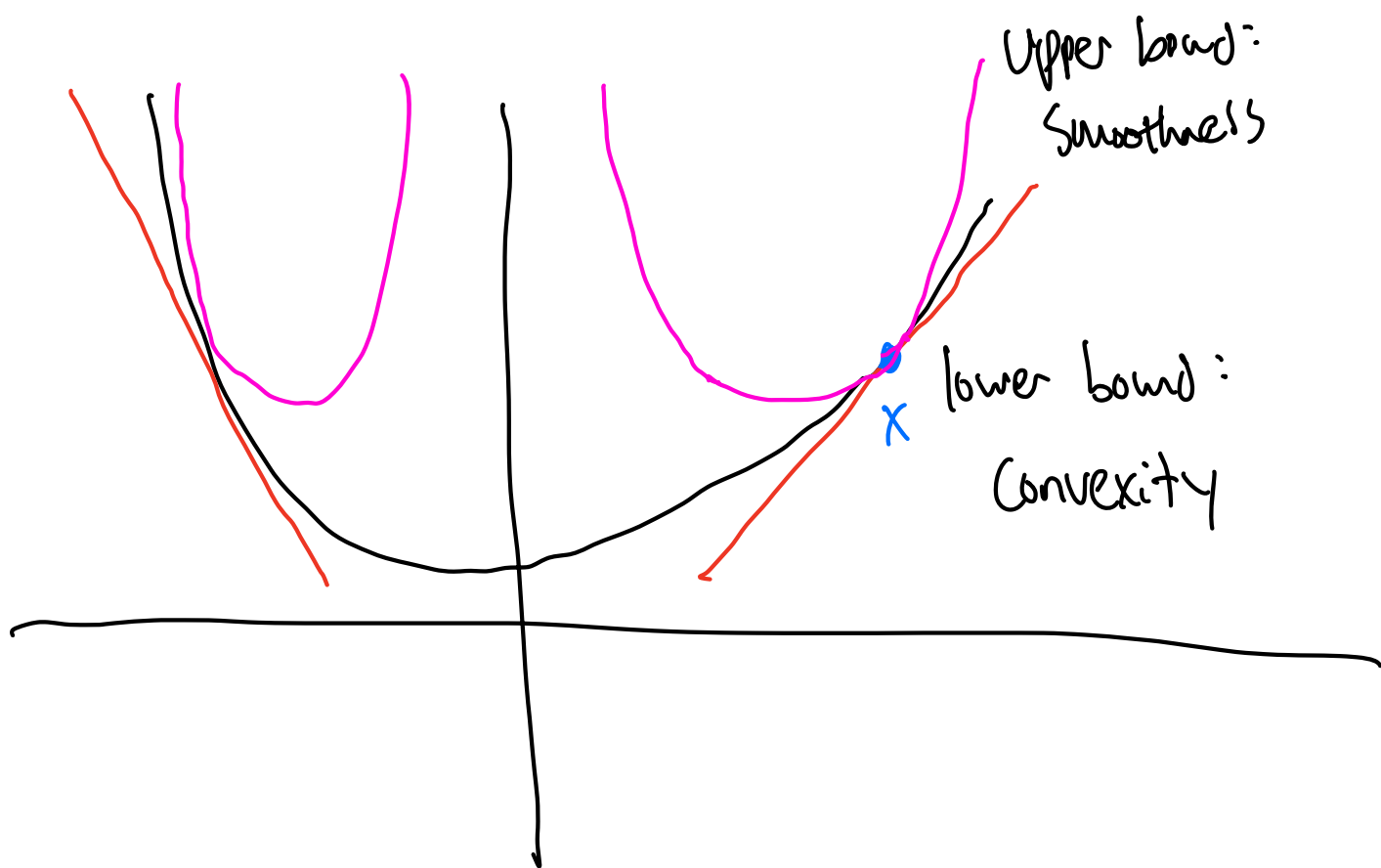
$$f(y) \approx f(x) + f'(x)(y - x)$$

Best quadratic fit

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

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 , $q' = f' = l'$

Also, q is L -Smooth.

$$\begin{aligned} \text{Hence, } x &\leftarrow x - \frac{1}{L} f'(x) \\ &= x - \frac{1}{L} q'(x) \end{aligned}$$

minimizes q @ each step.

