MATH 462 09.29. Lecture 9

Biz HW 2 posted. due Tuesday

Midterm 1 next Weds (30 min in class)

Friday HW session 10–11 Lunch outside 11–11:30

Convex Analysis
Convex Optimization

C. Analysis foundation : define, prove rates

C. opt → algorithms

Goal 1. Apply Algorithm (Gradient / Stochastic Gradient Descent.)

to $\hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(h_w(x_i), y_i)$

2. understand it

$$\hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(h_w(x_i), y_i)$$

GD

$$w_{k+1} = w_k - h \nabla_w \hat{L}(w_k)$$

$h > 0$ learning rate

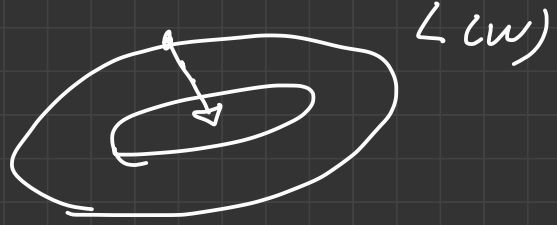
GD discretization of

$$\frac{d w(t)}{dt} = -\nabla \hat{L}(w(t))$$

GD ODE

h time step

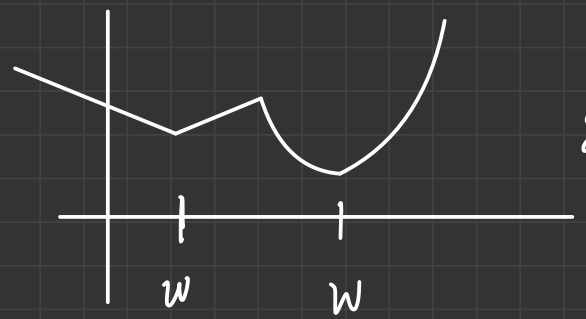
$$\frac{w(t_{k+1}) - w(t_k)}{t_{k+1} - t_k} \simeq -\nabla \hat{L}(w(t_k))$$



$t_{k+1} - t_k = h$

GD ① Abstract $\hat{L}(w)$ = black box fn gives $\nabla \hat{L}(w)$

Non-convex



go to local min

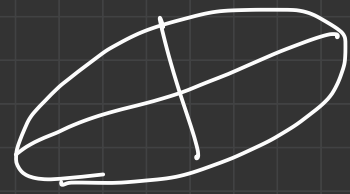
Bad h (too big) diverges. easy to fix: decrease h.

Analysis $\hat{L}(w)$ strongly convex & L-smooth
ALG
⇒ converge at Rate exponential.

Convergence Rate for GD.

parameter C = cond #

related to



Rate $\ell_k = \left(\frac{C}{C+1}\right)^k$

↳ error optimality gap

Case 1

$C = 1$ $\left(\frac{1}{2}\right)^k$ good!

Case 2

$C = 99$ $\left(\frac{99}{100}\right)^k$ slow

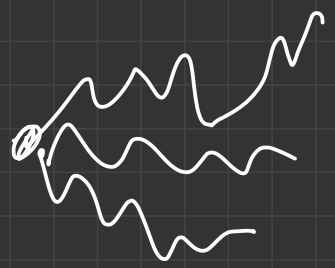
SGD understand it

3 defn all different with overlap

① Stochastic differential eqns

$$dx = v(x(t))dt + dW_t$$

↖ stoch. process



random paths.

(SDE → SGD)

$$dx = -\nabla L(x(t))dt + dW_t$$

not algorithm!

SGD Algorithms

ABSTRACT with noise Model (engineering)

$$x_{k+1} = x_k - \nabla L(x_k) + e_k$$

e_k: error term, random

could be mean zero.

$E[e_k] = 0$, $Var(e_k) = \sigma^2$

(Use to model for Mini-batch SGD for analysis)

PROVE convergence rate

Our setting SGD (mini-batch)

$$\hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(h_w(x_i), y_i)$$

DATA $S_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$

EX ImageNet 10^6 images ~ 1Mg. memory.

