MS&E 213 / CS 269O : Chapter 6
Non-smooth Convex Optimization *

By Aaron Sidford (sidford@stanford.edu)

November 18, 2020

1 The Problem

For many of the optimization algorithms we have discussed so far assumed that our objective function (or at least part of it) is differentiable. Even when we generalized our analysis of smooth convex function minimization to general norms or composite functions, we at least assumed that there was some way to make immediate progress on the function. Formally, we assumed that we could compute a *descent direction,* that is a direction to move in from the current point, that decreases the function value. Often, this was obtained by showing that for the objective function f it is possible to compute for any x a function $U_x : \mathbb{R}^n \to \mathbb{R}$ such that $U_x(y) = f(y)$ and $U_x(y) \geq f(y)$ for all $y \in \mathbb{R}^n$.

The primary question we address for the next several lectures is what to do when we no longer have these sorts of properties for the function we wish to minimize? How can we minimize a convex function when we cannot immediately make sufficient progress on decreasing the objective function, e.g. it is non-differentiable? Further, even if the function is differentiable, but with a very large value for smoothness, can we somehow obtain methods with improved convergence than what would be predicted by smoothness and strong convexity alone?

For the next few chapters we will address the question of how to solve

$$\min_{x \in \mathbb{R}^n} f(x)$$

when f is not necessarily differentiable, but is still convex. Some of the techniques we will develop generalize quite naturally (or are more clearly explained) in the case of *constrained optimization problem* when f is only defined over some set $S \subseteq \mathbb{R}^n$ that is convex and we may wish to solve the

$$\min_{x \in S} f(x) \,.$$

1.1 Why these Assumptions?

There are many reasons for considering problems of this type. First, they are prevalent and arise easily. For example, we have given efficient optimization methods for a variety of problems, e.g. smooth strongly-convex

*These notes are a work in progress. They are not necessarily a subset or superset of the in-class material, there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future, and citations are often omitted. These notes are intended converge to a superset of the class material that is TODO-free with a more complete set of citations and pointers to the literature. Your feedback is welcome and highly encouraged. If anything is unclear, you find a bug or typo, or if you would find it particularly helpful for anything to be expanded upon, please do not hesitate to post a question on the Piazza or contact me directly at sidford@stanford.edu.

functions, various composite functions, and that variants of max-type functions, i.e.

$$\min_{x \in \mathbb{R}^n} F(x) = \max_{i \in [m]} f_i(x) \tag{1.1}$$

where each f_i is convex and smooth.[1] However, the number of iterations of our methods for these problems depend polynomially on a notion of the condition number of the problem, e.g. the ratio of the smoothness to thee strong convexity. In general, this ratio could be arbitrarily large, e.g. $10n^{10}$, and in this case it may be desirable to have methods which depend on other measures of problem complexity. Further, in cases such (1.1), though efficient oracle complexities might be achievable, implementing the method could become computational expensive for large m.

Another common and popular and optimization problem for which we may wish to develop new techniques is *linear programming*. In this problem we have a *constraint matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$, *constraint vector* $b \in \mathbb{R}^m$, and *cost vector* $c \in \mathbb{R}^n$ and wish to solve

$$\min_{\mathbf{A}x \geq b} c^\top x \,.$$

This is one of the most fundamental problems in optimization and in some sense contains all of convex optimization as we will see. To get a better sense of the structure of this problem and the feasible region, consider the simpler geometric set known as a *half-space*.

Definition 1 (Half-space)**.** We call $S \subseteq \mathbb{R}^n$ a halfspace if for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ with $a \neq 0$ we have

$$S = \text{half}(a, b) \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^n \,:\, a^\top x \geq b \right\} .$$

Thus we see that the *feasible region,* $\mathbf{A}x \geq b$, is an intersection of half spaces, i.e. if we let $a_1, ..., a_m \in \mathbb{R}^n$ denote the rows of $\mathbf{A}$ written as vectors then

$$\{x \in \mathbb{R}^n \,:\, \mathbf{A}x \geq b\} = \cap_{i \in [m]} \text{half}(a_i, b_i) \,.$$

Such an intersection of finite number of half spaces is known as a *polytope*.

Another reason we consider this new assumptions we will consider is that they motivate a fundamental set of techniques in our broader optimization toolkit. Wheres in the previous section we could directly make objective function progress, here we will introduce new techniques to develop proxy functions for progress. Many of the algorithms we will introduce will work by locally making progress on some progress measure other than objective function value and we will need to argue about their ultimate connection to minimizing objective functions. This is a powerful technique for optimization more broadly and we will build on it in the last few sections when we discuss second-order optimization techniques.

1.2 Roadmap

We will build these new algorithms in several steps. First, in this chapter we take a closer look at the structure of convex functions. This will motivate new oracle assumptions we will make to solve these problems. In this chapter we prove the existence of these oracles for convex functions. In the next few chapters we will build on the results of this chapter, introducing new natural optimization problems and efficient algorithms to solve them.

