Convex Optimization – Lecture 2

Today:

- Convex Analysis
- Center-of-mass Algorithm

Convex Analysis

Convex Sets

Definition: A set $C \subseteq \mathbb{R}^n$ is *convex* if for all $x, y \in C$ and all $0 \leq \lambda \leq 1$,

$$\lambda x + (1 - \lambda) y \in C$$

Operations that preserve convexity:

- Intersection of convex sets is convex
- Scaling, translation, or generally affine transformations ($f(x) = Ax + b$)

Convex combination: The point $\sum_{i=1}^{k} \theta_i x_i$ such that $\theta_i \geq 0$ for all i, and $\sum_{i=1}^{k} \theta_i = 1$ is a *convex combination* of $x_1, \ldots, x_k$.
Claim: If C is convex and $x_i \in C$ for all i, then $\sum_{i=1}^{k} \theta_i x_i \in C$.

Convex hull: The convex hull of a set S, denoted $\mathrm{conv}(S)$, is the intersection of all convex sets containing S.
Equal to the set of all (finite) convex combinations of points in S:

$$\mathrm{conv}(S) = \left\{ \sum_{i=1}^{k} \theta_i x_i \mid \text{finite } k, \sum_i \theta_i = 1, \theta_i \geq 0, x_i \in S \right\}$$

Two special convex sets: hyperplanes and halfspaces

<u>Hyperplane:</u> $\{x \in \mathbb{R}^n \mid \langle a, x \rangle = b\}$, where $a \in \mathbb{R}^n$, $a \neq 0$, $b \in \mathbb{R}$.

<u>Halfspace:</u> $\{x \in \mathbb{R}^n \mid \langle a, x \rangle \leq b\}$

<u>Polyhedra:</u> A *polyhedron* is the intersection of a finite number of halfspaces.
May be unbounded.

A *polytope* is the convex hull of a finite number of points.
Always bounded.

Separating hyperplane theorem

Suppose $C, D \subseteq \mathbb{R}^n$ are convex and disjoint: $C \cap D = \emptyset$. Then there exist $a \neq 0, b$ s.t. $\langle a, x \rangle \leq b$ for all $x \in C$, and $\langle a, x \rangle \geq b$ for all $x \in D$.
Not true if either of the sets is not convex.

Special case: separating a point from a convex set.
Finding a convex set given a point outside it.

Supporting hyperplane theorem

A *supporting hyperplane* to a set $C \in \mathbb{R}^n$ at a boundary point x_0 is the set: $\{x \mid \langle a, x \rangle = b\}$, $\langle a, x_0 \rangle = b$, and for all $x \in C$, $\langle a, x \rangle \leq b$.
Theorem: If C is (open) convex, then there exists a supporting hyperplane at every boundary point of C.
Clarification: the converse statement is also true.

Convex Functions

Definition: Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \mapsto \mathbb{R}$ is *convex* if for all $x, y \in C$ and all $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

We say that f is *concave* if $-f$ is convex.

Extended-value function: Sometimes it is convenient to extend a convex funciton to all of $\mathbb{R}^n$ by defining its value to be $+\infty$ outside the domain. The *extended-value* function $\tilde{f} : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ is defined as:

$$\tilde{f}(x) = \begin{cases} f(x) & x \in C \\ \infty & x \notin C \end{cases}$$

We can then recover the domain of the original function f from the extension $\tilde{f}$ by $C = \{x \mid \tilde{f}(x) < \infty\}$.

First-order condition

The gradient is a *linear functional* that maps each column vector $x \in \mathbb{R}^n$ to the dual vector space $(\mathbb{R}^n)^*$.

<u>First order condition:</u> A differentiable f is convex iff C is convex and:

$$f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \qquad \text{for all } x_0, x \in C$$

This is the first-order Taylor approximation of f at x_0.
The condition states that this linear approximation is a global *underestimator* of the function.

