MATH466/MATH766
Math of machine learning
02/03-02/05 Lecture 7-8 Convex Optimization

References:

- Ch1 of Convex Optimization by Stephen Boyd and Lieven Vandenberghe
- https://sites.math.washington.edu/~burke/crs/408f/notes/nlp/unoc.pdf

Todays contents:

- Optimization overview
- Existence of minimizer
- Convexity and global minimizer, unique minimizer
- First and second order conditions of convex functions

Important concepts:

- coercivity
- convexity

Recommend reading:

- Ch 1.1, 1.3, 1.4 of Convex Optimization by Stephen Boyd and Lieven Vandenberghe
-

Warm up.

If $x_m < a$ for $m=1,2,\cdots$, then $\lim_{m\to+\infty} x_m$ ______ a

〃 $x_m \le a$ 〃 〃 〃, then $\lim_{m\to+\infty} x_m$ ______ a.

0. Optimization Overview

- A general problem formulation

$$\min_{x} f_0(x)$$
$$\text{subject to } f_i(x) \le b_i, \quad i = 1, \cdots, m \qquad (*)$$

optimization variable : $x = (x_1, \cdots, x_n)$

objective function : $f_0 : \mathbb{R}^n \to \mathbb{R}$

constraint functions : $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \cdots, m$

bounds : $b_i \in \mathbb{R}$, $i = 1, \cdots, m$

x^* is a global optimizer to $(*)$ if

① $f_i(x^*) \le b_i$ (feasible)

and ② for any feasible z, $f_0(z) \ge f_0(x^*)$ (global optimal)

x^* is a local optimizer to $(*)$ if

① $f_i(x^*) \le b_i$ (feasible)

and ② there exists $r > 0$

s.t. for any feasible z and $\|z - x\| < r$,

$f_0(z) \ge f_0(x^*)$ (local optimal)

① linear programming

When f_0 and f_i $(i=1,\cdots,m)$ are linear

" Solving (most) linear programs is a mature technology."

② convex programming

When f_0 and f_i $(i=1,\cdots,m)$ are convex

local $\xleftrightarrow{\text{cvx}}$ global $\xrightarrow{\text{cvx}}$ uniqueness

③ nonlinear programming

When f_0 or f_i $(i=1,\cdots,m)$ is not linear,

but not known to be convex.

This problem is not easy. In general

technics in cvx opt can help → local optimization : fast

give up seeking the optimal x, ...

global optimization : seeking global optimal

sacrifice efficiency

1. Existence of minimizer

$$\min_{x \in C} f(x)$$

1.1 When C is empty, solution doesn't exist

1.2 Weierstrass Extreme Value Theorem

- Continuous function: if $\lim_{m \to +\infty} x_m = x$, then $\lim_{m \to +\infty} f(x_m) = f(x)$
- Compact set (by Heine-Borel Thm): $S \subseteq \mathbb{R}^n$ is compact iff

  S is bounded: $\exists\, r > 0$ s.t. $\forall\, x \in S$, $\|x\| < r$

  and S is closed: for any x, if there exist $\{x_m\} \subseteq S$, $\lim_{m \to +\infty} x_m = x$

  then $x \in S$

  e.g. compact: $[-1, 1]$, $[0,1] \cup [2,3]$

  non-compact: $[0, +\infty)$ b/c not bounded

  $[0, 1)$ b/c not closed

- Weierstrass Extreme Value Theorem

  If $f: S \to \mathbb{R}$, S is compact and f is continuous on S,

  then there exists $x^* \in S$ s.t. $f(x^*) \le f(x)$ holds for any $x \in S$

  i.e. cts func on compact set attains its extreme value on the set.

  - Rigorous proof of this thm can be found in Real Analysis or advanced (proof-based) calculus textbook.

  - Continuity is necessary, o.w. counter e.g.

    $$f: [-\tfrac{\pi}{2}, \tfrac{\pi}{2}] \to \mathbb{R} \quad x \mapsto \begin{cases} 0 & \text{if } x = \pm\frac{\pi}{2} \\ \tan x & \text{o.w.} \end{cases}$$

  - compactness is necessary, o.w. counter e.g.

    $(-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$, $x \mapsto \tan x$

    $[1, +\infty) \to \mathbb{R}$, $x \mapsto \frac{1}{x}$

Recall: Many questions we saw in Supervised Learning are defined on $\mathbb{R}$, which is not compact.

How to fix this?