Algorithm BottleNeck

PAST data $\in \mathbb{R}$ $m = 1000$.

computers slow.

iteration count. bottleneck

Algorithms

Now computation cheap

Bottleneck DATA. Big.

Batch SGD.

$O(T)$ ← # iterations to get decent soln.

$O($# data points see per step$)$

SGD 10^7 steps

But each step 25% of $10^6 = m$ data points

e_k optimality gap $= O\left(\frac{1}{k}\right)$

What does $O(\frac{1}{k})$ look like?

Exp rate
1 step to error $\frac{1}{2}$
3 step error $\frac{1}{10}$
} FAST

"slow"
1000 step error $\rightarrow e/10$

$O(\frac{1}{k})$ rate

1. 9 steps $k=e \rightarrow e=0.1$
90 steps $e=\frac{1}{10} \rightarrow \frac{1}{100}=e$
900 steps $k=10$ 90
⋮ $\frac{1}{1000}$

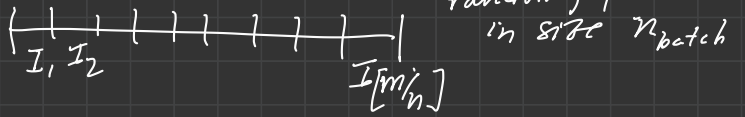
Mini-Batch SGD

$$\hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(h_w(x_i), y_i)$$

Bottleneck memory m too big.

$n = n_{batch}$ = dictated by memory

e.g. 256



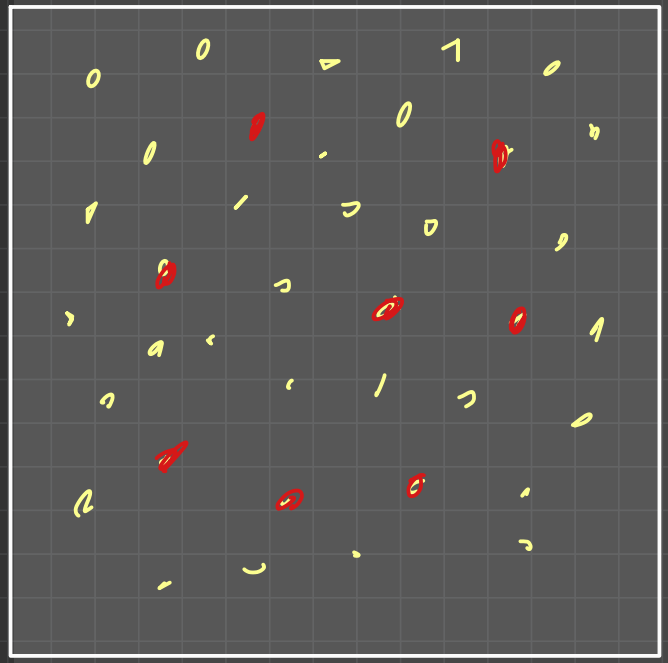
Before $\hat{L}(w, I) = \frac{1}{n_{batch}} \sum_{i \in I} \ell(h_w(x_i), y_i)$

(I: index set)

"GD"

$$w_{k+1} = w_k - h_k \nabla_w \hat{L}(w, I_k)$$

NOTE $h \to h_k$ I_k k^{th} partition.



