ML Theory Lecture 10

Just finished representation "chapter"; now optimization.

Plan: convexity overview; descent methods; Maurey; consistency of convex losses.

Maybe also non-convex?

General convexity remarks:

- Why? (a) convex losses (seem) here to stay;
  (b) beautiful & ~~[illegible]~~ pervasive.
  well outside opt.
  ↑ Since NN nonconvex...

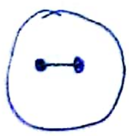
- This presentation: stressing simultaneous geometric/algebraic view; also duality.

- Refs: * Hiriart-Urruty / Lemaréchal "Fundamentals..." [HULL]
  * Borwein-Lewis "convex analysis..."; [BL]
  note, many key things in exercises (e.g., 3.3.?.f...)
  * Rockafellar "Convex Analysis" [ROC]

Convex Sets

Defn. S convex when $\{x, y\} \subseteq S \implies [x,y] \subseteq S$

$[x,y] = \{\alpha x + (1-\alpha) y : \alpha \in [0,1]\}$



Convex hulls & combinations

Convex comb: z is a <u>convex combination</u> of $(x_1, \dots, x_n)$ when $\exists\, (\alpha_1, \dots, \alpha_n)$, $\vec{\alpha} \in \Delta$ s.t. $\sum \alpha_i x_i = z$.

Δ simplex: $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$.

Convex hull:

$$\operatorname{conv}(S) = \bigcap_{\substack{C \supseteq S \\ C \text{ convex}}} C$$

[HULL Prop A.1.3.4]

$$= \{ z : z \text{ a convex combination of elements of } S \}$$

$$= \bigcup_{n \geq 1} \left\{ \sum_{i=1}^{n} \alpha_i x_i : (x_1, \dots, x_n) \in S, \vec{\alpha} \in \Delta \right\}.$$

<u>Prop</u> C is convex iff contains all convex combinations ($C = \operatorname{conv}(S)$). [HULL Prop A.1.3.3]

<u>Pf.</u> Induction on defn. //

Projection on convex sets

Remark. We will often need <u>closed</u> sets. Issues with closedness, interiors, etc. can be subtle; I'll try not to get too bogged down but feel free to ask.

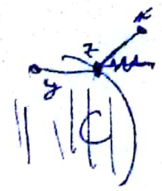
<u>Defn</u> $\Pi_C(x) = \operatorname{argmin} \{ \frac{1}{2} \|y - x\|^2 : y \in C \}$.

Note: defn <u>nonsense</u> as stated!

Theorem Suppose C closed convex nonempty. [HULL Section A.3.1]

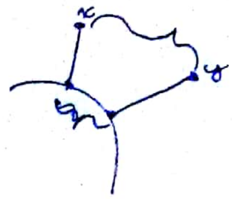
① $\Pi_C(x)$ exists, unique.

② $z = \Pi_C(x)$ iff $\forall y \ \langle y-z, x-z\rangle \leq 0$.



③ $\|\Pi_C(x) - \Pi_C(y)\|^2 \leq \langle \Pi_C(x) - \Pi_C(y), x-y\rangle$

and Π_C nonexpansive: $\|\Pi_C(x) - \Pi_C(y)\| \leq \|x-y\|$

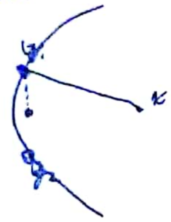


Pf. (Maybe skipped in class.) [Almost identical to HULL A.3.1]

① Exists: pick any $w \in C$, suffices to consider $\inf_{y\in C}\{\frac{1}{2}\|x-y\|^2 : \|x-y\| \leq \|x-w\|\}$

cont function over compact, min exists.

Unique:



[Draw ℓ_2 ball around x;
$\frac{1}{2}(y_1+y_2)$ closer than $y_1 \neq y_2$!]

In symbols, first note

$$\|A+B\|^2 = \|A\|^2 + \|B\|^2 + 2\langle A, B\rangle$$
$$\|A-B\|^2 = \|A\|^2 + \|B\|^2 - 2\langle A, B\rangle$$
$$\Rightarrow \|A+B\|^2 = 2\|A\|^2 + 2\|B\|^2 - \|A-B\|^2$$

$\Rightarrow$ given claimed projections z_1, z_2, midpoint $z = \frac{1}{2}(z_1+z_2)$ satisfies

$$\frac{1}{2}\|x-z\|^2 = \frac{1}{8}\|(x-z_1)+(x-z_2)\|^2 = \frac{1}{4}\|x-z_1\|^2 + \frac{1}{4}\|x-z_2\|^2 - \frac{1}{8}\|z_1-z_2\|^2$$

$\|z_1 - z_2\| = 0$ iff $z_1 = z_2 = z$.

