Overview of week

HW end of week / early next grades

A4 due Monday 11/2

today : optimization

rest of week : nonlinear models

Friday : project

Today's goals : SGD w/ mini-batches

convex sets & functions

why we like convexity

SGD w/ minibatch

I_t = uniform random index $\{1, \dots, n\}$ ← data pts

$\vec{d}_t = -\nabla \ell_{I_t}(\vec{w}_t)$

$\mathbb{E}[\vec{d}_t] = -\nabla C(\vec{w}_t)$

$= -\nabla \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \ell_i(\vec{w}_t)\right)}_{\text{average of } n}$

$\vec{w}_{t+1} = \vec{w}_t + h_t \vec{d}_t$

mini-batching : variance decreases $\sim 1/B$

⇒ closer to average

Pick B random indices (w/ replacement)

$I_t = \{I_{t1}, I_{t2}, \dots, I_{tB}\}$ $|I_t| = B$

$$\vec{d}_t = -\frac{1}{B}\sum_{I \in I_t} \nabla \ell_I(\vec{w}_t)$$

(SGD : $B=1$)

B → # of data pts in batch

Practical considerations, advantages of SGD

- less computation than GD $\mathcal{O}(B)$ vs. $\mathcal{O}(n)$
  memory } important for GPU and NNs
- parallelizes easily
  - different processes / computers working on different batches (Hogwild!)
- special sauce of SGD for NNs
  - "implicit bias" (bias-var) of SGD small $\|\vec{w}\|$ similar to ridge
  - noise helps avoid local min
  - debatable how important

Disadvantages:

- more iterates than GD
- noisier trajectories, not always descent

depends (convex)

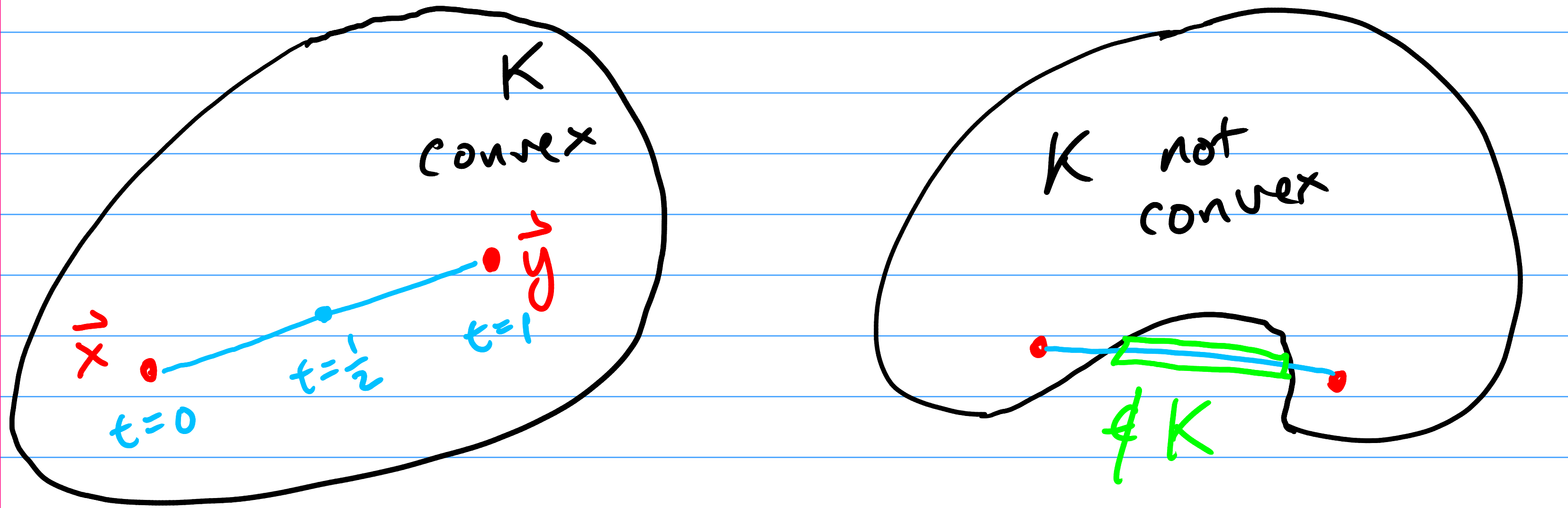
GD $\|\vec{w}_t - \vec{w}^*\| \leq \frac{1}{t}$