2 Convex Sets

In the remainder of this chapter we begin to address these questions by studying the structure of convex sets. We do this for two reasons. First, it is an interesting area of study broadly useful for mathematics and optimization. Second, it motivate the oracles and algorithms we study in this chapter and the next few.

[1] This is a not necessarily smooth function for which smooth strongly-convex rates of function minimization can be achieved. In he homework this was shown for $m = 2$, but the algorithm and analysis can be extended to this more general case easily.

We begin by defining convex sets and giving several basic properties about them. Note that the structure of convex sets is a well-studied area of mathematics and optimization and there are courses denoted entirely too it. However, here we give just a few basic properties that we may use repeatedly in our analysis.

Definition 2 (Convex Set)**.** We say a a set $S \subseteq \mathbb{R}^n$ is *convex* if for all $x, y \in S$ and $t \in [0,1]$ it is the case that $t \cdot x + (1-t) \cdot y \in S$

This definition says that a set is convex if and only if for any two points in the set, the line between them is also contained in the set. It is easy to see that convex sets are closed under intersection and the closure operation.

Lemma 3 (Intersections of Convex Sets are Convex)**.** *Let $\mathcal{C}$ be a (possibly infinite) set of convex subsets of $\mathbb{R}^n$. Then $\cap_{S \in \mathcal{C}} S$ is a convex set.*

Proof. Suppose that $x, y \in \cap_{S \in \mathcal{C}} S$ and $t \in [0,1]$ is arbitrary. Then $x, y \in S$ for all $S \in C$ and by convexity $t \cdot x + (1-t) \cdot y \in S$ for all $S \in \mathcal{C}$. □

Lemma 4 (Closure of Convex Set is Convex)**.** *Suppose $S \subseteq \mathbb{R}^n$ is a convex set. Then C the closure of S, i.e. the union of all limit points of S, is convex.*

Proof. Let $x, y \in C$. Then there exist sequences $x_i, y_i \in \mathbb{R}^n$ such that $x_i, y_i \in S$ for all $i \in \mathbb{Z}_{>0}$ and $\lim_{t \to \infty} x_t = x$ and $\lim_{t \to \infty} y_t = y$. Now, if $\alpha \in [0,1]$ is arbitrary we have that $z_i = \alpha x_i + (1-\alpha) y_i \in S$ for all $i \in \mathbb{Z}_{>0}$ by convexity. Consequently,

$$\alpha \cdot x + (1-\alpha) \cdot y = \lim_{i \to \infty} z_i \in C\,.$$

□

Furthermore, we can show that half-spaces are always convex.

Lemma 5 (Half-spaces are Convex)**.** *For all $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ the half-space* $\text{half}(a, b)$ *is convex.*

Proof. Suppose that $x, y \in \text{half}(a, b)$ this implies that $a^\top x \geq b$ and $a^\top y \geq b$. Consequently, for all $t \in [0,1]$ we have that

$$a^\top [t \cdot x + (1-t) \cdot y] = t \cdot a^\top x + (1-t) \cdot a^\top y \geq t \cdot b + (1-t) \cdot b = b\,.$$

□

Note that from the lemmas we have seen this shows that any intersection of a (possibly infinite) number of half-spaces is convex. Consequently, polytopes are convex. Eventually we will show that all closed convex sets can be written this way, as an intersection of a possibly infinite number of half-spaces. Furthermore, the study of half-spaces associated with convex sets will be one of the primary tools we use to design algorithms.

3 From Convex Sets to Convex Function

The convexity of sets is closely related to the convexity of functions and convex sets arise natural in reasoning about decreasing objective function values of convex functions. Here we provide several such relations.

One such connection between convex sets and convex functions is through (sub)level sets, i.e. the set of all point whose value is at least a given value of a function.

Definition 6 ((Sub)level Sets)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ and $v \in \mathbb{R}$ we call $\text{level}_{\leq}(f, v) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : f(x) \leq v\}$ *a (sub)level set of f and* $\text{level}_{<}(f, v) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : f(x) < v\}$ a *strict v-(sub)level set.*

For convex functions it can be shown that all level sets are convex.

Lemma 7. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then* $\text{level}_{\leq}(f, v)$ *and* $\text{level}_{<}(f, v)$ *are convex for all* $v \in \mathbb{R}$.

