Lecture 2: Convex Optimization

Lecturer: Roi Livni *Scribe: Idan Amir*

In this lecture we will provide the standard setup for convex optimization, and define the basic terminologies as well as provide the preliminary results and properties we will need to discuss the standard optimization algorithms in the convex setting.

2.1 Basic terminology, notations and definitions

We first define some basic definitions that will serve us later.

Definition 2.1. *An open ball $B(x,\epsilon)$ with radius $\epsilon > 0$ centered around x is defined as:*

$$B(x,\epsilon) = \{y : \|x - y\| < \epsilon\}$$

Definition 2.2. *A set X is* open *if for every $x \in X$, there exists $\epsilon > 0$ such that $B(x,\epsilon) \subseteq X$.*

Definition 2.3. *A set $X \subseteq \mathbb{R}^n$ is closed if its complement is open (i.e., $\mathbb{R}^n/X$ is open) equivalently:*

$$\forall \epsilon > 0, B(x,\epsilon) \cap X \neq \emptyset \Longrightarrow x \in X$$

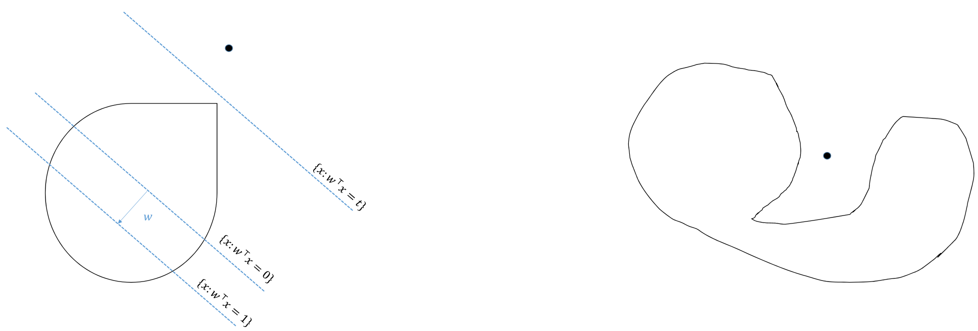
Definition 2.4. *The interior of X is the set of all points x which hold:*

$$intX = \{x : \exists \epsilon > 0 | B(x,\epsilon) \subseteq X\}$$

Definition 2.5. *The boundary of X is defined as $\partial X = X \setminus intX$.*

2.1.1 Properties of convex functions and convex sets

Theorem 2.6 (Separation Theorem)**.** *Let $X \subset \mathbb{R}^n$ be a closed convex set, and $x_0 \in \mathbb{R}^n \setminus X$. Then there exists $w \in \mathbb{R}^n$ and $t \in \mathbb{R}$ such that $\forall x \in X$, $w^T x \geq t$ and $w^T x_0 < t$.*



(a) A point separated from a convex set (b) A non convex set and a point that cannot be separated

Theorem 2.7 (Supporting Hyperplane Theorem)**.** *Let $X \subseteq \mathbb{R}^n$ be a convex set, and $x_0 \in \partial X$. Then there exists $w \neq 0, w \in \mathbb{R}^n$ such that $\forall x \in X, w^T x \geq w^T x_0$.*

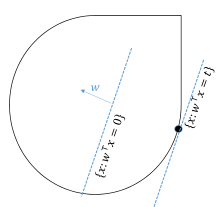


Figure 2.2: A supporting hyperplane (tangent) at a point on the boundary

Definition 2.8 (Subgradient)**.** *Let $f : X \to \mathbb{R}$ be a function over the domain $X \subseteq \mathbb{R}^n$, then $g \in \mathbb{R}^n$ is a subgradient of f at the point $x \in X$ if for every $y \in X$:*

$$f(x) - f(y) \leq g^T (x - y)$$

The subgradient set of f at the point x is denoted by $\partial f(x)$.