$x_i \in [0,1]^2$

$i = 1, \dots, m$

Goal $w = \frac{1}{m} \sum_{i=1}^{m} x_i$

EL $\min_{w \in \mathbb{R}^2} \frac{1}{m} \sum_{i=1}^{m} \frac{\|w - x_i\|^2}{2}$

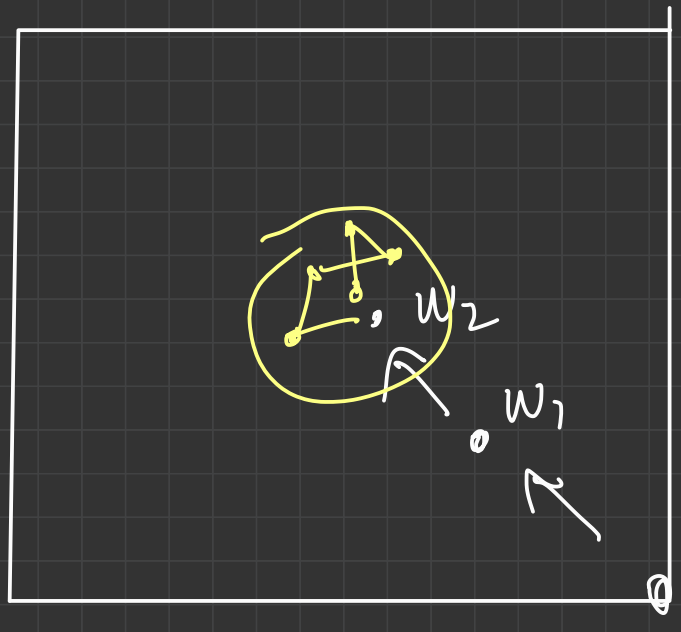
$\hat{L}(w)$

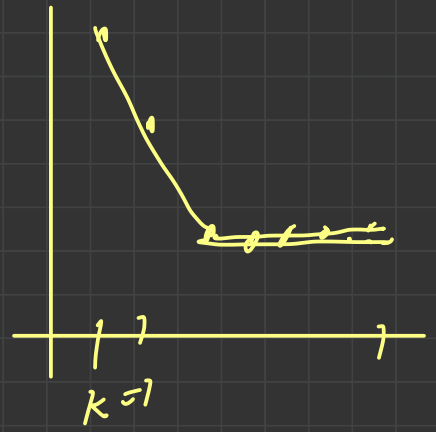
Define $\hat{L}(w, I)$ = as before

$n_{batch} = 8$

$\nabla L(w, I) \rightleftharpoons$ points $w - \frac{1}{8} \sum_{i \in I_b} x_i$

$$w_{k+1} = w_k - h_k \left(w - \text{average of } 8 \text{ random points} \right)$$





$h = \text{const}$

Convex Learning Problems (Reference: Ch 12 Shalev-Schwartz)

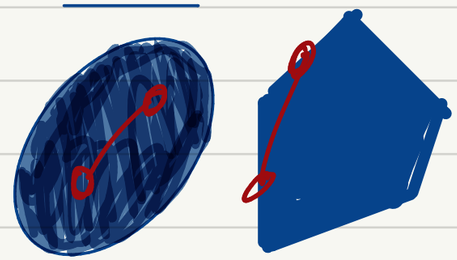
Definition (Convexity)

A set C in a vector space is convex if for any two vectors u, v in C the line segment between u and v is contained in C.