Proof. If $x, y \in \text{level}_{\leq}(f, v)$ and $t \in [0, 1]$ then by definition $f(x) \leq v$ and $f(y) \leq v$. Consequently, by the convexity of f we have that

$$f(t \cdot x + (1-t) \cdot y) \leq t \cdot f(x) + (1-t) \cdot f(y) \leq t \cdot v + (1-t) \cdot v = v\,.$$

Consequently, $t \cdot x + (1-t) \cdot y \in \text{level}_{\leq}(f, v)$ and f is convex. The proof that $\text{level}_{<}(f, v)$ is convex is analogous. □

Note that the set of all ϵ-optimal points is simply, $\text{level}_{\leq}(f, f_* + \epsilon)$ and consequently Lemma 7 shows that the problem of finding an ϵ-optimal point is the same as thee problem of finding a point in a convex set. Further, since $\text{level}_{\leq}(f, f(x))$ is the set of all points whose value is at most that of x and $\text{level}_{<}(f, f(x))$ is the set of all points whose value is at most that of x, we see that reasoning about *descent directions*, i.e. directions to move from x which don't increase or decrease function value, are problems about finding points in convex sets. Leveraging this fact will motivate cutting plane methods that we will discuss in Chapter 8.

Interestingly, the converse is not true. There are functions where all level sets (and strict level sets are convex) and nevertheless the function is convex. As a simple example of this, consider the function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = 0$ and $f(x) = 1$ for all $x \neq 0$, this function is clearly not convex (consider any pair of points where one of them is 0) and since its only level sets are $\mathbb{R}$ and $\{0\}$, all its level sets are convex. More broadly, any 1-dimensional function where its values increase monotonically as the distance from the minimizer increases has convex level sets. The set of functions with convex level sets is known as quasi-convex functions and below we give an alternative definition of them and prove this equivalence.

Definition 8 (Quasi-convex Functions)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex if for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ it is the case that $f(t \cdot x + (1-t) \cdot y) \leq \max\{f(x), f(y)\}$.

Lemma 9. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex if and only if all of its level sets are convex.*

Proof. Homework □

Although, the convexity of level sets does not imply that a function is convex, it is possible to characterize the convexity of functions through the convexity of sets. To show this connection we define a natural set associated with a function, known as its *epigraph*.

Definition 10 (Epigraph)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ its *epigraph* $\text{epi}(f) \subseteq \mathbb{R}^{n+1}$ is defined as

$$\text{epi}(f) \stackrel{\text{def}}{=} \{(x, v) \,|\, x \in \mathbb{R}^n, v \in \mathbb{R}, f(x) \leq v\}\,.$$

In the following lemma we prove that a function is convex if and only if its epigraph is a convex set.

Lemma 11. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if* $\text{epi}(f)$ *is a convex set.*

Proof. If f is convex, then if $(x, v_x) \in \text{epi}(f)$ and $(y, v_y) \in \text{epi}(f)$ then $f(x) \leq v_x$ and $f(y) \leq v_y$ and for all $t \in [0, 1]$ we have

$$f(t \cdot x + (1-t) \cdot y)) \leq t \cdot f(x) + (1-t) \cdot f(y) \leq t \cdot v_x + (1-t) \cdot v_y$$

and consequently $t \cdot (x, v_x) + (1-t) \cdot (y, v_y) \in \text{epi}(f)$.

On the other hand if $\text{epi}(f)$ is convex, then for all $x, y \in \mathbb{R}^n$ we have $(x, f(x)), (y, f(y) \in \text{epi}(f)$ and therefore, for all $t \in [0, 1]$ we have $t \cdot (x, f(x)) + (1-t) \cdot (y, f(y)) \in \text{epi}(f)$ so $f(t \cdot x + (1-t) \cdot y) \leq t \cdot f(x) + (1-t) \cdot f(y)$. □

Consequently, by understanding properties of convex sets we can understand properties of convex functions. In the next sections we motivate the structural properties and oracles about convex sets that we will use and prove that properties and existence of the oracles.

4 Oracles and Separating Convex Sets

So what structure of convex sets can we exploit to design fast optimization algorithms in the absence of smoothness and strong convexity? There are two ways to motivate the structure we will consider.

The first stems from our previous discussion of the intersection of half-spaces being convex sets. We could try to make a converse statement and argue that all convex sets are intersections of half-spaces and therefore we can understand convex sets by finding the half-spaces that border them. This motivates the notion of separating hyperplanes that we will introduce and prove exist in the next section.

Another nice way to motivate our approach is to again consider differentiable convex functions. Recall that if f is a differentiable convex function then for all $x, y \in \mathbb{R}^n$ we have that $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$. Consequently, if $f(y) \leq f(x)$ then it must be the case that $\nabla f(x)^\top y \leq \nabla f(x)^\top x$. Furthermore, this implies that $\text{level}_{\leq}(f, f(x)) \subseteq \text{half}(-\nabla f(x), -\nabla f(x)^\top x)$. Consequently, we have that the gradient always induces a half-space that is on the boundary of the level set for where the gradient is computed. Even if a function is differentiable but very non-smooth we see that the gradient may not let us make a lot of function progress necessarily, but it does gives us useful information about where the minimizer of f might lie.