② ($\Rightarrow$) for any w,

$$0 \leq \|x-w\|^2 + \|x-z\|^2$$
$$= \|x-z+z-w\|^2 + \|x-z\|^2$$
$$= \|z-w\|^2 + 2\langle x-z, z-w\rangle$$
$$\Rightarrow \langle x-z, w-z\rangle \leq \tfrac{1}{2}\|z-w\|^2$$

Thus let y given, set $w := \alpha y + (1-\alpha)z$ for $\alpha \in [0,1]$:

$$\alpha\langle x-z, y-z\rangle = \langle x-z, w-z\rangle \leq \tfrac{1}{2}\|z-w\|^2 = \tfrac{\alpha^2}{2}\|z-y\|^2$$

$\xRightarrow{\alpha \downarrow 0}$ $\langle x-z, y-z\rangle \leq 0$.

($\Leftarrow$) $0 \geq \langle y-z, x-z\rangle = \langle y-x+x-z, x-z\rangle = \langle y-x, x-z\rangle + \langle x-z, x-z\rangle$
$\geq -\|y-x\|\,\|x-z\| + \|x-z\|^2$

$\Rightarrow \|x-z\| \leq \|y-x\|$ (when $x \neq z$; if $x = z$, ~~directly~~ done.)

③ Instantiating previous twice,

$$\langle \Pi_C(y) - \Pi_C(x), x - \Pi_C(x)\rangle \le 0$$
$$+ \quad \langle \Pi_C(x) - \Pi_C(y), y - \Pi_C(y)\rangle \le 0$$
$$\langle \Pi_C(x) - \Pi_C(y), y - x + (\Pi_C(y) - \Pi_C(x))\rangle \le 0$$

$$\Rightarrow \|\Pi_C(x) - \Pi_C(y)\|^2 \le \langle \Pi_C(x) - \Pi_C(y), x - y\rangle. \quad \text{(first part)}$$
$$\le \|\Pi_C(x) - \Pi_C(y)\| \cdot \|x - y\| \quad \text{(second part by division).}$$

Building from this, get separating hyperplanes: [HULL Section A.4.1]

* C closed nonempty convex, $x \notin C$:
  $\exists\, s$ s.t. $\langle s, x\rangle > \sup_{y \in C} \langle s, y\rangle$.

(Pf. take $s = x - \Pi_C(x)$.)

* C_1, C_2 closed nonempty convex, ~~at least one~~ C_2 bounded;
  $\exists s$ $\sup_{y \in C_1} \langle s, y\rangle < \min_{x \in C_2} \langle s, x\rangle$

* Supporting hyperplanes; [HULL Section A.4.2]
  C is closed nonempty convex $\Leftrightarrow$ $C = \bigcap_{H \text{ is a supporting hyperplane}} H$

  alternatively, this intersection gives closed convex hull.

Rem
* Supporting hyperplanes give a sense of a dual view of convex sets.
  Indeed, points on boundary share normals iff connected;
  convex sets seem not too far from spheres?
  (See also John lemma + various relations to spheres.)

Further operations on convex sets: [HULL Section A.1.2]

* Minkowski sum: $A+B = \{x+y : x\in A,\ y\in B\}$

convex when A,B convex.

* intersection $\bigcap_{C\in\mathcal{J}} C$ convex when each $C\in\mathcal{J}$ convex
(Pf: consider $\{x,y\}\in \bigcap C$; then $[x,y]\in C\ \forall C\in\mathcal{J}$.)

Convex functions

$f:\mathbb{R}^d\to\mathbb{R}$; $\quad \mathrm{epi}(f)$ ("epigraph") $=\{(x,y):\ y\geq f(x)\}$.

function f convex when $\mathrm{epi}(f)$ convex set.

Equivalently: ~~f(αx+...)~~ $f(\alpha x+(1-\alpha)y)\leq \alpha f(x)+(1-\alpha)f(y)$

Remark. Useful to allow $f=\infty$.
Sometimes also $-\infty$; need to be careful with defn, else $\infty-\infty$. See Rockafeller for more. [ROC Theorem 4.2, Chapter 4]

[Another "technical" thing we're avoiding is closedness of epigraph for now...]

Definition: Sublevel set $S_r(f) := \{x\in\mathbb{R}^d;\ f(x)\leq r\}$. [HULL Section B.1.1]

Rem $S_r(f)$ convex $\forall r$, but this isn't sufficiently for convexity!

