Convexity bootcamp part 1: basic objects

Administrative:

- **New room:** Siebel 1214 starting on 9/21.
- Homework 1 is out.
- Project milestone 0 due.

Convexity basics

[Note to future matus: do everything from perspective of duality.]

These lectures will be a little more enyclopedic and dry than the last few.

(Tons of pictures will be drawn in class, omitted in these notes.)

- **References.** These lectures will collect properties related to convexity, but prove only few statements. Excellent references are as follows:
  - I believe the quickest way to get up to speed are the first four chapters of J. Borwein and Lewis (2000), however this deserves a **warning**: the book pushes many key results deep inside the exercises (see for instance exercise 3.3.9.f, which is an essential property). The book is also quite intense and goes for shortest proofs without explaining much.
  - For a much broader, cohesive, and geometric view, I recommend Hiriart-Urruty and Lemaréchal (2001) (note that the same authors also have a two-volume convex opt set of books).
  - The classical reference is Rockafellar (1970), which is what I cite and refer too often since it really tries to pin down many things as tightly as possible, and also it has some generality that was later dropped (e.g., extensive consideration of convex functions taking on $-\infty$).
  - In infinite dimensions, things are a mess, due in part to relative interiors and constraint qualifications breaking down. I have used and can recommend Zălinescu (2002), Rockafellar (1974), J. M. Borwein and Zhu (2005).
- **Why care about convexity?** One may say that machine learning is shifting to non-convex problems. Whether or not this is true, convexity still has the following value.
  - Even in non-convex ML problems, we still generally use convex *losses*.
  - Convexity is one of the most well-developed pieces of geometry we have.
  - Convexity has a very interesting local-to-global property which can perhaps be generalized or relaxed.
- **Convex sets**. A set $C \subseteq \mathbb{R}^d$ is **convex** when it contains the line segments between any pair of points; in symbols,
$$x, x' \in C \qquad \Longrightarrow \qquad \{\alpha x + (1-\alpha)x' : \alpha \in [0,1]\} \subseteq C.$$
Examples.

– Any halfspace $H := \{x \in \mathbb{R}^d : a^\top x \geq b\}$ is convex. This is because

$$x, x' \in H \qquad \implies \qquad \forall \alpha \in [0,1] \centerdot a^\top(\alpha x + (1-\alpha)x') = \alpha a^\top x + (1-\alpha)a^\top x' \geq \alpha b + (1-\alpha)b.$$

– Any intersection of convex sets is convex: for any line segment, its endpoints must be in each constituent of the intersection, and they're all convex so the line segment must be in each constituent as well.
– A **polyhedron** is an intersection of finitely many halfspaces (**polytope** if compact). By the two preceding points, a polyhedron is convex. Note that polyhedra can be written $\{x \in \mathbb{R}^d : Ax \geq b\}$ where A is now a matrix.

Operations on convex sets. Convex sets have associated scalar multiplication and addition rules (latter being *Minkowski sum*): given convex sets C_1, C_2 and scalar α,

$$\alpha C := \{\alpha x : x \in C\}, \qquad C_1 + C_2 := \{x_1 + x_2 : x_1 \in C_1, x_2 \in C_2\}.$$

A **convex combination** of (necessarily finitely many) vectors $(v_1, \ldots, v_k)$ is the vector $\sum_i \alpha_i v_i$ where $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$. Define the **convex hull** $\text{conv}(S)$ of a set of points S as the collection of all convex combinations of elements of S, meaning

$$\text{conv}(S) := \left\{ \sum_{i=1}^k \alpha_i v_i : k \in Z_{++}, (v_i)_{i=1}^k \subseteq S, \alpha_i \geq 0, \sum_i \alpha_i = 1 \right\}.$$

Technical note. Sometimes it will be necessary to talk about *closure*. If you do not have much background in topology/analysis, just treat this as a black box: we generally want sets to be closed, it will give better behavior.