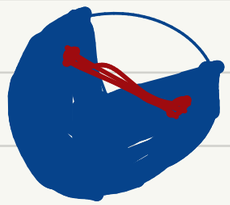
That is, for any $\alpha \in [0,1]$

$$\alpha u + (1-\alpha) v \in C$$

Convex



non convex



Note $f : \mathbb{R}^n \to \mathbb{R}$

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

gradient

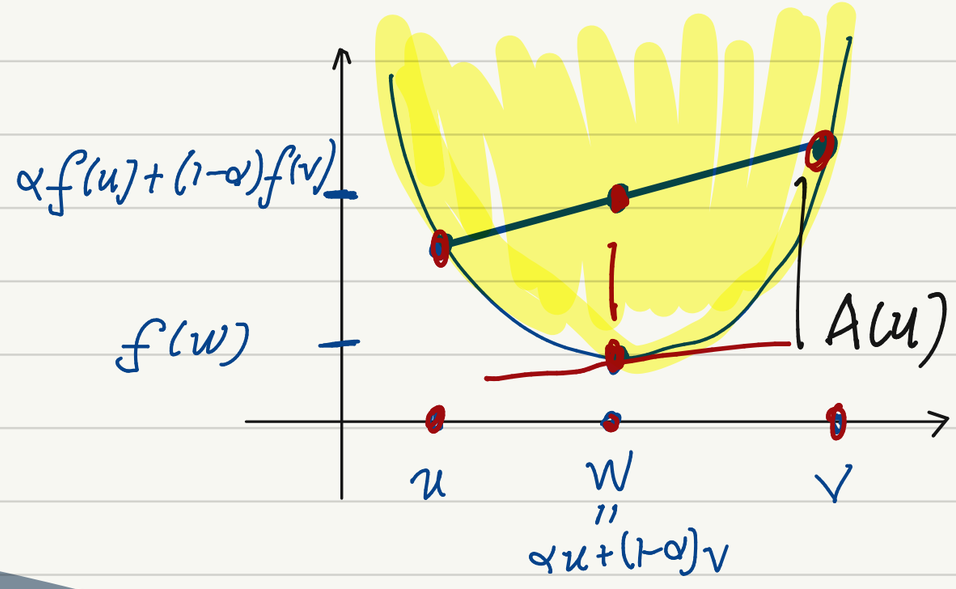
vector pointing in direction of greatest increase for $f(x)$

Defn (Convex function)

A function from a convex set C, $f: C \to \mathbb{R}$ is convex if, for every $u, v \in C$ and $\alpha \in [0,1]$

$$f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v)$$

w



Exercise

$A(u) = g \cdot (u - w) + f(w)$

Check $A(w) = g \cdot 0 + f(w) = f(w)$.

check $u = 7$ $\quad 49/2 = f(7)$

$A(4) = 3 \cdot 4 + 9/2$

Gap at 7 is $24.5 - 17.5$?

$f(w) = w^2/2$

$w = 3$

$A(u) = 3(u-3) + 9/2$

Properties of convex functions

Ref Boyd Convex Opt.

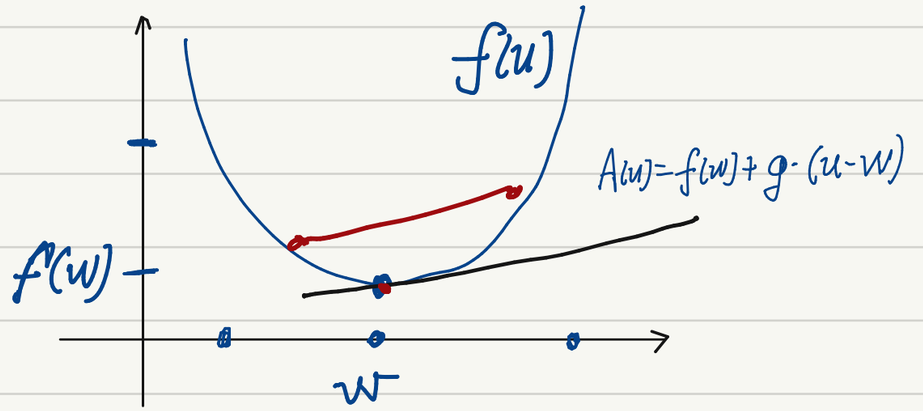
- Every local minimum is a global minimum (exercise)
- supporting hyperplane (tangent) property.