Consequently, one natural oracle to assume for minimizing a convex function is a *separation oracle* defined below.

Definition 12 (Function Separation Oracle)**.** A *separation oracle* for a function $f : \mathbb{R}^n \to \mathbb{R}$ is an oracle which when queried at a point $x \in \mathbb{R}^n$ outputs a vector $g \in \mathbb{R}^n$ such that $\text{level}_{\leq}(f, f(x)) \subseteq \text{half}(g, g^\top x))$ with $g = 0$ if and only if $f_* = f(x)$.

Note that statement regarding when $g = 0$ is present to rule out the trivial output of $g = 0$ in which case $\text{half}(g, g^\top x) = \mathbb{R}^n$. Also note that often separation oracles are defined in terms of sets, rather than functions. However, we use this terminology due to the close connection between the two.

In the last section of these notes we will show that a separation oracle in principle always exists if f is convex and in Chapter 8 we will discuss optimization algorithms which use a function separation oracle.

Another, natural idea is to simply attempt to generalize the notion gradients of differentiable convex functions to non-differentiable convex functions. We know $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and convex if and only if $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in \mathbb{R}^n$ and consequently, we could simply assume that for all x there exists a vector g_x with $f(y) \geq f(x) + g_x^\top (y - x)$. Such a vector is known as a *subgradient* and an oracle that outputs subgradients is known as a subgradient oracle. We define each below. (note that we define these oracles even for functions with restricted domains S as we may consider constrained function minimization instances later in the course.)

Definition 13 (Subgradient)**.** For $f : S \to \mathbb{R}$ and $S \subseteq \mathbb{R}^n$ we say that $g_x \in \mathbb{R}^n$ is a subgradient of f at $x \in S$ if for all $y \in S$ it is the case that $f(y) \geq f(x) + g_x^\top (y - x)$. Further, for all $x \in \mathbb{R}^n$ we let $\partial f(x)$ denote the set of all subgradients of f at x.

Definition 14 (Subgradient)**.** For $f : S \to \mathbb{R}$ and $S \subseteq \mathbb{R}^n$ we say that $g_x \in \mathbb{R}^n$ is a subgradient of f at $x \in S$ if for all $y \in S$ it is the case that $f(y) \geq f(x) + g_x^\top (y - x)$. Further, for all $x \in \mathbb{R}^n$ we let $\partial f(x)$ denote the set of all subgradients of f at x.

Definition 15 (Subgradient Oracle)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ as *subgradient oracle* is an oracle that given query point $x \in \mathbb{R}^n$ returns $g_x \in \partial f(x)$.

In the next chapter, we will show how subgradient oracles suffice to perform convex optimization. In this chapter, we prove that subgradients and subgradient oracles exist on the interior of the domain of f if f is convex on its domain.

5 Separating Hyperplane Theorem

To show the existence of separation and subgradient oracles, we leverage a general property of convex sets, namely the existence of of separating hyperplanes.

Definition 16 (Separating Hyperplane)**.** For a set $S \subseteq \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$ we say that for $g \in \mathbb{R}^n$ and $c \in \mathbb{R}$ the hyperplane $H_=(g,c) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : g^\top x = c\}$ is *separating* if for all $x \in S$ it is the case that $g^\top x_0 \leq c \leq g^\top x$, i.e. $x_0 \in H_\leq(g,c)$ and $S \subseteq H_\geq(g,c)$.

The main theorem we will prove regarding convex sets is that whenever S is convex and x_0 is either on the boundary of S or not in S then there is a separating hyperplane through x_0. In the particular case when x_0 is on the boundary of S, such a hyperplane is called a supporting hyperplane. Below, we recall the definition of the boundary of a set and provide the theorem.

Definition 17 (Boundary)**.** For $S \subseteq \mathbb{R}^n$ the boundary of S, denote ∂S, is the set of points x such that there is both a sequence of points in S and a sequence of points not in S that converge to x, i.e. x is in the closure of S and its complement.

Theorem 18 (Separating Hyperplane Theorem)**.** *For any convex $S \subseteq \mathbb{R}^n$ and $x_0 \notin S$ or $x_0 \in \partial(S)$ there exists $g \neq \vec{0} \in \mathbb{R}^n$ such that $S \subseteq H_\geq(g, g^\top x_0)$.*

Note that for many convex sets computing the separating hyperplanes guaranteed to exist by Theorem 18 can be quite easy. For example, for any polytope, i.e. $S = \{x \in \mathbb{R}^n : \mathbf{A}x \geq b\} = \cap_{i\in[m]}\{x \in \mathbb{R}^n | a_i^\top x \geq b_i\}$ with non-zero a_i if we are given a point $x_0 \notin S$ or $x_0 \in \partial(S)$, we can find a separating hyperplane by outputting g as any a_i with $a_i^\top x \leq b_i$. Further, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and $x_0 = (x, v_x) \notin \text{epi}(f)$ or $(x, v_x) \in \partial(\text{epi}(f))$ then $g = (-\nabla f(x), 1)$ is a valid separating hyperplane. To see this, note that the assumption (and that convex f are continuous) implies that $f(x) \geq v_x$. Consequently, if $(y, v_y) \in \text{epi}(f)$, i.e. $f(y) \leq v_y$ then by convexity and the assumptions

$$v_y \geq f(y) \geq f(x) + \nabla f(x)^\top (y - x) \geq v_x + \nabla f(x)^\top (y - x)$$

which can be written a $g^\top (y, v_y) \geq g^\top (x, v_x)$.