The set of symmetric positive semi-definite matrices is convex. It is a closed set, but the set of symmetric positive definite matrices (no "semi-") is open. Both sets are **convex cones**, meaning they are convex sets that are also **cones** (S is a cone iff $\alpha S \subseteq S$ for all $\alpha > 0$).

- **Convex functions.** Given $f : \mathbb{R}^d \to \mathbb{R}$ define its **epigraph** $\text{epi}(f)$ as the set you get by pouring liquid onto the graph of the function from above:

$$\text{epi}(f) := \left\{ (x, y) : x \in \mathbb{R}^d, y \geq f(x) \right\}.$$

f is **convex** when $\text{epi}(f)$ is a convex set. There are many equivalent definitions of convexity, here's another:

$$f : \mathbb{R}^d \to \mathbb{R} \text{ convex} \qquad \iff \qquad \forall x, x' \in \mathbb{R}^d, \alpha \in [0,1] \centerdot f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha) f(x').$$

(Proof by picture.)

There will be many more ways to phrase convexity; to prove something is convex, it's useful to consider many tools.

Infinite values. It is useful to let convex functions take on infinite values, particularly for optimization: if we want to minimize something but rule out some region of space, we can just make the functions take on the value ∞ on that part of the space. An example is the **indicator function**

$$\iota_S(x) := \begin{cases} 0 & \text{when } x \in S, \\ \infty & \text{when } x \notin S. \end{cases}$$

S is convex iff ι_S is convex. Let $\text{dom}(f)$ denote the points where f is finite, meaning $\text{dom}(f) := \{x \in \mathbb{R}^d : f(x) < \infty\}$; $\text{dom}(f)$ is sometimes called the "effective domain" to disambiguate it from the "domain" $\mathbb{R}^d$. **Technical note:** there starts to be some discrepency in the definitions of convexity when

the convex function is also allowed to take on $-\infty$; the place to look for this is Rockafellar (1970), since many later texts only allow for $-\infty$.

Another note on closure. Convex functions have a notion of closure as well: f is closed when $\text{epi}(f)$ is a closed set. If $f : \mathbb{R}^d \to \mathbb{R}$, then f is closed, but the situation $f : \mathbb{R}^d \to \mathbb{R} \cup \infty$ can lead to non-closed functions (consider ι_S when S is convex but open). As before, if unfamiliar with closure, treat this as a "technical black box", and in general it's easier for us when functions are closed.

Sublevel sets. A useful collection of sets to associate with a function $f : \mathbb{R}^d \to \mathbb{R}$ are its **sublevel sets**

$$S_c(f) := \{x \in \mathbb{R}^d : f(x) \leq c\}.$$

When f is convex, then $S_c(f)$ is convex for every $c \in \mathbb{R}$ *(picture proof: look at epi(f), intersect it with a horizontal hyperplane (intersection of two convex sets, thus convex).)* The converse is not true (consider $g(x) := \sqrt{|x|}$), and describe **quasiconvex** functions.

Operations preserving convexity. If f, g are convex, so is $f + g$; for this it's easiest to check the algebraic definition of convexity. Additionally, for any collection $\mathcal{F}$ of convex functions, $g(x) := \sup_{f \in \mathcal{F}} f(x)$ is convex; a cute proof here is to use the epigraph definition of convexity, and to note

$$\text{epi}(g) := \cap_{f \in \mathcal{F}} \text{epi}(f).$$

In convexity, it is often useful to think of things in many equivalent ways, as some lead to drastically shorter proofs.

- **Convex functions and (sub)differentiation.**

  First order Taylor approximations always lie below convex functions: $f : \mathbb{R}^d \to \mathbb{R}$ is convex iff

  $$\forall x, x' \in \mathbb{R}^d \centerdot f(x) \geq f(x') + \left\langle \nabla f(x'), x - x' \right\rangle,$$

  If the Hessian $\nabla^2 f$ is positive semi-definite over the interior of the domain of f (written $\nabla^2 f \succeq 0$ over $\text{int}(\text{dom}(f))$) then f is also convex over $\text{int}(\text{dom}(f))$.