For convex functions, the local information (function value and gradient at a point) gives global information (underestimator).
Characterization of an optimal point:

$$\nabla f(x_0) = 0 \quad \Rightarrow \quad x_0 \in \operatorname*{argmin}_{x \in C} f(x)$$

subgradient

We can define a similar first-order condition for non-differentiable functions.
Definition: $g \in \mathbb{R}^n$ is a *subgradient* of f at x_0 iff:

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle \qquad \text{for all } x \in C$$

Definition: The set of all subgradients is called a *subdifferential*:

$$\partial f(x_0) = \{g \mid g \text{ is a subgradient of } f \text{ at } x_0\}$$

We will use $\nabla f(x_0)$ to denote a single subgradient.

For differentiable functions, the subgradient is unique and given by the gradient.
For non-differentiable functions, the subgradient may not be unique.

Theorem: $f : C \mapsto \mathbb{R}$ is convex $\Leftrightarrow$ there exists a subgradient at each $x \in C$.
This means that convexity is characterized as having subgradients.

Similar to the differentiable case, we have a characterization of an optimum:
Theorem: $0 \in \partial f(x^*) \Leftrightarrow x^* \in \operatorname{argmin}_{x \in C} f(x)$.

Proof:
$\Rightarrow$ From the definition of subgradients we have:

$$f(x) \geq f(x^*) + \langle 0, x - x^* \rangle = f(x^*) \qquad \text{for all } x \in C$$

$\Leftarrow$ Since x^* is a minimizer we have for all $x \in C$:

$$f(x^*) \leq f(x) \quad \Rightarrow \quad f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$$

and therefore, $0 \in \partial f(x^*)$.

This means that every local minimum is also a global minimum of f.

Epigraph

The *epigraph* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$\mathrm{epi}(f) = \left\{(x,t) \in \mathbb{R}^{n+1} \mid f(x) \leq t\right\}$$

Claim: f is a convex *function* $\Leftrightarrow$ $\mathrm{epi}(f)$ is a convex *set*.

A subgradient defines a supporting hyperplane to the epigraph.
May not be unique.

Sublevel sets

The α-sublevel set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$S_\alpha = \{x \mid f(x) \leq \alpha\}$$

Claim: If f is a convex function $\Rightarrow$ S_α is a convex set for all α.
Proof: If $(x, y) \in S_\alpha$, then $f(x) \leq \alpha$ and $f(y) \leq \alpha$, and so $f(\lambda x + (1-\lambda)y) \underset{\text{convexity}}{\leq} \lambda f(x) + (1 - \lambda)f(y) \leq \alpha$ for $0 \leq \lambda \leq 1$, and hence $\lambda x + (1 - \lambda)y \in S_\alpha$.

The converse is not true! But defines quasiconvex functions:
Definition: $f : \mathbb{R}^n \to \mathbb{R}$ is quasiconvex if its sublevel sets S_α are convex for all α.

A subgradient defines a supporting hyperplane to the sublevel set:
<u>Claim:</u> Denote $f(x_0) = \alpha$. If $\nabla f(x_0) \neq 0$ then $S_\alpha \subseteq \{x \mid \langle \nabla f(x_0), x - x_0 \rangle \leq 0\}$.
<u>Proof:</u> $x \in S_\alpha$ means that $f(x) \leq f(x_0)$. In addition, since f is convex, we have that: $f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$. Now,

$$\begin{aligned} \langle \nabla f(x_0), x \rangle &\leq \langle \nabla f(x_0), x_0 \rangle + (f(x) - f(x_0)) \\ &\leq \langle \nabla f(x_0), x_0 \rangle \end{aligned}$$

This is very important for optimization. If we are at x_0 and want to improve (reduce) f, then the subgradient excludes half of the space.

Center-of-mass Algorithm

Extends Bisection to higher dimensions.
Special type of *cutting-plane* methods (query point may vary).

Problem:

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

f, C convex.

Assumptions:

- Function is bounded: $|f| \leq B$.
- Access to first- and zero-order oracles.

Center-of-mass Algorithm [Levin, Newman 1965]

Let $G^{(1)} = C$.