In addition to prove Theorem 18, we will prove the following strict separating hyperplane theorem which hows that whenever S is convex and closed and $x \notin S$ then there exists $g \neq 0$ and c such that $g^\top x_0 < c < g^\top x$ for all $x \in S$, or equivalently $S \subseteq H_\geq(g, g^\top x_0 + \epsilon)$ for $\epsilon = c - g^\top x_0 > 0$.

Theorem 19 (Strict Separating Hyperplane Theorem)**.** *For any closed convex $S \subseteq \mathbb{R}^n$ and $x_0 \notin S$ there exists a* strict separating hyperplane*, i.e. $g \neq \vec{0} \in \mathbb{R}^n$ and $\epsilon > 0$ such that $S \subseteq H_\geq(g, g^\top x_0 + \epsilon)$.*

Interestingly, this Theorem is stronger than what is needed to prove Theorem 18. Below we prove Theorem 18 using Theorem 18. Note that in this proof, Theorem 19 where ϵ is replaced with 0 would have sufficed.

of Theorem 18. Note that the convex closure of S is convex and contains S. Further, if $x_0 \notin S$ or $x_0 \in \partial(S)$ then one of these is still true when S is replaced with it closure. Consequently it suffices to prove the claim for the convex closure of S and we simply assume that S is closed.

In the case that S is closed and $x_0 \notin S$ the result follows immediately from Theorem 18. To see this, note that Theorem 18 implies that there exists $g \neq \vec{0} \in \mathbb{R}^n$ and $\epsilon > 0$ such that $S \subseteq H_\geq(g, g^\top x_0 + \epsilon)$. However, since $H_\geq(g, g^\top x_0 + \epsilon) \subseteq H_\geq(g, g^\top x_0)$ the result follows.

Consequently, it suffices to consider the case where S is closed and $x_0 \in \partial(S)$. Since, $x_0 \in \partial(S)$ there exist an infinite sequence of points $x_1, x_2 \ldots$ such that $\lim_{k\to\infty} x_k = x_0$ and $x_i \notin S$ for all $i \geq 1$. Further, by Theorem 19 there exist $g_i \neq \vec{0} \in \mathbb{R}^n$ and $\epsilon_i > 0$ such that $S \subseteq H_\geq(g_i, g_i^\top x_0 + \epsilon_i) \subseteq H_\geq(g_i, g_i^\top x_0)$. Further, we can assume without loss of generality that each $\|g_i\|_2 = 1$ as $H_\geq(g_i g_i^\top x_0 + \epsilon_i) = H(\frac{1}{\|g_i\|_2}, \frac{1}{\|g_i\|}[g_i^\top x_0 + \epsilon_i])$.

This implies that there is a convergent subsequence of the g_i and consequently we also assume without loss of generality that $\lim_{i\to\infty} g_i$ converges and let $g_0 = \lim_{i\to\infty} g_i$.

In summary, we have that for all $x \in S$ it is the case that $g_i^\top x \geq g_i^\top x_i$ and $\lim_{i\to\infty} x_i = x_0$ and $\lim_{i\to\infty} g_i = g_0$. This implies that

$$0 \leq \lim_{i\to\infty} g_i^\top (x - x_i) = g_0^\top x - g_0^\top x_0$$

and since this holds for all $x \in S$ we have the desired result that $S \subseteq H_{\geq}(g_0, g_0^\top x_0)$. □

Interestingly, this theorem can also be used to obtain characterization of all closed convex sets!

Lemma 20. *If S is a closed convex set then there exists C a possibly infinite set of halfspaces, $h = H_{\geq}(a, b)$ such that $S = \cap_{h\in C} h$.*

Proof. For all $x \notin S$ let $h_x = H_{\geq}(g_x, g_x^\top x + \epsilon_x)$ where $g_x \neq 0$ and $\epsilon_x > 0$ are given by Theorem 19, i.e. $S \subseteq h_x$, and let $C = \{h_x | x \notin S\}$. Now if $x \notin S$ then $x \notin \cap_{h\in C} h$ as $x \notin h_x$. Further, if $y \in S$ then $y \in h_x$ for all $h_x \in C$ as $S \subseteq h_x$ for al $x \in S$. Consequently, $S = \cap_{h\in C} h$. □

This essentially shows that all closed convex sets are essentially polytopes, where the number of halfspaces are allowed to go to infinity. Further, the proof of this lemma shows that the only sets for which Theorem 19 holds are closed convex sets, since the intersections of halfspaces are always closed and convex.