  However, there are many interesting convex functions which fail to be differentiable. Instead, the notion of a **subgradient** arises by taken the above differentiaion/tangent inequality as a definition: define the **subdifferential** of f at $x \in \text{dom}(f)$ (set of all subgradients), $\partial f(x)$, as

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

  (This definition is for $x \in \text{dom}(f)$; for $x \notin \text{dom}(f)$, set $\partial f(x) = \emptyset$.) Here are some key properties.

  - $\partial f(x)$ is always a closed convex set.
  - $\partial f(x)$ is nonempty whenever $x \in \text{int}(\text{dom}(f))$. (A general characterization is more difficult: consider ι_S for any closed convex S, which has subderivatives along $\text{dom}(\iota_S)$, whereas

    $$s \mapsto s \ln(s) - s$$

    has $\partial f(0) = \emptyset$.) **Note:** do not underestimate this result, it is one of the most important lemmas. In particular, it will be one of the key pieces in establishing duality.
  - f is differentiable at x iff $|\partial f(x)| = 1$ (in which case $\partial f(x) = \{\nabla f(x)\}$).
  - Given matrix $A : \mathbb{R}^{n \times d}$ and convex functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$,

    $$\partial(f + g \circ A)(x) = \partial f(x) + A^\top \partial g(Ax).$$

    If f and g take on the value $+\infty$, this equality becomes more delicate, and needs a **constraint qualification**, e.g., $\text{int}(A\text{dom}(f)) \cap \text{int}(\text{dom}(g)) \neq \emptyset$, as in Exercise 3.3.9.b of J. Borwein and Lewis (2000). Constraint qualifications will be discussed in the next lecture.

– Suppose $s_1 \in \partial f(x_1)$ and $s_2 \in \partial f(x_2)$, where f is convex; applying the definition of subgradient twice,

$$f(x_2) \geq f(x_1) + \langle s_1, x_2 - x_1 \rangle,$$
$$f(x_1) \geq f(x_2) + \langle s_2, x_1 - x_2 \rangle.$$

Summing these, canceling terms, and rearranging,

$$\langle s_2 - s_1, x_2 - x_1 \rangle \geq 0,$$

meaning subgradients increase along fixed lines.

With this we can prove a nice version of **Jensen's inequality**.

Lemma (Jensen's inequality)**.** Let convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and random variable X be given with $\mathbb{E}(X) \in \text{int}(\text{dom}(f))$. Then

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)).$$

Proof. Set $u := \mathbb{E}(X)$. Since $u \in \text{int}(\text{dom}(f))$, then $\partial f(u)$ is nonempty, so pick any $g \in \partial f(u)$. Then

$$\mathbb{E}(f(X)) \geq \mathbb{E}(f(u) + \langle g, X - u \rangle) = f(u) + \big\langle g, \mathbb{E}(X) - u \big\rangle = f(u).$$

To break away from the tedium of these dry notes so far, note that convexity is everywhere and Jensen's inequality is endlessly useful.

Theorem (AM-GM inequality)**.** Let nonnegative reals $(r_1, \ldots, r_k)$ and $(a_1, \ldots, a_k)$ be given with $\sum_i a_i = 1$. Then

$$\prod_i r_i^{a_i} \leq \sum_i a_i r_i.$$

Proof. By Jensen's inequality (since $-\ln$ is convex),

$$\ln \prod_i r_i^{a_i} = \sum_i a_i \ln(r_i) \leq \ln(\sum_i a_i r_i).$$

Lastly, note how useful subgradients are for optimization.

Theorem (first-order optimality conditions)**.** Let convex $f : \mathbb{R}^d \to \mathbb{R}$ be given. Then $\bar{x}$ is a minimizer of f (meaning $f(\bar{x}) = \inf\{f(x) : x \in \mathbb{R}^d\}$ iff $0 \in \partial f(\bar{x})$.