Defn Let $f: C \to \mathbb{R}$ where C is an open, convex set define, for vectors g, the affine function

$$A_g(u) = f(w) + g \cdot (u - w)$$

$A(w) = f(w)$

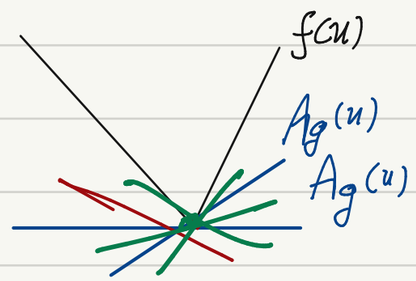
$g = \nabla f(w)$



Lemma: the function $f: C \to \mathbb{R}$, C open, convex, is convex iff for every $w \in C$, there exists g such that

$$f(u) \geq A_g(u) \quad \text{for all } u \in C \qquad (*)$$

$A_g(w) = f(w)$



A vector g that satisfies $*$ is called a subgradient of f at w. $\partial f(w) = \{\text{all subgradients of } f \text{ at } w\}$

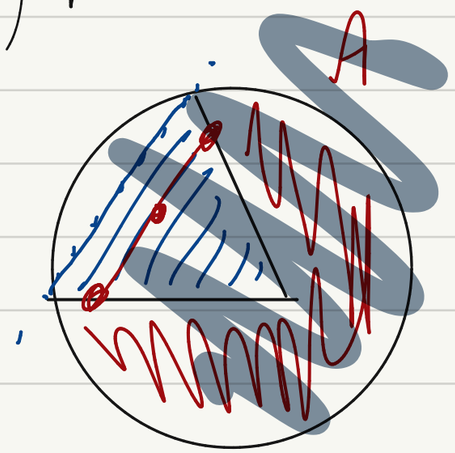
$f(x) = |x|$

$\partial f(0) = \{ [-1, 1] \}$

Definition Convex Hull (of a set) A

smallest convex set containing A

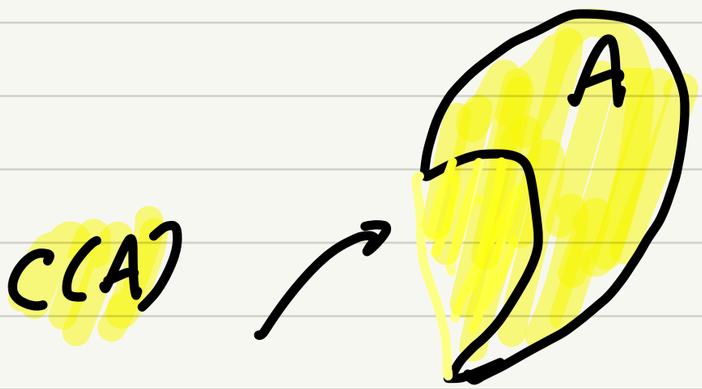
$$co(A) = \left\{ \bigcap C \mid C \supset A, \; C \text{ convex} \right\}$$