In the remainder of this chapter we prove the strict separating hyperplane theorem and use this to prove existence of the separation and subgradient oracles for convex sets. This is part of a richer theory about duality between convex sets and half-spaces that induce them which we will only touch upon. Rather, in the lectures to come, our emphasis will be to design efficient algorithms.

6 Proof of Separating Hyperplane Theorem

Here we prove Theorem 19 which shows that closed convex sets always have strict separating hyperplanes for point not in the set.

So how should we compute a separating hyperplane and thereby prove it exists? Note that for a set S and $x \notin S$ we are looking for a plane that slices through the line between x and every point in S. Thus to slice all these lines it seems like we want to find an extreme point of S, i.e. one that is closest to x in some way. To analyze this we need to determine what happens when we minimize a convex function over a convex set.

We perform this analysis in several parts. First we state the following somewhat standard lemma from analysis that says when continuous functions obtain their minimum or maximum value on subsets of $\mathbb{R}^n$.

Lemma 21 (Multivariable Extreme Value)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $S \subseteq \mathbb{R}^n$ is closed, bounded, and non-empty then there exists $x_* \in S$ such that $f(x) \geq f(x_*)$ for all $x \in S$.*

Next, using this lemma we prove that strongly convex functions always obtain their minimum value over closed convex sets.

Lemma 22. *If $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and μ-strongly convex with respect to some norm $\|\cdot\|$ and $\mu > 0$ and $S \subseteq \mathbb{R}^n$ is a non-empty closed set, then there exists $x_* \in S$ such that $f(x) \geq f(x_*)$ for all $x \in S$.*

Proof. Let $x_0 \in S$ be arbitrary. Since f is differentiable and μ-strongly convex for we have that for all $y \in S$ it is the case that $f(y) \geq f(x_0) + \nabla f(x_0)^\top (y - x_0) + \frac{\mu}{2}\|y - x_0\|^2$. Consequently, if $\|y - x_0\| > \frac{2}{\mu}\|\nabla f(x_0)\|_*$ by Cauchy Schwarz

$$f(y) \geq f(x_0) - \|\nabla f(x_0)\|_* \cdot \|y - x_0\| + \frac{\mu}{2}\|y - x_0\|^2 > f(x_0)$$

and therefore if we let $B \stackrel{\text{def}}{=} \{y : \|y - x_0\| \leq \frac{2}{\mu}\|\nabla f(x_0)\|_*\}$ then

$$\inf_{x \in S} f(x) = \inf_{x \in S \cap B} f(x)$$

However, since B is closed we have that $S \cap B$ is closed and since $x_0 \in S \cap B$ we have that $S \cap B$ is nonempty and since B is bounded so is $S \cap B$. Consequently by the Multivariable Extreme Value Theorem (Lemma 21) the result follows. □

Next we characterize the minimizer of a convex differentiable function over closed convex S.

Lemma 23 (Constrained Convex Minimizers)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function and $S \subseteq \mathbb{R}^n$ is a non-empty closed convex set, then x_* is a minimizer of f over S, i.e. $x_* \in S$, i.e. $f(x_*) \leq f(x)$ for all $x \in S$, if and only if $\nabla f(x_*)^\top (x - x_*) \geq 0$ for all $x \in S$.*

Proof. Since $f(x) \geq f(x_*) + \nabla f(x_*)^\top (x - x_*)$ by convexity we have that if $\nabla f(x_*)^\top (x - x_*) \geq 0$ for all $x \in S$ then $f(x) \geq f(x_*)$ for all $x \in S$.

Consequently, it just remains to show that when $\nabla f(x_*)^\top (x - x_*) < 0$ for some $x \in S$ then $\exists y \in S$ with $f(y) < f(x_*)$. To prove this, let $x_t = t \cdot x + (1 - t)x_*$ and let $g(t) \stackrel{\text{def}}{=} f(x_t)$. Note that by convexity $x_t \in S$ for all $t \in [0, 1]$ and we have

$$\lim_{\delta \to 0} \frac{g(0 + \delta) - g(0)}{\delta} = g'(0) = \nabla f(x_*)^\top (x - x_*) < 0$$

and since $g(0 + \delta) - g(0) = f(x_* + \delta(x - x_*)) - f(x_*)$ we see that for small enough $\delta \in [0, 1]$ it is the case that $y = x_* + \delta(x - x_*)$ satisfies $y \in S$ and $f(y) < f(x_*)$. □

These lemmas give us everything we need to prove separation oracles exist. Suppose we have a strongly convex differentiable function f such that $f(x_0) < f(x)$ for all $x \in S$. Then from the above lemmas we see that there is some x_*such that $\nabla f(x_*)^\top (x - x_*) \geq 0$ for all $x \in S$. However, by convexity we know that $f(x_0) \geq f(x_*) + \nabla f(x_*)^\top (x_0 - x_*)$ and therefore as $f(x_0) < f(x_*)$ we have $\nabla f(x_*)^\top x_0 < \nabla f(x_*)^\top x_*$. Consequently, we have that $H_{\geq}(\nabla f(x_*), \nabla f(x_*)^\top x_0 + \epsilon)$ would be a strict separating hyperplane for some $\epsilon > 0$.

