10-725: Convex Optimization **Spring 2023**

Lecture 1: January 17

Lecturer: Siva Balakrishnan

1.1 Administrative Highlights

Most information can be found on the Syllabus, including information about grading, HWs, quizzes and tests.

The first half of the course (until Spring break) will be taught in person, and the second half will be taught via Zoom. We will be using some combination of Piazza, Canvas and Gradescope.

For almost all questions your first point of contact should be the Education Associate: Daniel Bird (`dpbird@andrew.cmu.edu`).

HW 1 will be released in one week and will be due two weeks after release.

The course will be fairly fast paced – we will assume some familiarity with real analysis, calculus and linear algebra. If you fall short on any of these things, it?s certainly possible to catch up; but don't hesitate to talk to us.

1.2 What the course is about

The course is broadly about optimization. Despite the fact that it's a course in the ML department – this is not solely a course about optimization for deep learning. At least the first half of the course will primarily focus on *convex* optimization. These ideas are extremely broadly useful.

There are lots of different motivations for the topics we will learn about in this course. Optimization problems are everywhere in ML, Statistics, and tons of other disciplines.

1. A deep understanding of optimization will aid in designing algorithms to solve different types of optimization problems, and in understanding their relative merits. This will be our primary focus in this course.
2. Just formulating an optimization problem often gives a much deeper understanding of the problem at hand – for instance, the statistical analysis of most estimators crucially builds on insights (and characterizations) obtained by formulating the estimator as a solution to an optimization problem.

3. Finally, knowing the tricks of the optimization trade often aids in creating new optimization problems (ones with better algorithmic properties – i.e. are easier to solve, or better statistical properties).

Todays lecture will focus on introducing optimization problems, and convex optimization problems (which will be the focus of the first half of the course). Then we'll turn our attention to defining and understanding convex sets. This is all from Chapters 1 and 2 of the Boyd-Vandenberghe (henceforth BV) book.

1.3 (Mathematical) optimization problems

An optimization problem of the form,

$$\begin{aligned} &\min \ f_0(x) \\ &\text{subject to } f_i(x) \leq b_i, \quad i \in \{1, \ldots, m\}. \end{aligned}$$

Just some terminology:

1. **Optimization variables:** $x \in \mathbb{R}^d$.
2. **Objective function:** $f_0 : \mathbb{R}^d \mapsto \mathbb{R}$.
3. **Constraint functions:** $f_i : \mathbb{R}^d \mapsto \mathbb{R}$.
4. **Feasible solution:** x satisfies all the constraints.
5. **Optimal solution:** x^*, has smallest value of f_0 amongst all vectors which satisfy constraints.
6. **Optimal value:** $p^* = \inf\{f_0(x) : f_i(x) \leq 0, i \in \{1, \ldots, m\}\}$.
   - p^* may not be attained, i.e. there may not be an x^* for which $f_0(x^*) = p^*$.
   - $p^* = \infty$ if problem is infeasible.
   - $p^* = -\infty$ if problem is unbounded from below.

1.3.1 Examples

It is worth keeping in mind some examples of optimization problems, just so we have some concrete places to map the terminology we will learn. Here are some of my favorite optimization problems:

1. Maximum likelihood
2. Least squares
3. Empirical risk minimization
4. Optimal Transport

1.3.2 Standard form

It is not significantly different, but some authors (particularly BV), refer to programs in standard form as also additionally allowing equality constraints, i.e.

$$\begin{aligned} \min \ & f_0(x) \\ \text{subject to } & f_i(x) \leq b_i, \quad i \in \{1, \ldots, m\} \\ & h_i(x) = 0, \quad i \in \{1, \ldots, p\}. \end{aligned}$$

1.3.3 Implicit versus explicit constraints

The above optimization problems have some explicit (inequality and equality) constraints. It is worth noting that in general they also have *implicit constraints*, i.e. that,

$$x \in \mathcal{D} = \operatorname{dom}(f_0) \cap \bigcap_{i=1}^{m} \operatorname{dom}(f_i) \cap \bigcap_{i=1}^{p} \operatorname{dom}(h_i).$$

That is to say, these functions may not be defined everywhere, in which case our optimization problem is implicitly only over vectors where all the criterion and constraint functions are defined.