Theorem 2.9 (Existence of subgradients)**.** *Let $X \subseteq \mathbb{R}^n$ be a convex set and denote $f : X \to \mathbb{R}$, then:*

1. *If $\partial f(x) \neq \emptyset$, $\forall x \in X$ then f is convex.*
2. *If f is convex then $\partial f(x) \neq \emptyset$, $\forall x \in intX$.*
3. *If f is convex and differentiable at x then $\nabla f(x) \in \partial f(x)$.*

As a first step we will define the epigraph of a function.

Definition 2.10. $epi(f) = \{(x,t) \subseteq X \times \mathbb{R} : f(x) \leq t\}$

proof of theorem 2.9.

1. We proceed by proving the first claim by assuming that $\partial f(x)$ is a non-empty set for all $x \in X$ we will show that for any $x, y \in X$:

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

Denote $g \in \partial f\left((1-\lambda)x + \lambda y\right)$. By definition of the subgradient we get the following two inequalities:

(a) $f\left((1-\lambda)x + \lambda y\right) - f(x) \leq \lambda g^T(y-x)$

(b) $f\left((1-\lambda)x + \lambda y\right) - f(y) \leq (1-\lambda) g^T(x-y)$

Multiplying the first inequality by $(1-\lambda)$, the second by λ and summing between the two, we get:

$$f\left((1-\lambda)x + \lambda y\right) - (1-\lambda)f(x) - \lambda f(y) \leq 0$$

this concludes the proof of our first claim.

2. For the second claim we will assume a convex function f and show that $\partial f(x) \neq \emptyset$, $\forall x \in intX$. Given $x \in X$ it is trivial that $(x, f(x)) \in \partial epi(f)$. We apply the supporting hyperplane theorem to the boundary point $(x, f(x))$ and the convex set $epi(f)$ (see Exercise 2.1). Therefore, there exist $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that for all y and $r \geq \max\{f(y), 0\}$, (since $(y,r) \in epi(f)$):

$$a^T x + bf(x) \leq a^T y + b \cdot r$$

This implies that $b \geq 0$ by taking $r \to \infty$. In addition, assuming that $b = 0$ and setting $y = x - \epsilon a$ results in the contradiction $a^T x \leq a^T x - \epsilon \|a\|^2$. Note that there exist such $\epsilon > 0$ that guarantees

$x + \epsilon a \in X$ since $x \in intX$. Dividing then by $b \geq 0$, taking $r = 1$ and rearranging terms:

$$f(y) - f(x) \geq -\frac{1}{b}a^T(y - x)$$

therefore $-\frac{1}{b}a \in \partial f(x)$.

3. For our third and final claim we will assume that f is convex and differentiable at x and show that $\nabla f(x) \in \partial f(x)$. By convexity,

$$\begin{aligned} f(y) &\geq \frac{f\left((1-\lambda)x + \lambda y\right) - (1-\lambda)f(x)}{\lambda} \\ &= f(x) + \frac{f\left(x + \lambda(y - x)\right) - f(x)}{\lambda} \end{aligned}$$

Taking the limit $\lambda \to 0$,

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

hence $\nabla f(x) \in \partial f(x)$.

$\square$

Exercise 2.1. *Show that f is convex $\iff$ $epi(f)$ is a convex set.*

Exercise 2.2. *Show that for any convex function f there exists a subset $W_f \subseteq \mathbb{R}^n \times \mathbb{R}$ of pair such that:*

$$f(x) = \sup_{(w,c) \in W_f} \{w \cdot x + c\}.$$

2.1.2 First-order optimality conditions

Theorem 2.11. *Let f be a convex function over the domain X, then the following statements are equivalent:*