So how to we get such a strongly convex function f? A natural choice would be to pick $f(x) = \frac{1}{2}\|x - x_0\|_2^2$. However, there is a general way to construct such a f. Suppose g is a differentiable μ-strongly convex function. Then $f(x) \stackrel{\text{def}}{=} g(x) - \left[g(x_0) + \nabla g(x_0)^\top (x - x_0)\right]$ is a μ-strongly convex function that obtains its minimum at x_0, since $\nabla f(x_0) = \nabla g(x_0) - \nabla g(x_0) = \vec{0}$. This is known as a *Bregman Divergence* and we will study them more later. However, for now we prove our separating hyperplane theorems with $f(x) = \frac{1}{2}\|x - x_0\|_2^2$.

To simplify our notation we define the following.

Definition 24. For closed convex set $S \subseteq \mathbb{R}^n$ and $x_0 \notin S$ we define the projection operator $\pi_S : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\text{argmin}_{x \in S} \frac{1}{2}\|x - x_0\|_2^2$$

and call the $\pi_S(x)$ the projection of x_0 onto x.

We remark that projection obeys a type of Pythagorean Theorem, showing that the angle between x_S, $\pi_S(x_0)$, and x_0 for any $x_s \in S$ and x_0 is obtuse.

Lemma 25. *For all closed convex sets S and $x_0 \in \mathbb{R}^n$ and $x_S \in S$ we have that*

$$\|x_S - \pi_S(x_0)\|_2^2 + \|\pi_S(x_0) - x_0\|_2^2 \leq \|x_S - x_0\|_2^2$$

Proof. Letting $A \stackrel{\text{def}}{=} \|x_S - x_0\|_2^2 - \|\pi_S(x_0) - x_0\|_2^2 - \|x_S - \pi_S(x_0)\|_2^2$ we have

$$\begin{aligned} A &= \|x_S\|_2^2 - 2x_S^\top x_0 + \|x_0\|_2^2 - \left[\|\pi_S(x_0)\|_2^2 - 2x_0^\top \pi_S(x_0) + \|x_0\|_2^2\right] \\ &\quad - \left[\|x_S\|_2^2 - 2x_S^\top \pi_S(x_0) + \|\pi_S(x_0)\|_2^2\right] \\ &= 2\left[x_S^\top \pi_S(x_0) + x_0^\top \pi_S(x_0) - x_S^\top x_0 - \|\pi_S(x_0)\|_2^2\right] \\ &= 2\left[(\pi_S(x_0) - x_0)^\top (x_S - \pi_S(x_0))\right] \geq 0 \end{aligned}$$

Where we used Lemma 23 to conclude the final line. □

We now formally show that this gives us strict separating hyperplanes for closed convex sets by proving Theorem 19.

Theorem 26. *Let $S \subseteq \mathbb{R}^n$ a closed convex set and $x_0 \notin S$. Then for all $x \in S$ we have*

$$(\pi_S(x_0) - x_0)^\top (x - \pi_S(x_0)) \geq 0$$

Consequently, for $g = (\pi_S(x_0) - x_0)$ and $c = g^\top x_0 + \|g\|_2^2$ we have that $g^\top x \geq c$ for all $x \in S$ and $H(g, c - \delta)$ is a strict separating hyperplane for x_0 and S for all $\delta \in (0, \|g\|_2^2)$.

Proof. Since $\nabla\left(\frac{1}{2}\|x - x_0\|_2^2\right) = x - x_0$, Lemma 23 implies that for all $x \in S$

$$(\pi_S(x_0) - x_0)^\top (x - \pi_S(x_0)) \geq 0\,.$$

Consequently, for $g = \pi_S(x_0) - x_0$ we have that for al $x \in S$

$$g^\top x \geq g^\top \pi_S(x_0) = g^\top x_0 + g^\top(\pi_s(x_0) - x_0) = g^\top x_0 + \|g\|_2^2\,.$$

The result then follows by picking $\epsilon = \|g\|_2^2$ and noting that $\epsilon > 0$ as $\pi_S(x_0) - x_0 \neq 0$ due to the fact that $\pi_S(x_0) \in S$ and $x_0 \notin S$. □

7 Oracle Existence for Convex Functions

Here, we conclude the chapter by using the separating hyperplane theorem to prove that separation oracles and subgradient oracles exist for convex functions. First we prove the claim for separation oracles.