If we wanted to be more explicit we might write the standard form optimization problem as:

$$\begin{aligned} \min_{x \in \mathcal{D}} \ & f_0(x) \\ \text{subject to } & f_i(x) \leq b_i, \quad i \in \{1, \ldots, m\} \\ & h_i(x) = 0, \quad i \in \{1, \ldots, p\}. \end{aligned}$$

1.3.4 Convex Optimization Problems – Standard Form

A problem of the form,

$$\begin{aligned} \min_{x \in \mathcal{D}} \ & f_0(x) \\ \text{subject to } & f_i(x) \leq b_i, \quad i \in \{1, \ldots, m\} \\ & h_i(x) = 0, \quad i \in \{1, \ldots, p\}, \end{aligned}$$

where

1. $\mathcal{D}$ is a convex set.
2. $f_0, f_1, \ldots, f_m$ are convex functions.
3. $h_i(x) = a_i^T x + b_i$, are affine functions.

To make sense of this definition we'll need to understand what convex sets are, and what convex functions are. This will be what we will spend most of this and the next lecture on.

For now it is worth noting (and re-visiting once the definitions are in place), that the explicit constraints define a convex set, and their intersection with the domain $\mathcal{D}$ is also a convex set. If we denote this convex set $\mathcal{C}$ then our convex optimization problem can be equivalently, succinctly described as:

$$\min_{x \in \mathcal{C}} \; f_0(x),$$

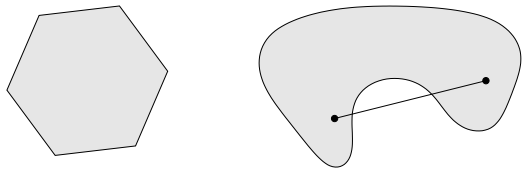
i.e. a *convex optimization problem* is simply the problem of minimizing a *convex function* over a *convex set*.

1.3.5 The Key Feature of Convex Optimization Problems

The most important structural feature of convex optimization problems is that *every local minima is a global minima*. This in turn makes local search algorithms effective for convex optimization.

We'll need to define some things in order to make sense of this claim. First, lets briefly define convex sets and functions:

Definition 1.1 (Convex Set) *A set C is convex, if for every $x_1, x_2 \in C$ and $0 \leq \theta \leq 1$ we have that, $\theta x_1 + (1 - \theta) x_2 \in C$.*

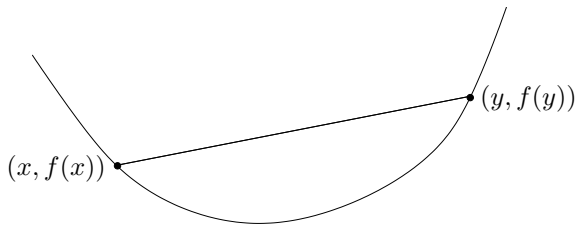


Definition 1.2 (Convex Function) *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex function if,*

1. *$dom(f)$ is a convex set,*

2. *for every* $x, y \in dom(f)$*, and* $0 \leq \theta \leq 1$ *we have that,*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y).$$



Next, we'll need to understand what local optima are:

Definition 1.3 (Local & Global Optima) *A point* x *is a local optima, if* x *is feasible, and minimizes* f_0 *in a local neighborhood, i.e. for some* $\rho > 0$*,*

$$f_0(x) \leq f_0(y),$$

for all y *which are feasible, and* $\|x - y\|_2 \leq \rho$*. A point* x^* *is a global optima, if* x^* *is feasible and*

$$f_0(x) \leq f_0(y),$$

for all y *which are feasible.*

Theorem 1.4 *For a convex optimization problem any local optima is a global optima.*

Proof: Let x be a local optima. Suppose for contradiction of global optimality, that there is some x^* which is feasible, and has the property that,

$$f_0(x^*) < f_0(x).$$

Now, lets examine a new point,

$$x_0 = \left(1 - \frac{\rho}{\|x - x^*\|_2}\right) x + \frac{\rho}{\|x - x^*\|_2} x^*.$$

Notice that,

1. x_0 is feasible, since it is a convex combination of two feasible points x and x^*, and the set of feasible points is a convex set.

2. It is within a ρ-neighborhood of the local optima x, i.e.

$$\|x - x_0\|_2 = \frac{\rho}{\|x - x^*\|_2}\|x - x^*\|_2 = \rho.$$

3. Finally, observe that the objective value at x_0 by using the convexity of f_0 can be upper bounded as,

$$\begin{aligned} f_0(x_0) &\leq \left(1 - \frac{\rho}{\|x - x^*\|_2}\right) f_0(x) + \frac{\rho}{\|x - x^*\|_2} f_0(x^*) \\ &= f_0(x) + \frac{\rho}{\|x - x^*\|_2}(f_0(x^*) - f_0(x)) < f_0(x), \end{aligned}$$

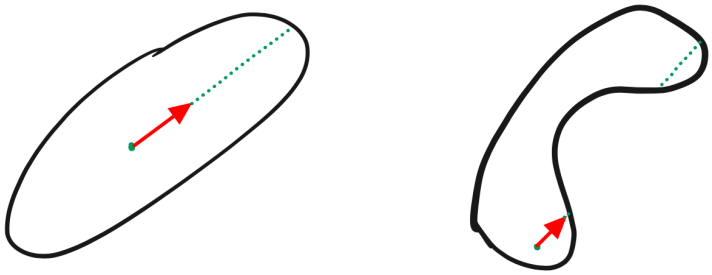
since $f_0(x^*) < f_0(x)$. However, since x_0 is in the ρ-neighborhood of x, this final claim contradicts the local optimality of x.