SGD $\|\vec{w}_t - \vec{w}^*\| \leq \frac{1}{\sqrt{t}}$

Convex sets:

__Defn__ A set $K \subseteq \mathbb{R}^d$ is __convex__ if

$$\underbrace{(1-t)\vec{x} + t\vec{y}}_{\text{line between } \vec{x} \text{ and } \vec{y}} \in K \quad \text{for any } \vec{x}, \vec{y} \in K, \quad 0 \le t \le 1$$

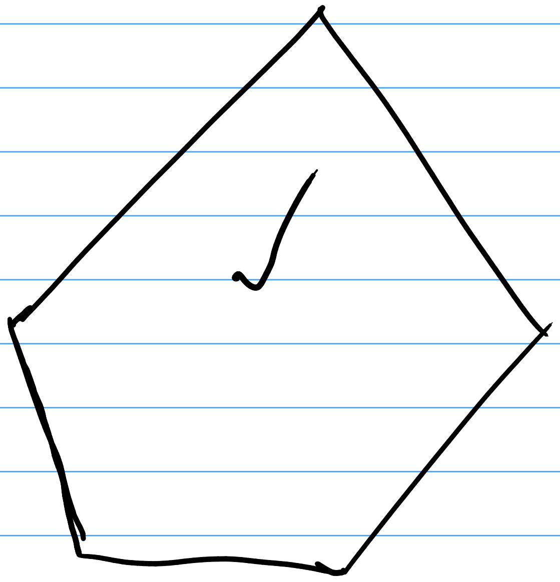


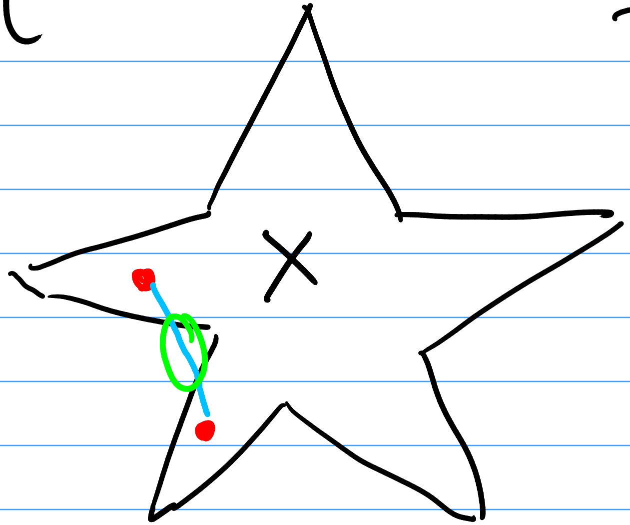
Is it convex?

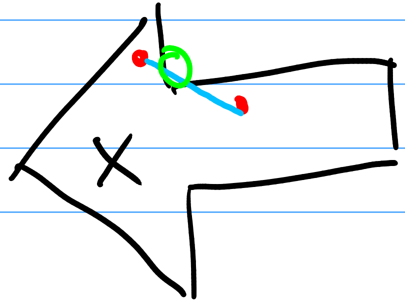
$$\mathbb{R}^d_{\geq 0} = \{\vec{x} : x_i \geq 0\}$$

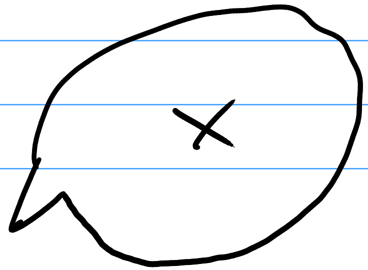
✓

$$\underbrace{(1-t)}_{\geq 0}\vec{x} + \underbrace{t}_{\geq 0}\vec{y} = \vec{z}$$