1.3. Coercive functions

- Def. (coercive) A function $f: \mathbb{R}^n \to \mathbb{R}$ is coercive if for every sequence $\{x_m\} \subseteq \mathbb{R}^n$, such that $\lim_{m\to+\infty} \|x_m\| = +\infty$ it holds that $\lim_{m\to+\infty} f(x_m) = +\infty$
- Prop. A continuous function is coercive iff for any $\alpha \in \mathbb{R}$, the set $S_\alpha := \{x \in \mathbb{R}^n \mid f(x) \le \alpha\}$ is compact

Pf. "$\Rightarrow$": "closedness" (taking limit preserves sign)

for any $\alpha \in \mathbb{R}$, if $\{x_m\} \subseteq S_\alpha$, $\lim_{m\to+\infty} x_m = x$

then by continuity $f(x) = \lim_{m\to+\infty} f(x_m) \le \alpha$, $x \in S_\alpha$

"boundedness" (by contradiction)

if there exists $\alpha_0 \in \mathbb{R}$ s.t. S_{α_0} is not bounded

then there exists a sequence $\{x_m\}$

s.t. $\lim_{m\to+\infty} \|x_m\| = +\infty$ and $\lim_{m\to+\infty} f(x_m) \le \alpha_0$

Contradiction to f being coercive.

"$\Leftarrow$" left as exercise

- Coercivity + Continuity $\Rightarrow$ Existence of global optimizer

Prop. Let $f: \mathbb{R}^n \to \mathbb{R}$ be continuous and coercive, then there exists $x^* \in \mathbb{R}^n$ s.t. $f(x^*) \le f(x)$ for any $x \in \mathbb{R}^n$.

Pf left as exercise (use the above prop. and W. extreme value thm)

2. Global minimizer and Uniqueness from Convexity

Q: local information $\xrightarrow{?}$ global minimizer

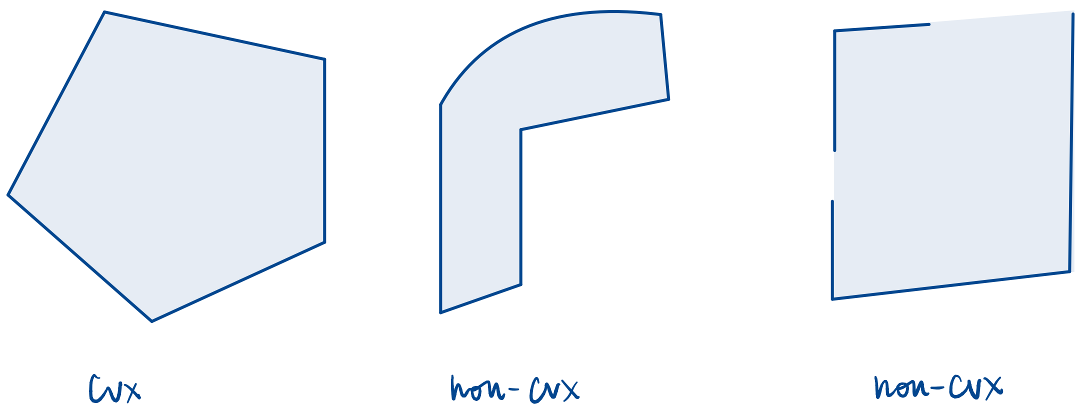
2.1 Convex set and convex function

- Def. (convex set) A set C is convex if

  for any $x_1, x_2 \in C$, and any $\theta \in [0,1]$, $\theta x_1 + (1-\theta) x_2 \in C$

  the line segment between any two points in C lies in C.

e.g.



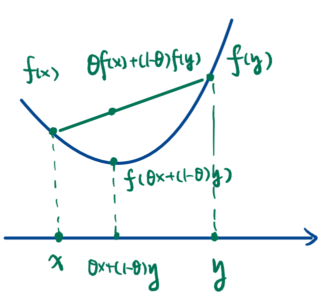
- Def. (convex function) A function f is convex if

  $\mathrm{dom}(f)$ is a convex set

  and for any $x, y \in \mathrm{dom}(f)$, $\theta \in (0,1)$,

  it holds that $f((1-\theta)x + \theta y) \le (1-\theta) f(x) + \theta f(y)$

Interpretation



A function f is strictly convex if

$\mathrm{dom}(f)$ is a convex set

and for any $x, y \in \mathrm{dom}(f)$, $\theta \in (0,1)$, $x \neq y$

it holds that $f((1-\theta)x + \theta y) < (1-\theta) f(x) + \theta f(y)$

Examples

- $f: \mathbb{R} \to \mathbb{R}$, $x \mapsto e^{ax}$, $a \in \mathbb{R}$

  $f: \mathbb{R}_{++} := \{x \in \mathbb{R} \mid x > 0\} \to \mathbb{R}$, $x \mapsto -\log x$

  $f: \mathbb{R}_{++} \to \mathbb{R}$, $x \mapsto x \log x$

- $f: \mathbb{R}_{++} \to \mathbb{R}$, $x \mapsto x^a$, $a \geq 1$ or $a \leq 0$

  $f: \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|^p$, $p \geq 1$

- $f: \mathbb{R}^n \to \mathbb{R}$. $x \mapsto \|x\|$ where $\|\cdot\|$ is a norm on $\mathbb{R}^n$