As a consequence we see that there cannot be any feasible x^* which satisfies $f_0(x^*) < f_0(x)$.

■

1.4 Convex Sets

We have already defined convex sets so let us briefly reflect on why they are so important in optimization. Here is picture you should have in your head, suppose we are optimizing some function over a set C and the function is simple (linear) and takes smaller values in the direction of the arrow. In case the domain is convex, we can follow the "good direction" and when we hit a "wall" declare that we're done. If it's not convex, we have a problem – there could be some "juicy" points (with much better objective value) somewhere "across the wall", and there is no easy way to optimize.



Our next goal will be to describe some examples.

Convex Hull: For a given collection of points $x_1, \ldots, x_k \in \mathbb{R}^k$, a convex combination of the points is a linear combination,

$$\theta_1 x_1 + \ldots + \theta_k x_k,$$

with $\theta_i \geq 0$, and $\sum_{i=1}^k \theta_i = 1$. For a set C, the *convex hull* $\text{conv}(C)$ is the set of all convex combinations of elements of C. This is always a convex set (and is the smallest convex set that contains C).

Many more examples (in each case, would be a good exercise to figure out how you would verify convexity):

1. Trivial ones: empty set, point, line
2. **Norm ball:** $\{x : \|x\| \leq r\}$, for any given norm $\|\cdot\|$ and radius $r \geq 0$.
3. **Hyperplane:** $\{x : a^T x = b\}$ for a given a, b.
4. **Halfspace:** $\{x : a^T x \leq b\}$ for a given a, b. Note that halfspaces are fundamental convex sets. We will think about them in more detail when discussing the separating and supporting hyperplane theorems. They are also at the heart of convex duality.
5. **Affine space:** $\{x : Ax = b\}$, for given A, b.

Here is a slightly more interesting example.

Theorem 1.5 *The set of optimal solutions X_{opt} to a convex optimization problem is a convex set.*

Proof: Suppose we consider, $x_1, x_2 \in X_{\text{opt}}$. Since they are both optimal we must have that $f_0(x_1) = f_0(x_2)$. Now, consider $x_0 = \theta x_1 + (1-\theta)x_2$, where $0 \leq \theta \leq 1$. x_0 is feasible, since the set of feasible solutions is convex. Further, by convexity of the objective we see that,

$$f_0(x_0) \leq \theta f_0(x_1) + (1-\theta) f_0(x_2) \leq f_0(x_1),$$

and so $x_0 \in X_{\text{opt}}$ also. ∎

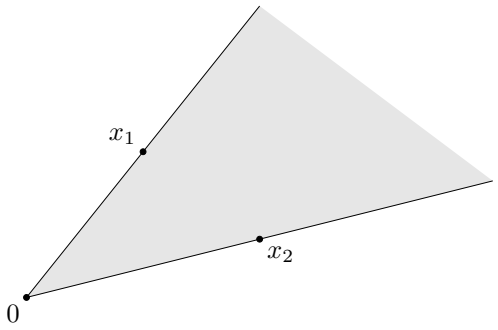
Some more examples (again, useful to make sure you know how to verify the convexity of these sets):

1. **Polyhedra:** The set $\{x : Ax \leq b\}$ for given A, b (or equivalently, sets of the form $\{x : Ax \leq b, Cx = d\}$).
2. **Simplices:** $\{x_1, \ldots, x_k\}$ are affinely independent if there is no $\lambda_1, \ldots, \lambda_k$, with $\sum_{i=1}^k \lambda_i = 0$ such that $\sum_{i=1}^k \lambda_i x_i = 0$ except all zeros. For a collection of affinely independent points $x_1, \ldots, x_k$, the corresponding simplex is simply the convex hull $\text{conv}\{x_1, \ldots, x_k\}$.

   A prominent example is the probability simplex, which is the convex hull of the d basis vectors $e_1, \ldots, e_d$.

1.4.1 Convex Cones

A set C is a cone if for every $x \in C$, $\theta x \in C$ for any $\theta \geq 0$, i.e. for any point in C the ray joining that point to the origin must also be in C.