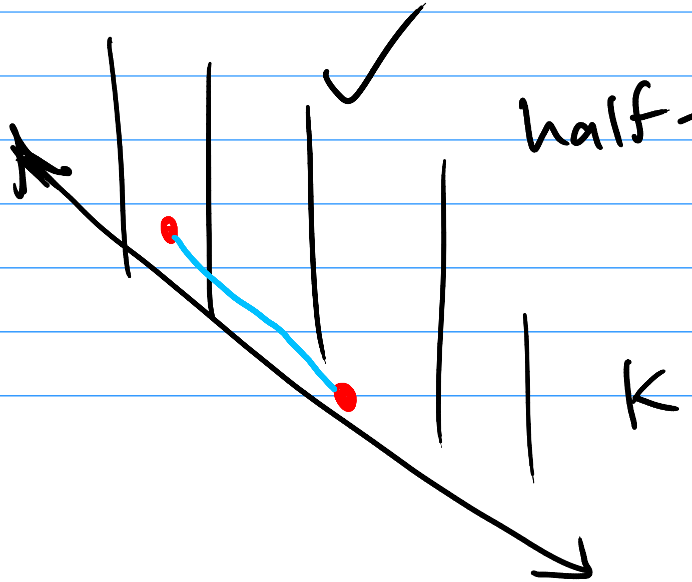








half-space

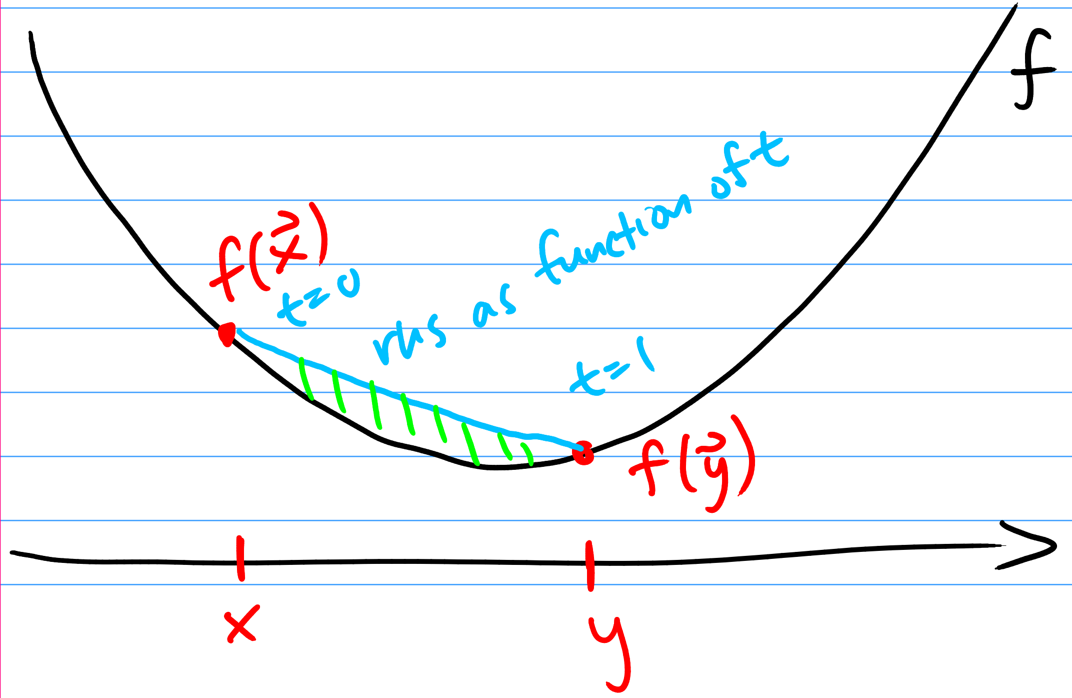


Convex functions

Defn $f : \mathbb{R}^d \to \mathbb{R}$ is <u>convex</u> iff

$$f(\underbrace{(1-t)\vec{x} + t\vec{y}}_{\text{any pt. on line}}) \le (1-t) f(\vec{x}) + t f(\vec{y})$$

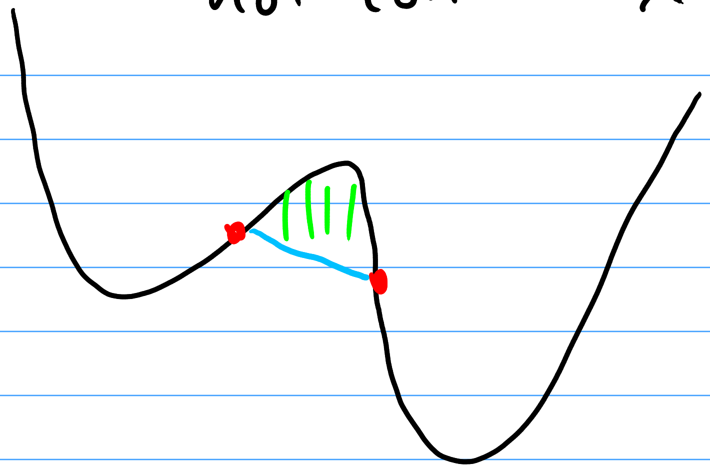
for any $\vec{x}, \vec{y} \in \mathrm{dom}(f)$, $0 \le t \le 1$