  $f: \mathbb{R}^n \to \mathbb{R}$ $\quad x \mapsto \max\{x_1, \cdots, x_n\}$

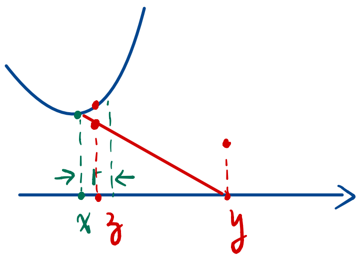
  $f: \mathbb{R}^n \to \mathbb{R}$ $\quad x \mapsto \log(e^{x_1} + e^{x_2} + \cdots + e^{x_n})$ differentiable approximation of $\max\{x_1, \cdots, x_n\}$.

  $f: \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}$ $\quad (x, y) \mapsto x^2/y$

Consider $\min_{x \in C} f(x)$

2.2. local + convex $\rightarrow$ global

Thm 1. Let $f: \mathbb{R}^n \to \mathbb{R}$ be convex. If x^* is a local minimizer of f then x^* is a global minimizer of f.



Pf by contradiction.

Assume there exists y s.t. $f(x^*) > f(y)$

Let $z := x + t(y-x) = (1-t)x + ty$. $t \in (0,1]$

then $\|z-x\| = t\|y-x\|$, $f(z) \le (1-t)f(x) + tf(y) < f(x)$

Contradiction to x^* being a local minimizer $\square$

2.3. strictly convex $\rightarrow$ uniqueness

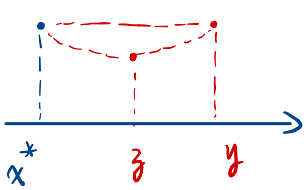
Thm 2. Let $f: \mathbb{R}^n \to \mathbb{R}$ be strictly convex. If x^* is a global minimizer of f then x^* is the unique minimizer of f.

Pf by contradiction

Assume there exists $y \neq x^*$ s.t. $f(y) = f(x^*)$

For any $z = (1-t)x^* + t(y)$, $t \in (0,1)$

$f(z) < (1-t)f(x^*) + tf(y) = f(x^*)$



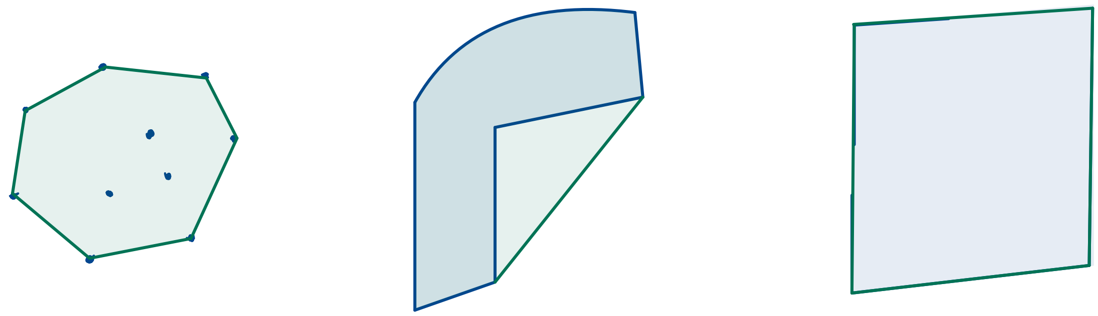
Contradiction to x^* being a global minimizer. $\square$

3. More about convexity

3.1. convex hull

$\mathrm{conv}\, C := \{ \theta_1 x_1 + \cdots + \theta_k x_k \mid x_i \in C,\ \theta_i \geq 0,\ i=1,\cdots,k,\ \theta_1+\cdots+\theta_k = 1 \}$

e.g.



convex hull $\mathrm{conv}\, C$ is always convex.

3.2 Operations that preserve convexity of a set

- S_1, S_2 cvx $\Rightarrow S_1 \cap S_2$ cvx   How about $S_1 \cup S_2$?
- $S \subseteq \mathbb{R}^n$ cvx, $f: \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax+b$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$

  $\Rightarrow f(S) = \{ f(x) \mid x \in S \} \subseteq \mathbb{R}^m$ cvx

3.3 Operations that preserve convexity of a function

- $f_1, \cdots, f_m : \mathbb{R}^n \to \mathbb{R}$ convex, $w_1, \cdots, w_m \geq 0$

  $\Rightarrow f := \sum_{i=1}^{m} w_i f_i$ is convex
- $f: \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$, f is convex