Ex. $y\mapsto \frac{1}{2}\|x-y\|^2$ convex

$S := \{y\in\mathbb{R}^d:\ \|x-y\|\leq\|y-w\|\}$ (as we used) is a sublevel set.

$\iota_S(x)=\begin{cases}0 & x\in S\\ \infty & \text{o.w.}\end{cases}$ is convex (very useful!)
$\quad\llcorner$ to add constraints.

$\frac{1}{2}\|x-y\|^2+\iota_S$ was used in proof...

Subdifferentials

[Here I use HULL only for intuition; it avoids ∞.]

Convex functions lie above first-order Taylor: $f(x) \geq f(y) + \langle \nabla f(y), y-x \rangle$.

Useful to consider non differentiable.

Define.

Subdifferential $\partial f(x) := \{ s \in \mathbb{R}^d : \forall y \in \mathbb{R}^d,\ f(y) \geq f(x) + \langle s, y-x \rangle \}$

(Supporting hyperplane!)

Define

$\mathrm{dom}(f)$: $\{x : f(x) < \infty\}$.

By convention, $\partial f(x) = \emptyset$ when $x \notin \mathrm{dom}(f)$.

Prop.

* $\partial f(x)$ is closed convex set. [ROC Theorem 23.2]
* $\partial f(x)$ nonempty if $x \in \mathrm{int}(\mathrm{dom}(f))$ [indeed, if $x \in \underline{\mathrm{ri}}(\mathrm{dom}(f))$.]

[$\mathrm{bd}(\mathrm{dom}(f))$ can go either way; consider $x \mapsto \begin{cases} x\ln(x) - x & x \geq 0 \\ \infty & x < 0 \end{cases}$ at 0.]

[ROC Theorem 23.4]

* f differentiable at x iff $|\partial f(x)| = 1$ (and then $\partial f(x) = \{\nabla f(x)\}$).
* $s_1 \in \partial f(x_1)$, $s_2 \in \partial f(x_2)$ $\Rightarrow$ $\langle s_1 - s_2, x_1 - x_2 \rangle \geq 0$

[This means "nondecreasing slopes along lines".
Proof is easy: by defn $f(x_1) \geq f(x_2) + \langle s_2, x_2 - x_1 \rangle$, $f(x_2) \geq f(x_1) + \langle s_1, x_1 - x_2 \rangle$, then subtract one from other.]

Ex. $f(x) = |x|$.

$$\partial f(x) = \begin{cases} \{-1\} & x < 0, \\ [-1, +1] & x = 0, \\ \{+1\} & x > 0. \end{cases}$$

Theorem (Jensen's inequality).

Suppose $\mathbb{E}(X) \in \operatorname{int}(\operatorname{dom}(f))$. Then

$$f(\mathbb{E}X) \le \mathbb{E}\, f(X).$$

[Generalizes convexity defn. from single points.]

Pf. Pick $s \in \partial f(\mathbb{E}X)$. (Can do this: $\mathbb{E}(X) \in \operatorname{int}(\operatorname{dom}(f))$.)

$$\Rightarrow \mathbb{E}\, f(X) \ge \mathbb{E}\Big(f(\mathbb{E}(X)) + \langle s, X - \mathbb{E}X\rangle \Big) = f(\mathbb{E}X) + \langle s, \mathbb{E}(X - \mathbb{E}X)\rangle = f(\mathbb{E}X). \quad //$$

Remark Jensen is very powerful. We've used it to optimize a few expressions already (need act layout).

~~Another~~ Another example application is AM-GM inequality:

$(r_1, \dots, r_n)$ nonnegative reals, $\vec{\alpha} \in \Delta$:

$$\prod r_i^{\alpha_i} \le \sum_i \alpha_i r_i$$

Pf: $\ln \prod r_i^{\alpha_i} = \sum_i \alpha_i \ln r_i \le \ln\left(\sum \alpha_i r_i\right)$; apply $\exp(\cdot)$ both sides.

($\ln$ is concave ($-\ln$ is convex))

Strict & strong convexity: [HULL Section B.4]

f is strictly convex when $\forall x \ne y$, $\alpha \in (0,1)$,

$$f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y).$$

f is strongly convex with modulus λ when $\forall x, y$, $\alpha \in (0,1)$

$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y) - \frac{c\alpha(1-\alpha)}{2} \|x-y\|^2$$

Ex e^x strict but not strong; $\frac{c}{2}\|x\|^2$ strong.