Lemma 27. *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex then there exists a separation oracle for f.*

Proof. Let $x \in \mathbb{R}^n$ be arbitrary. If $f(x) = f_*$ then it is valid for the oracle to output $\vec{0}$. However, if this is not the case, then there exists a point x_* with $f(x_*) < f(x)$. For all $\epsilon > 0$ let $x_\epsilon = x + \epsilon(x - x_*)$. Choosing $t_\epsilon = (1 + \epsilon)^{-1}$ so that $t_\epsilon \cdot x_\epsilon + (1 - t_\epsilon)x_* = x$ we see that by convexity

$$f(x) \leq t_\epsilon \cdot f(x_\epsilon) + (1 - t_\epsilon) \cdot f(x_*) < t_\epsilon \cdot f(x_\epsilon) + (1 - t_\epsilon) \cdot f(x)$$

and therefore $f(x) < f(x_\epsilon)$. Consequently, we see that $x \in \partial(\text{level}_{\leq}(f, f(x)))$ as $x \in \text{level}_{\leq}(f, f(x))$ and $\lim_{\epsilon \to 0^+} x_\epsilon = x$ with $x_\epsilon \notin \text{level}_{\leq}(f, f(x))$ for all $\epsilon > 0$. Therefore, by Theorem 18 there exists a $g \neq 0$ with $\text{level}_{\leq}(f, f(x)) \subseteq H_{\geq}(g, g^\top x)$ as desired. □

Next, we prove the claim for subgradients of convex functions defined over convex sets. Note up to this point we have only defined convexity for functions defined over all of $\mathbb{R}^n$, however the definition extends naturally to functions defined over convex sets, i.e. we say that $f : S \to \mathbb{R}$ is convex for convex S if for all $x, y \in S$ and $t \in [0, 1]$ it is the case that $f(t \cdot x + (1 - t) \cdot y) \leq t \cdot f(x) + (1 - t) \cdot f(y)$. We conclude by showing that such functions have subgradients at all points in the interior of S, where we recall the definition of the interior below.

Definition 28 (Interior)**.** For $S \subseteq \mathbb{R}^n$ we denote the interior of S by $\text{int}(S) = S \setminus \partial(S)$ that is, the set of points in S that are not reachable by a convergent sequence of points not in S.

Note that equivalently we can view the interior of a convex set as a point in the set such that the set contains a ball centered around the point.

Lemma 29. *Let $S \subseteq \mathbb{R}^n$ be a convex set and $f : S \to \mathbb{R}$ be convex function. Then $\partial f(x) \neq \emptyset$ for all $x \in \text{int}(S)$ and consequently a subgradient oracle exists whenever S is open.*

Proof. Let $x \in \text{int}(S)$ be arbitrary. Note that $(x, f(x)) \in \text{epi}(f)$ by definition and $(x, f(x) + t) \notin \text{epi}(f)$ for all $t < 0$. Consequently, $(x, f(x)) \in \partial(\text{epi}(f))$ and by the separating hyperplane theorem there is a vector $g \in \mathbb{R}^n$ and $v_g \in \mathbb{R}$ with either $g \neq 0$ or $v_g \neq 0$ such that for all $(y, v_y) \in \text{epi}(f)$ we have

$$g^\top y + v_g \cdot v_y \geq g^\top x + v_g \cdot f(x) \tag{7.1}$$

Further, since $(x, f(x) + 1) \in \text{epi}(f)$ this implies that

$$g^\top x + v_g \cdot [f(x) + 1] \geq g^\top x + v_g \cdot f(x)$$

and rearranging yields that $v_g \geq 0$. Further, since $(y, f(y)) \in \text{epi}(f)$ for all $y \in S$ applying (7.1) again yields that

$$v_g \cdot f(y) \geq v_g \cdot f(x) + [-g]^\top (y - x) \tag{7.2}$$

Note that if $v_g \neq 0$ then dividing both sides of (7.2) and leveraging that $v_g \geq 0$ yields that $-v_g^{-1} g \in \partial f(x)$. Consequently, it suffices to how $v_g \neq 0$. However, since $x \in \text{int}(S)$ it is the case that $x_t \stackrel{\text{def}}{=} x - tg \in S$ for small enough $t > 0$. Substituting $y = x_t$ in (7.2) yields that $v_g \cdot [f(x_t) - f(x)] \geq t \cdot \|g\|_2^2$. Consequently, if $g \neq 0$ then $v_g \neq 0$. On the other hand, if $g = 0$ then $v_g = 0$ by the separating hyperplane theorem. Consequently, in either case $v_g \neq 0$ and the result follows. □