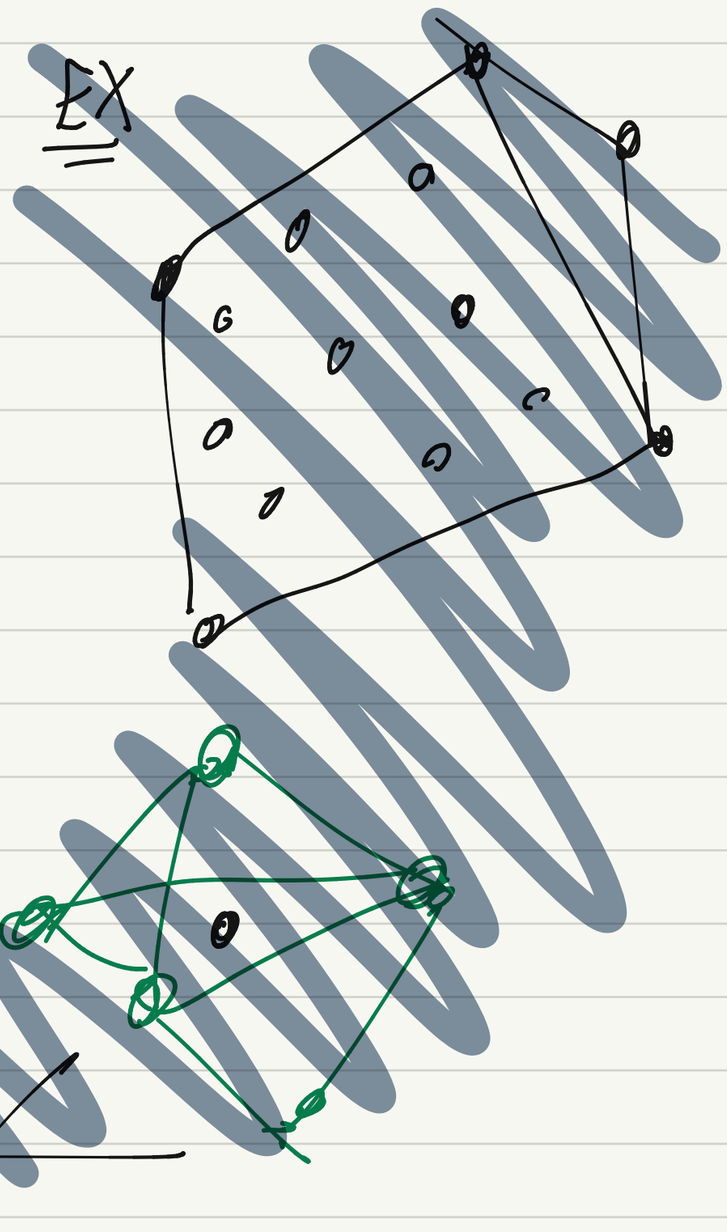
can write every $c \in co(A)$ as $A \subset \mathbb{R}^d$

$$c = \left\{ \sum_{i=1}^{n} w_i a_i \right. \quad \vec{w} \text{ weight vector} \quad \left. \sum_{i=1}^{n} w_i = 1 \quad w_i \geq 0 \right\}$$

Mazur's Lemma



EX

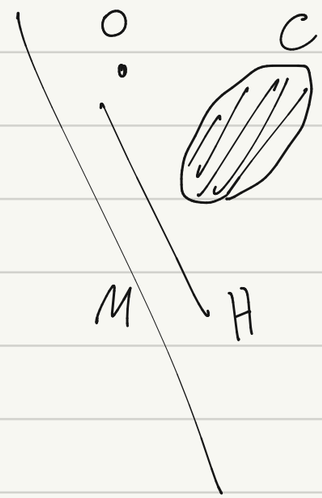


Separation Thm

difference between finite & ∞ dim.

V vector space (could be ∞-dimensional)

[e.g. ℓ^2 $\{(x_1,\dots,x_n,\dots)\}$, $\|x\|_2^2 = \sum x_i^2 < \infty$]



Hahn Banach Thm

V vector space

C open convex set non-empty

M non-empty affine subspace, $C \cap M = \emptyset$

Then there exists a separating hyperplane H, given by $A(x)$

$A(x) = a \cdot x + b$

separates if $A(c) \geq 0 \quad \forall c \in C$

$A(a) \leq 0 \quad \forall a \in A$

(non strict)

strict: > 0 / < 0

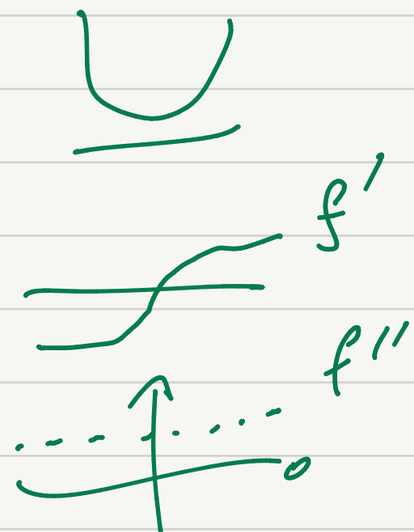
Analysis

More properties of convex functions.

Lemma Let $f:\mathbb{R}\to\mathbb{R}$ be twice differentiable. 1

the following are equivalent (TFAE)

- 1. f is convex
- 2. f' is monotonically non-decreasing
- 3. f'' is non-negative.