Cones are not convex in general, so we will refer to *convex cones* as cones which are additionally convex. It is easy to see that convex cones additionally satisfy the property that if $x_1, x_2 \in C$ then for any $\theta_1, \theta_2 \geq 0$, $\theta_1 x_1 + \theta_2 x_2 \in C$. These are called *conic combinations*, i.e. for $x_1, \ldots, x_k$, a conic combination is any point of the form $\theta_1 x_1 + \ldots + \theta_k x_k$ with $\theta_i \geq 0$ is called a conic combination. The conic hull of a set C collects all conic combinations of points in C, and is the smallest *convex* cone containing C.

There are several important cones:

1. **Norm Cone:** $\{(x, t) : \|x\| \leq t\}$. For the ℓ_2 norm this cone is called the second-order cone (sometimes called the ice-cream cone).

2. **PSD Cone:** Denoted $\mathbb{S}^d_+ = \{X \in \mathbb{S}^d : X \succeq 0\}$, i.e. X is a symmetric matrix, with all positive eigenvalues.

For any cone C, the *polar* cone C° is defined as the collection of vectors which make an atleast 90-degree angle with all vectors in C, i.e.

$$C^\circ = \{x : x^T y \leq 0, \text{for all } y \in C\}.$$

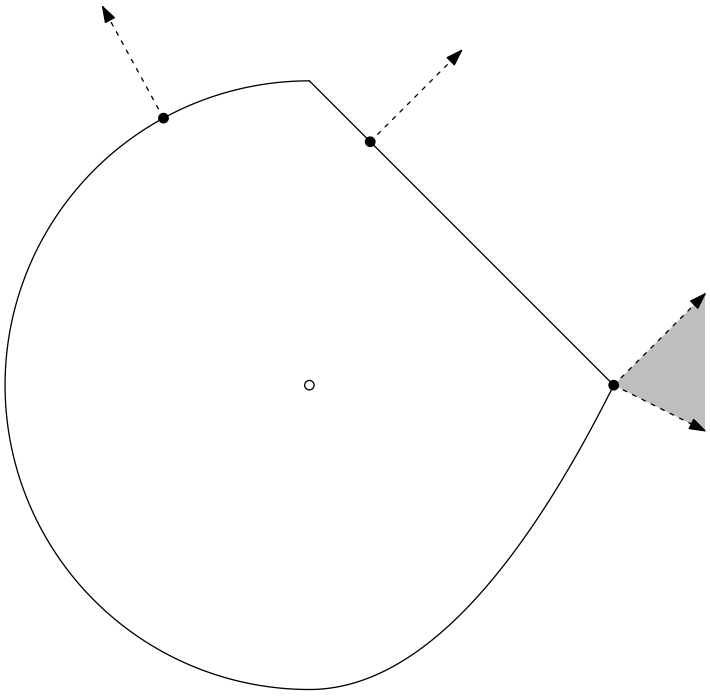
1.4.1.1 The Tangent Cone and Normal Cone

There is a fundamental reason why cones will be important to us. We will use them to characterize optimality. Two cones are important in this context: the normal cone and its polar cone (which has its own name, the tangent cone).

1. **Normal Cone:** Given a set C, and a point $x \in C$ the normal cone of C at x is defined as:

$$N_C(x) = \{g : g^T(y - x) \leq 0, \text{for all } y \in C\}.$$

It is important to make sense of the following figure (for clarity in the figure, the normal cone $N_C(x)$ has been translated to x).



There are three different types of points for which we should understand what the normal cone looks like: (1) Interior points (the normal cone is empty), (2) Boundary points where the boundary is smooth (the normal cone is a single ray) (3) Boundary points where the boundary is not smooth (the normal cone is "fat").

Even if C is not convex this cone is a convex cone (think about how you might show this).

2. **Tangent Cone:** For *convex sets* the polar of the normal cone is the tangent cone, i.e. $T_C(x) = N_C(x)^\circ$. In this case, the tangent cone is a convex cone.

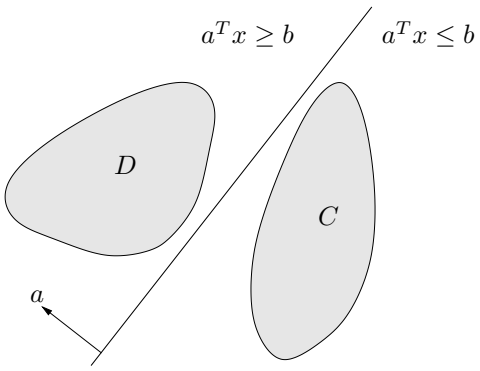
More generally (i.e. for non-convex sets) the tangent cone is defined to be the set of feasible (limiting) directions, i.e. roughly directions along which you can move and stay in the set C. This is possibly the more intuitive way of thinking about the tangent cone at a point (it is simply the set of feasible directions we can move and stay in the set). For general sets C, the tangent cone need not be convex.