For $t = 1, \ldots, T$ do

$i)$ Compute center of mass: $c_t = \frac{\int_{x \in G^{(t)}} x dx}{\int_{x \in G^{(t)}} dx}$

$ii)$ Compute subgradient at c_t, obtain $g_t \in \partial f(c_t)$, and let:

$$G^{(t+1)} = G^{(t)} \cap \{x \in \mathbb{R}^n \mid \langle g_t, x - c_t \rangle \leq 0\}$$

Output: $\tilde{x} \in \operatorname{argmin}_{1 \leq t \leq T} f(c_t)$.

We next analyze the convergence of this algorithm.

Grunbaum's Theorem (1960)
Let $G \subseteq \mathbb{R}^n$ be a bounded convex set, with center of mass c, then for any hyperplane passing through c (i.e., $\{g \mid \langle g, x - c\rangle = 0\}$), we have:

$$\mathrm{Vol}\left(G \cap \{x \in \mathbb{R}^n \mid \langle g, x - c\rangle < 0\}\right) \leq \left(1 - \frac{1}{e}\right) \mathrm{Vol}(G)$$

Therefore, after t iterations of the algorithm:

$$\mathrm{Vol}(G^{(t)}) \leq \left(1 - \frac{1}{e}\right)^t \mathrm{Vol}(C)$$

Claim:

$$f(\tilde{x}) - f^* \leq 2B\left(1 - \frac{1}{e}\right)^{t/n}$$

This implies that in order to get ϵ-suboptimality it is enough to run for $T = 2.2n\log(2B/\epsilon)$ iterations (queries to the oracles).
Solve: $2B\left(1 - \frac{1}{e}\right)^{t/n} \leq \epsilon$ for t.
This means *linear convergence.*

Proof:
Let x^* be a minimizer of f (for simplicity, we assume it's unique).
Due to the update rule, we have that:

$$\langle g_t, x - c_t \rangle > 0 \quad \text{for all } x \in (G^{(t)} \setminus G^{(t+1)})$$

Now, since $g_t \in \partial f(c_t)$ then: $f(x) \geq f(c_t) + \langle g_t, x - c_t \rangle$. So $f(x) > f(c_t)$ for all $x \in (G^{(t)} \setminus G^{(t+1)})$. Therefore, we never exclude the optimal point, so $x^* \in G^{(t)}$ for all t.
Next, for $0 \leq \epsilon \leq 1$ define the set $C_\epsilon = \{(1-\epsilon)x^* + \epsilon x \mid x \in C\}$ (shrinking C around x^*).
Note that $\mathrm{Vol}(C_\epsilon) = \epsilon^n \mathrm{Vol}(C)$.
Combining this with $\mathrm{Vol}(G^{(t)}) \leq \left(1 - \frac{1}{e}\right)^t \mathrm{Vol}(C)$ we can defer that:

$$\epsilon > \left(1 - \frac{1}{e}\right)^{t/n} \quad \Rightarrow \quad \mathrm{Vol}(G^{(t+1)}) < \mathrm{Vol}(C_\epsilon)$$

This implies that for $\epsilon > \left(1 - \frac{1}{e}\right)^{t/n}$ there must be a time $1 \leq r \leq t$, and $x_\epsilon \in C_\epsilon$ such that $x_\epsilon \in G^{(r)}$ and $x_\epsilon \notin G^{(r+1)}$.
From the argument above: $x \in (G^{(t)} \setminus G^{(t+1)}) \Rightarrow f(c_t) < f(x_\epsilon)$.

On the other hand, from convexity we have:

$$\begin{aligned} f(\underbrace{(1-\epsilon)x^* + \epsilon x}_{x_\epsilon}) &\leq (1-\epsilon)f(x^*) + \epsilon f(x) \\ &= f(x^*) - \epsilon f(x^*) + \epsilon f(x) \\ &\leq f(x^*) + 2\epsilon B \end{aligned}$$

So together:

$$f(c_t) < f(x^*) + 2B\epsilon$$

Substituting for ϵ completes the proof.

<u>Comment:</u> Finding the center-of-mass is hard in general (even for polyhedra), so this is fast but each iteration is expensive.