$f:\mathbb{R}^n\to\mathbb{R}$ C^2. TFAE

1. f is convex
3. $D^2f(x)$ is non-negative definite
2. $(\nabla f(x)-\nabla f(y))\cdot(x-y)\geq 0$

note: 2 extends to non-differentiable case replacing gradient with subgradients

3. $D^2f(x)=\left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]$

$D^2f(x)\geq 0$ pos. definite

cond. local min

crit pt. $\nabla f(x)=0$

Algebra of Convex fns

Examples of Convex functions:

- $Q(x) = x^T P x + bx$ quadratic where P is positive definite

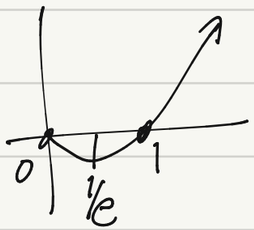
- $f(x) = x^2/2$

- $f(x) = x \log x$ (on $x > 0$)

  check $f'(x) = \frac{x}{x} + \log x$

  $= 1 + \log x$ increasing

  $f''(x) = \frac{1}{x} > 0$



$\log x = -1$

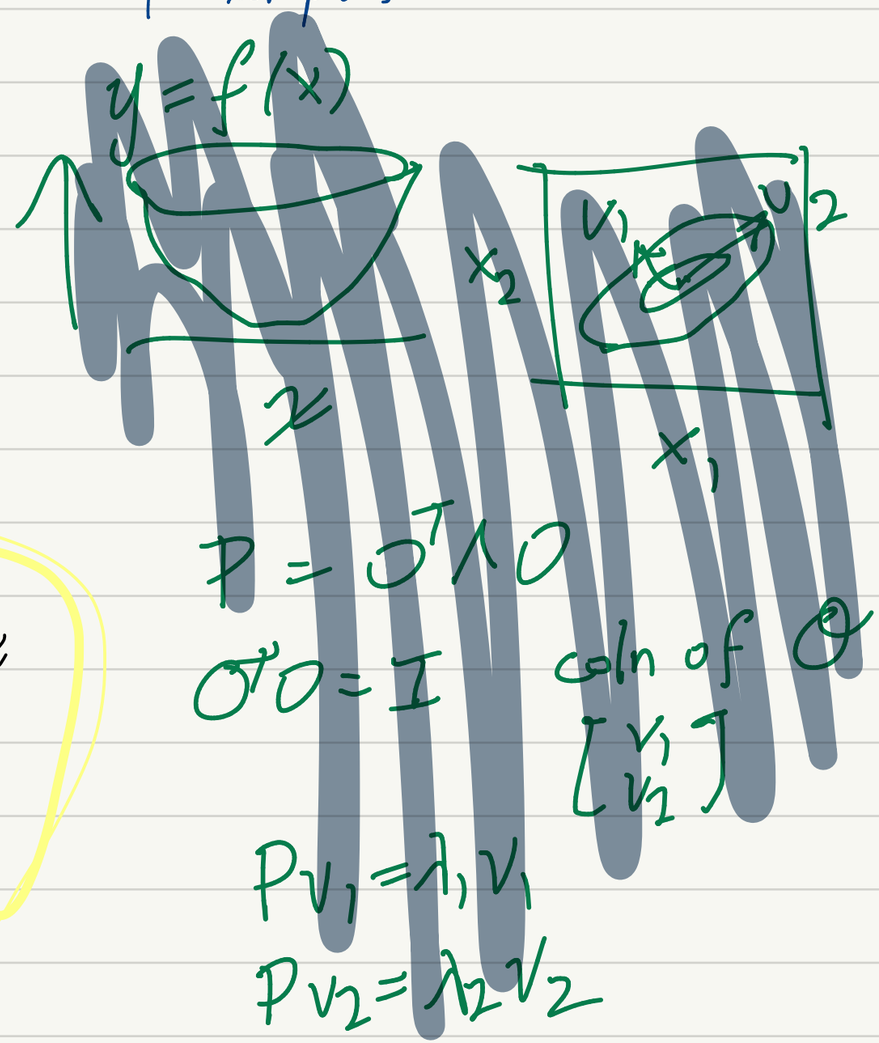
$x = \frac{1}{e}$