Proof. This holds since

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

- **Strict and strong convexity**. f is **strictly convex** when

$$\forall x \neq x', \alpha \in (0,1) \,.\, f(\alpha x + (1-\alpha)x') < \alpha f(x) + (1-\alpha) f(x')$$

Similarly,

$$\forall x, g \in \partial f(x), x' \neq x \,.\, f(x) > f(x') + \big\langle g, x' - x \big\rangle.$$

Lastly, $f : \mathbb{R}^d \to \mathbb{R}$ if $\nabla^2 f \succ 0$ everywhere. Jensen's inequality can also be adapted to a strict inequality case.

f is **strongly convex** with parameter λ) (or simply **λ-strongly-convex**) when

$$\forall x \neq x', \alpha \in (0,1) \,.\, f(\alpha x + (1-\alpha)x') + \frac{\lambda \alpha (1-\alpha)}{2} \|x - x'\|_2^2 < \alpha f(x) + (1-\alpha) f(x').$$

A less awkward way of writing this is to modify the subgradient inequality:

$$\forall x, \forall g \in \partial f(x), \forall x' \,.\, f(x') \geq f(x') + \big\langle g, x' - x \big\rangle + \frac{\lambda}{2} \|x - x'\|_2^2.$$

Lastly, $f : \mathbb{R}^d$ is λ-strongly-convex when $\nabla^2 f \succeq \lambda I$.

Strong convexity appears constantly in machine learning for roughly the following reason. In machine learning we often optimize a function of the form

$$\mathcal{R}(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_i(w) + \frac{\lambda}{2}\|w\|_2^2,$$

where ℓ_i is a per-example loss function, and the other term is a **regularizer** (this whole expression is called **regularized empirical risk minimization**). If each ℓ_i is convex, then $\mathcal{R}$ is λ-strongly-convex. Moreover, consider any vector $w \in \mathbb{R}^d$ satisfying $\mathcal{R}(w) \le \mathcal{R}(0)$, and suppose $\ell_i \ge 0$ (which is generally true). Then

$$\mathcal{R}(0) \ge \mathcal{R}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(w) + \frac{\lambda}{2}\|w\|_2^2 \ge 0 + \frac{\lambda}{2}\|w\|_2^2.$$

This can be written as follows.

Proposition. If each $\ell_i \ge 0$, then every $w \in S_{\mathcal{R}(0)}(\mathcal{R})$ satisfies

$$\|w\|_2 \le \sqrt{\frac{2\mathcal{R}(0)}{\lambda}}.$$

If each ℓ_i is also continuous, then as a corollary we get minimizers for $\mathcal{R}$ (e.g., this gives that $S_{\mathcal{R}(0)}$ is compact, so $\mathcal{R}(S_{\mathcal{R}(0)})$ is compact, and a minimizer is attained; this is spelled out in (J. Borwein and Lewis 2000, Proposition 1.1.3)).

Note that λ-strongly-convex is often paired with β-smooth, which grants the reverse inequaliy

$$f(x') \le f(x) + \langle g, x' - x\rangle + \frac{\beta}{2}\|x - x'\|_2^2,$$

where $g \in \partial f(x)$. Both this and strong convexity can work with other norms, though things are clearest with $\|\cdot\|_2$.

References

Borwein, Jonathan M., and Qiji J. Zhu. 2005. *Techniques of Variational Analysis*. 1st ed. Springer.

Borwein, Jonathan, and Adrian Lewis. 2000. *Convex Analysis and Nonlinear Optimization*. Springer Publishing Company, Incorporated.

Hiriart-Urruty, Jean-Baptiste, and Claude Lemaréchal. 2001. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated.

Rockafellar, R. Tyrrell. 1970. *Convex Analysis*. Princeton University Press.

———. 1974. *Conjugate Duality and Optimization*. SIAM Publications.

Zălinescu, Constantin. 2002. *Convex Analysis in General Vector Spaces*. World scientific.