Critical points

L-smooth, not convex

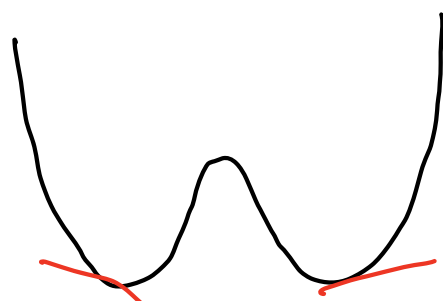
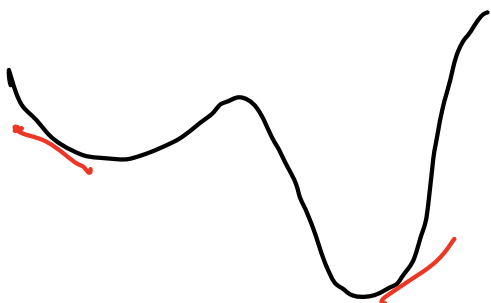
Algo: assume $f(x_0) - f(x^*) \leq \Delta$
iterate $\forall t \in [T]$

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

Claim: $T \geq \frac{2L\Delta}{\epsilon^2} \Rightarrow \min_t |f'(x_t)| \leq \epsilon$
 $\approx$ local minimum

Proof: Suppose otherwise. Progress / iter
 $\geq \frac{1}{2L} |f'(x_t)|^2 \geq \frac{\epsilon^2}{2L}$ 

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 $T \geq \frac{L^2}{\epsilon^2}$ iters,

$$f(x_t) \leq f(x^*) + \epsilon$$



global opt

Proof: $\Delta \leq \frac{L}{2}$, $\overbrace{\frac{L}{2} (x_0 - x^*)^2}^{\leq 1}$

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

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

$$\begin{aligned} f(x) - f(x^*) &\leq f'(x)(x - x^*) \\ (\text{convexity}) \quad &\leq |f'(x)| \leq \epsilon \end{aligned}$$

... and beyond

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

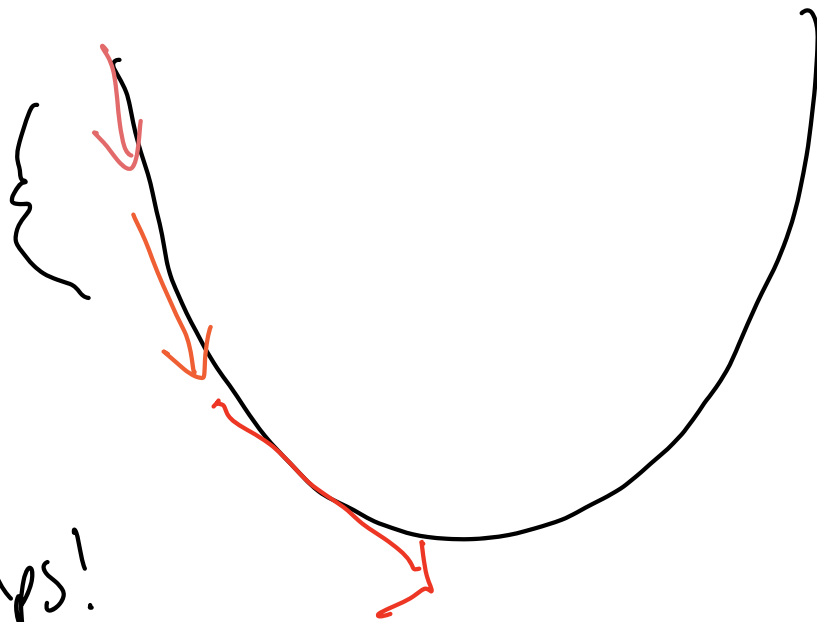
1) Same algo but smarter analysis $\Rightarrow \frac{L}{\epsilon}$

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

Key idea: momentum

remember these

Smoothness =
history helps!



Why is this important?

for that GPT,
 $\delta = 10^{12}$

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

(critical points $\hat{=}$ global optima)

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_D}(\mathbf{x}) \right)$$

"gradient"

"hint" of where to find $\mathbf{x}^*$