& sum of convex is convex

properties of convex functions

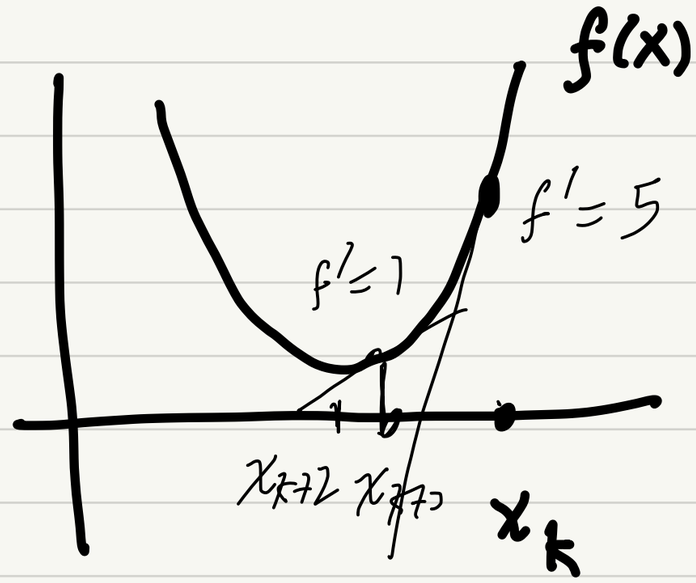
- if g is convex and $f(x)$ is affine, then $g \circ f$ is convex
- max of convex functions is convex

See Boyd & Vandenberg for more properties & examples



Key properties for convex optimization.

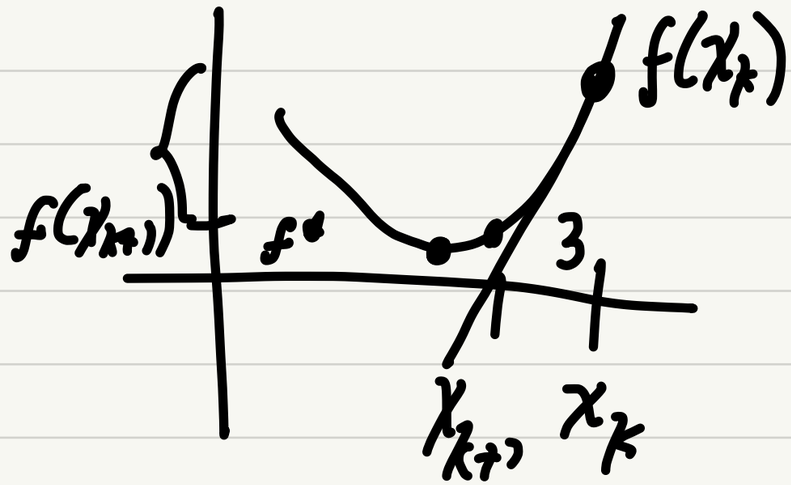
Lipschitz & "Smooth"



Defn Let $f: C \subset \mathbb{R}^d \mapsto \mathbb{R}^k$. Suppose

$$\| f(w_1) - f(w_2) \| \leq \rho \| w_1 - w_2 \| \quad \forall w_1, w_2 \in C$$

then say f is ρ-Lipshitz continous over C.



Note if f is $\nabla f(x)$ exists & $\| \nabla f(x) \| \leq \rho \quad \forall x \in C$
then f is ρ-Lipschitz.

Why? Mean-Value-Thm.

$$f(w_1) - f(w_2) = \nabla f(z) \cdot (w_1 - w_2)$$

for some z.