We will re-visit optimality conditions at some point, but for now we'll just summarize the punchline: in a convex optimization problem, a point x will be optimal if the negative gradient belongs to $N_C(x)$, i.e. roughly if the direction we'd like to move makes atleast a 90-degree angle with every direction that we *can* move in.

1.5 The Separating and Supporting Hyperplane Theorems

Theorem 1.6 (Separating Hyperplane) *If C and D are non-empty convex sets which are disjoint, i.e. $C \cap D = \emptyset$, then there exists a separating hyperplane, i.e. a, b such that,*

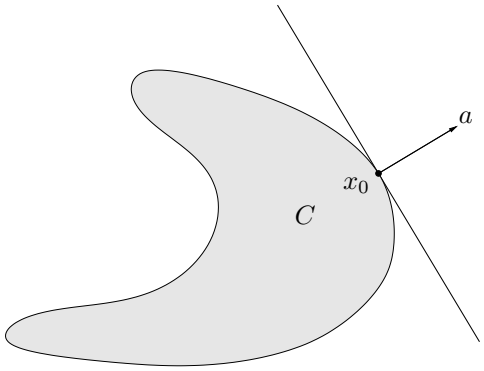
$$a^T x \leq b, \text{ for all } x \in C,$$
$$a^T x \geq b, \text{ for all } x \in D.$$



Theorem 1.7 (Supporting Hyperplane) *If C is a non-empty convex set, and $x_0 \in$ boundary(C), then there is a vector a such that,*

$$a^T(x - x_0) \leq 0, \text{ for all } x \in C.$$

The latter has an interesting converse, if the set C is closed (check what this means if you're not familiar with it), and has a non-empty interior, and has a supporting hyperplane at every point then C must be convex.



The proofs of these theorems (at least in the case where the sets are closed and bounded) is straightforward (and explicit) – see BV, Section 2.5 if you are curious.

1.6 Operations which Preserve Convexity

There are some important operations which preserve convexity of sets:

1. **Intersection:** The intersections of convex sets is a convex set.

2. **Scaling and Translation:** If C is convex, then

$$aC + b := \{ax + b : x \in C\},$$

   is convex for any a, b.

3. **Affine Images and Pre-Images:** Let us define $f(x) = Ax + b$ to be an affine function. If C is a convex set, then,

$$f(C) = \{f(x) : x \in C\}$$

   is also a convex set. Also,

$$f^{-1}(C) = \{x : f(x) \in C\},$$

   is a convex set.

There are a couple more that are more involved but useful to know (we may not have time to cover this in lecture, in which case we will re-visit it when we need it).

1. **Perspective:** The perspective function $P : \mathbb{R}^d \times \mathbb{R}_{++} \mapsto \mathbb{R}$ (where $\mathbb{R}_{++}$ is the strictly positive reals), is defined as:

   $$P(x, t) = x/t.$$

   If $C \subseteq \text{dom}(P)$ is a convex set, then its image $P(C)$ is a convex set, and similarly if D is convex then $P^{-1}(D)$ is convex.

2. **Linear-Fractional:** The linear fractional function for a given A, b, c, d is given by:

   $$f(x) = \frac{Ax + b}{c^T x + d}.$$

   If $C \subseteq \text{dom}(f)$ is a convex set, then its image $f(C)$ is a convex set, and similarly if D is convex then $f^{-1}(D)$ is convex.

Conditional Probability Set: This is an example of using the linear-fractional image to characterize convexity. Let U, V be random variables over $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$. Let $C \subseteq \mathbb{R}^{nm}$ be a set of joint distributions for U, V, i.e., each $p \in C$ defines joint probabilities

$$p_{ij} = \mathbb{P}(U = i, V = j)$$

Let $D \subseteq \mathbb{R}^{nm}$ contain corresponding *conditional distributions*, i.e., each $q \in D$ defines

$$q_{ij} = \mathbb{P}(U = i | V = j)$$

Assume C is convex. Let's prove that D is convex. Write

$$D = \left\{ q \in \mathbb{R}^{nm} : q_{ij} = \frac{p_{ij}}{\sum_{k=1}^{n} p_{kj}}, \text{ for some } p \in C \right\} = f(C)$$

where f is a linear-fractional function, hence D is convex.