Props

f strictly convex $\Rightarrow \begin{cases} \nabla^2 f(x) \succ 0 & \text{on interior}(\operatorname{dom}(f)) \text{ (when it exists)} \\ x \ne y, s \in \partial f(x) \Rightarrow & f(y) > f(x) + \langle s, y-x\rangle \end{cases}$

f strongly convex $\Rightarrow \begin{cases} \nabla^2 f(x) \succeq \lambda \cdot I & \text{(when exists)} \\ s \in \partial f(x) & f(y) \ge f(x) + \langle s, y-x\rangle + \frac{\lambda}{2}\|x-y\|^2 \\ f - \frac{\lambda}{2}\|\cdot\| & \text{convex} \end{cases}$

Ex Suppose $f \geq 0$, convex.

Then $R(\cdot) = f + \frac{\lambda}{2}\|\cdot\|^2$ s.c.

Note $R(w) \leq R(0)$ means

$$\frac{\lambda}{2}\|w\|^2 \leq R(w) \leq R(0) = f(0) \Rightarrow \|w\| \leq \sqrt{\frac{2f(0)}{\lambda}}.$$

$\Rightarrow$ Regularization gives bounded sublevel sets!

~~Convexity maps~~

Operations on convex functions

* $g(x) := \sup_{f \in \mathcal{F}} f(x)$ convex when each $f \in \mathcal{F}$ convex.

  (Pf. $\mathrm{epi}(g) = \bigcap_{f \in \mathcal{F}} \mathrm{epi}(f)$.)

  Have roughly $\partial g(x) = \overline{\mathrm{conv}}\left(\bigcup_{\substack{f \in \mathcal{F} \\ f(x) = g(x)}} \partial f(x)\right)$, but need some care when infinite.

* $g = f + h \circ A$ convex; $\partial g = \partial f + A^* \partial h(A\cdot)$

  need care when dom $\neq \mathbb{R}^q$.

  e.g. $A\,\mathrm{dom}\,f \cap \mathrm{cont}\,h \neq \emptyset$

  [BL Theorem 3.3.5]

$\boxed{\text{First-order conditions; primal optimality}}$

Theorem $0 \in \partial f(\bar{x}) \iff f(\bar{x}) = \inf_x f(x)$.

Pf. $0 \in \partial f(\bar{x}) \iff \forall y.\ f(y) \geq f(\bar{x}) + \langle 0, y - x \rangle \iff \forall y.\ f(y) \geq f(\bar{x})$. //

Rem. Convex functions have tremendous regularity:

* local to global: subgradients anywhere give global information
* locally lipschitz where bounded above
* differentiable almost everywhere

Let's go for a very useful corollary of this.

Consider $f + \iota_S$. We have $\bar{x}$ opt $\iff$ $\partial(f+\iota_S)(\bar{x})$.

What is $\partial \iota_S(\bar{x})$?

Suffices to consider $\bar{x} \in S$ (otherwise $(f+\iota_S)(\bar{x}) = \infty$):

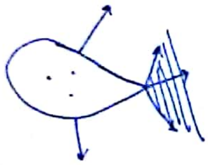
$$\partial \iota_S(\bar{x}) = \{ g \in \mathbb{R}^d : \forall y,\ \iota_S(y) \geq \iota_S(x) + \langle g, y-x \rangle \}$$
$$= \{ g \in \mathbb{R}^d : \forall y \in S,\ 0 \geq \langle g, y-x \rangle \}.$$

can ignore lhs $=\infty$

This gives a useful set "attached" to S:

Defn. Normal cone of S at x is

$$N_S(x) = \{ g \in \mathbb{R}^d : \forall y \in S,\ 0 \geq \langle g, y-x \rangle \}.$$



Theorem. Suppose $S \subseteq \mathbb{R}^d$ closed convex and $\overbrace{\mathrm{dom}(f) \cap \mathrm{int}(S) \neq \emptyset}^{CQ}$.

Then $f(\bar{x}) = \inf_{x \in S} f(x) \iff \partial f(\bar{x}) \cap -N_S(\bar{x}) \neq \emptyset$.

[BL Exercise 3.3.13.(d)]

Pf. $\bar{x}$ optimal

$\iff 0 \in \partial(f + \iota_S)(\bar{x})$

$\iff 0 \in \partial f(\bar{x}) + \partial \iota_S(\bar{x})$ by CQ ("constraint qualification")

$\iff 0 \in \partial f(\bar{x}) + N_S(\bar{x})$

$\iff \partial f(\bar{x}) \cap -N_S(\bar{x}) \neq \emptyset.$ //