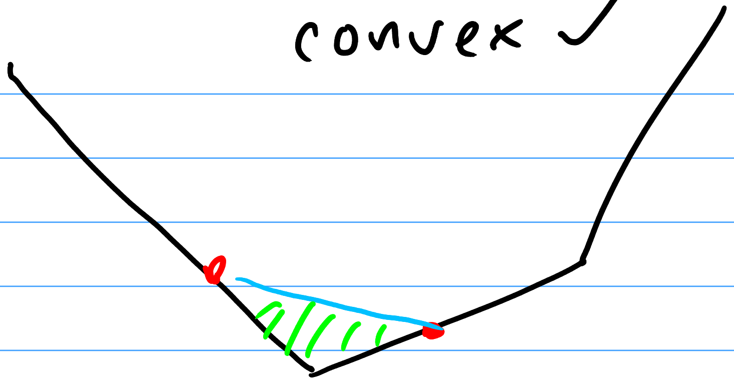
If f is convex:

"f lies underneath line segment connecting any two points"

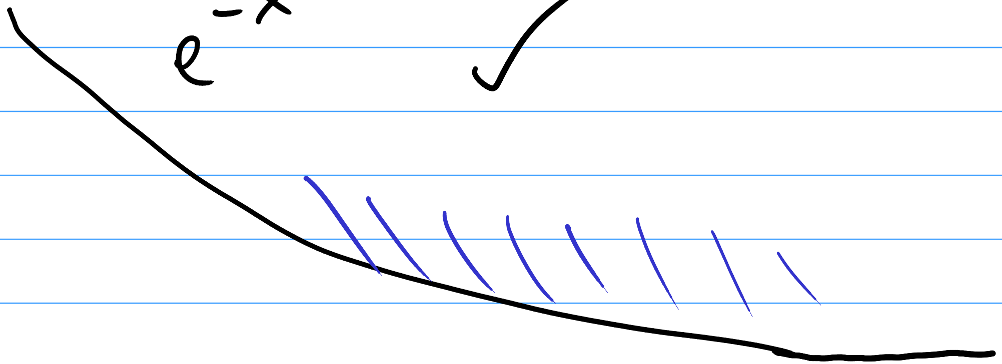
f not convex ✗

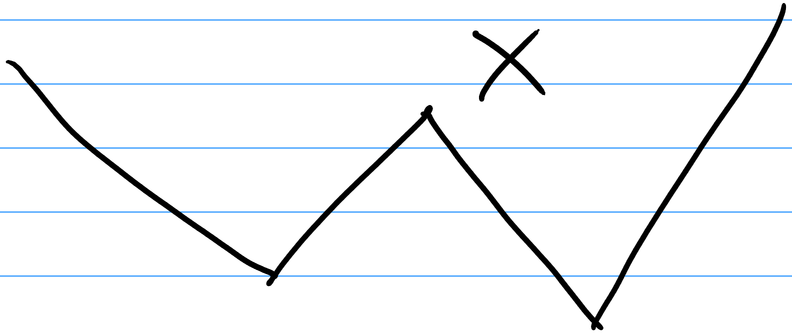


convex ✓

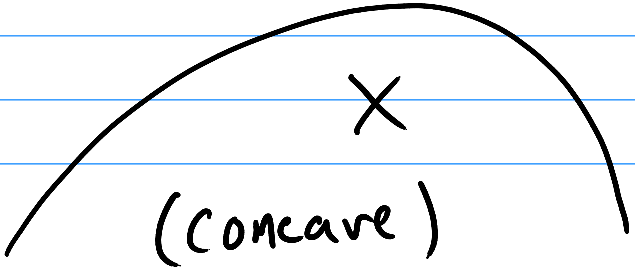


e^{-x} ✓





✗



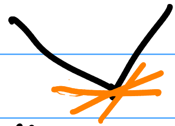
✗

(concave)

Why do we like them?

- all local minima are global
- algorithms are efficient and find these minima
  - GD, SGD

A4 {
  - coordinate descent
  - tricks for nonsmooth
    "proximal"
    "sub-gradient"

 ℓ_1-norm

  - accelerated versions $1/t \to 1/t^2$
    averaging

ex/ $\|X\vec{w}-\vec{y}\|^2$ convex

$L(\vec{w}) + \lambda R(\vec{w})$

$\|\vec{w}\|$ (real) norms

$f(\vec{w}) + g(\vec{w})$ f, g convex