1. *x is a local minimum.*
2. *x is a global minimum.*
3. $0 \in \partial f(x)$

proof of theorem 2.11. First we show that (3) $\to$ (2): Assume that $0 \in \partial f(x)$, therefore by definition of the subgradient $\forall y \in X$,

$$f(x) - f(y) \leq 0^T(x-y) = 0$$

$$\Downarrow$$

$$f(x) \leq f(y)$$

this implies that x is the global minimum. We proceed by showing that (2) $\to$ (3): Assume that x is a global minimum, then $\forall y \in X$,

$$f(x) - f(y) \leq 0 = 0^T(x-y) \Longrightarrow 0 \in \partial f(x)$$

Lastly we prove that (1) $\to$ (2): Assume that x is a local minimum, then for any chosen $y \in X$ there exists a small enough $t > 0$ such that,

$$\begin{aligned} f(x) &\leq f\left(x + t\left(y-x\right)\right) \\ &= f\left(\left(1-t\right)x + ty\right) \end{aligned}$$

by convexity of f we obtain,

$$f(x) \leq (1-t)f(x) + tf(y)$$

$$\Downarrow$$

$$f(x) \leq f(y)$$

which concludes our proof. $\square$

Theorem 2.12 (First order optimality condition)**.** *Let K be a convex set and f a differentiable convex function over the domain K. Then $x^\star \in K$ holds $x^\star = \arg\min_{x\in K} f(x)$ if and only if,*

$$\forall y \in K \quad \nabla f(x^\star)^T(x^\star - y) \leq 0$$

proof of theorem 2.12. Assume $x^\star = \arg\min_{x\in K} f(x)$, then for any chosen $y \in K$ and $0 < t < 1$,

$$f(x^\star) \leq f\left(x^\star + t\left(y - x^\star\right)\right)$$

Dividing by t and rearranging the terms we obtain,

$$\frac{f(x^\star) - f\left(x^\star + t\left(y - x^\star\right)\right)}{t} \leq 0$$

And conclude the first part of our proof by taking the limit $t \to 0$,

$$\nabla f(x^\star)^T (x^\star - y) \leq 0$$

Note that $x^\star + t\left(y - x^\star\right) \in K$ since K is convex. For the other direction suppose $\nabla f(x^\star)^T (x^\star - y) \leq 0$ for all $y \in K$. Using the properties of differentiable convex functions combined with our assumption we obtain,

$$\begin{gathered} f(x^\star) - f(y) \leq \nabla f(x^\star)^T (x^\star - y) \leq 0 \\ \Downarrow \\ f(x^\star) \leq f(y) \end{gathered}$$

$\square$

2.2 General Setup

We next set out to define the standard setting of *Convex Optimization.* Consider the following constrained (convex) optimization problem,

$$\begin{aligned} \text{minimize} \quad & f(w) \\ \text{s.t.} \quad & w \in \mathcal{W} \end{aligned}$$

where we assume, as discussed, that f is convex and $\mathcal{W}$ is convex. We will also often make the following two assumptions:

1. f is bounded over $\mathcal{W}$. Namely, for a constant $C > 0$ $f : \mathcal{W} \to [-C, C]$.

2. $\mathcal{W}$ has bounded diameter (we will also sometimes assume that it contains 0 in its interior). Namely, For a known r and R, the set $\mathcal{W}$ satisfies:

$$B(0, r) \subseteq \mathcal{W} \subseteq B(0, R)$$

These assumptions alone are not sufficient for efficient optimization. Moreover, any optimization problem can be formalized as a convex problem (in fact a Linear Program). Also, note that we have not defined how the function f is represented, or what are the types of computations that we assume we are allowed to perform on f (we also did not explain what we may or may not assume about $\mathcal{W}$ but we will talk about this later). So it is not clear in what sense we can claim that we can solve convex problems.

To describe the type of access we have for the function f we now define the following two oracles – which can be thought of as procedures that we assume that we are able to implement:

1. **Zeroth-order oracle**: Given $w \in \mathcal{W}$ we assume that there exists a procedure $\mathcal{O}_o(w)$ that returns the value $f(w)$.

2. **First-order oracle**: given $w \in \mathcal{W}$ we assume that there exists a procedure $\mathcal{O}_1(w)$ that returns $g \in \partial f(w)$.

References