  $\Rightarrow g: \mathbb{R}^m \to \mathbb{R}$, $x \mapsto f(Ax+b)$ is convex
- $f_1, \cdots, f_m : \mathbb{R}^n \to \mathbb{R}$ convex

  $\Rightarrow f := \max\{f_1, \cdots, f_m\}$ is convex
- $h: \mathbb{R} \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}$, $f = h \circ g$

  g convex, h convex and nondecreasing $\Rightarrow f$ convex

3.4. First and Second order conditions of convex functions

intuition for $f: \mathbb{R} \to \mathbb{R}$ (not rigorous at all)

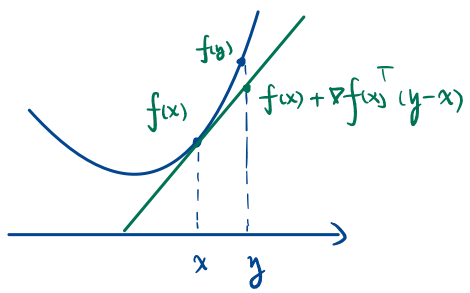
$$f \text{ cvx} \iff f' \text{ non-decreasing} \iff \text{if } y > x \quad \frac{f(y) - f(x)}{y - x} \geq f'(x) \iff f''(x) \geq 0$$

i.e. $\quad f(y) \geq f(x) + f'(x)(y-x)$

- Prop. 1. Suppose f is differentiable

  Then f is cvx iff $\mathrm{dom} f$ is convex and for all $x, y \in \mathrm{dom} f$

  $$f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y-x)}_{\text{first-order Taylor approximation}} \text{ holds}$$



For cvx func. the first-order Taylor approximation (local information) is a global underestimator of the function.

Pf for $f: \mathbb{R} \to \mathbb{R}$:

"$\Rightarrow$" f is cvx $\Rightarrow$ $\mathrm{dom} f$ is cvx, thus for any $x, y \in \mathbb{R}$, $0 < t \leq 1$

$$x + t(y-x) \in \mathrm{dom} f$$

By convexity $\quad f(x + t(y-x)) \leq (1-t) f(x) + t f(y)$

Therefore $\quad f(y) \geq f(x) + \dfrac{f(x + t(y-x)) - f(x)}{t}$ for any t

Taking $t \to 0$ yields $\quad f(y) \geq f(x) + f'(x)(y-x)$.

"$\Leftarrow$" For any $x \neq y$, $0 \leq \theta \leq 1$, $z = \theta x + (1-\theta) y$

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

$$\Rightarrow \theta f(x) + (1-\theta) f(y) \geq f(z) + f'(z)[\underbrace{\theta(x-z) + (1-\theta)(y-z)}_{=0}] \qquad \square$$

Proof for high-dimension left as exercise

(hint: choose a direction and convert to 1D)

Prop 2. If f is twice-differentiable then f is cvx iff

dom f is cvx and $\nabla^2 f(x) \succeq 0$ for all $x \in \text{dom}(f)$

Pf for $f: \mathbb{R} \to \mathbb{R}$:

"$\Rightarrow$" By prop. 1, for any $x, y \in \text{dom}(f)$, $x \neq y$

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

$$\Rightarrow \quad f'(x)(y-x) \leq f(y) - f(x) \leq f'(y)(y-x)$$

$$\Rightarrow \quad \frac{f'(y) - f'(x)}{y-x} \geq 0 \quad \xrightarrow{\lim y \to x} \quad f''(x) \geq 0.$$

"$\Leftarrow$" By mean value version of Taylor's theorem

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2} f''(z)(y-x)^2, \quad \text{for some } z \in [x,y]$$

$$\geq f(x) + f'(x)(y-x) \quad \text{since } f''(z) \geq 0$$

- f is strictly convex

$\Leftrightarrow$ dom f is convex and for any $x, y \in \text{dom} f$, $x \neq y$, $f(y) > f(x) + \nabla f(x)^T (y-x)$

$\Leftarrow$ dom f is convex and for any $x \in \text{dom} f$, $\nabla^2 f(x) \succ 0$

Three main approaches to prove the convexity of a function

1. By definition

e.g. $\|\cdot\|$ norm. $\quad \|\theta x + (1-\theta) y\| \leq \|\theta x\| + \|(1-\theta) y\|$ (tri inequ)

$$= \theta \|x\| + (1-\theta) \|y\| \quad \text{(homo)}$$

2. Check 2nd order condition

e.g. $f(x) = \frac{1}{2} \|x\|_2^2$, $\nabla^2 f(x) = I \succeq 0 \quad \forall x$

3. Use operators that preserve convexity

e.g. $g(x) = \frac{1}{2} \|Ax + b\|_2^2 = f(Ax+b)$ where $f = \frac{1}{2} \|\cdot\|_2^2$.