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

By Aaron Sidford (sidford@stanford.edu)

May 28, 2017

1 The Problem

So far most optimization algorithms have assumed that our objective function is differentiable. Even when we generalized further to the case of smoothness in other norms or composite functions we at least assumed that there was some way to make immediate progress on the function, i.e. that we could compute a *descent direction,* that is a direction that decreases the value. Note that the analysis in most of the lectures we have seen so far assumed that to minimize f we could compute for all 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 $x \in \mathbb{R}^n$.

The primary question we address for the next several lectures is what to do when we no longer have these properties. How do we minimize a convex function when we cannot immediately make sufficient progress on decreasing the objective function, e.g. it is non-differentiable. Note that there are cases where we might want to use the techniques we will introduce even when functions are differentiable, as they may be so non-smooth or non-strongly convex that the algorithms we have seen before will converge fairly slowly.

Consequently, the main subject we wish to address in this section is how do we solve

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

when f is no longer differentiable, but is still convex. Some of the techniques generalize quite naturally (or are more clearly explained) in the case when f is only defined over some set $S \subseteq \mathbb{R}^n$ that is convex and we may wish to solve the *constrained optimization problem*

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

1.1 Why these Assumptions?

There are two primary reasons for considering problems of this tight. The first is that hey are quite prevalent and arise easily. As we have already seen

$$\min_{x \in \mathbb{R}} \max_{i \in [n]} f_i(x)$$

[*] These notes are a work in progress. They are not necessarily a subset or superset of the in-class material and there may also be occasional *TODO* comments which demarcate material I am thinking of adding in the future. These notes will converge to a superset of the class material that is TODO-free. 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 discussion board or contact me directly at sidford@stanford.edu.

is convex function if the f_i are convex, but it may not be differentiable if the f_i are not differentiable and although we have given techniques for smooth minimization, these may slow down tremendously when the number of f_i are large.

Another common example of this is *linear programming* where we have a *constraint matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ 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 *halfspace*.

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

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

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

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

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

We may also wish to solve problems like

$$\min_{x\in\mathbb{R}^n} \|\mathbf{A}x - b\|_2^2 + \lambda \cdot \|x\|_1$$

a common form of regularized regression that shows up in machine learning.

Second, the reason we consider this new set of assumptions 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 these notes we take a closer look at the structure of convex functions. This will motivate the sort of oracle assumptions we will make to solve these problems. In the next few chapters we then build on these structural facts of convex sets, introducing new natural optimization problems and efficient algorithms to solve them.

2 Convex Sets

In this set of notes we begin to address these questions by probing deeper into the 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 definitions and the algorithms we will see in the next unit.

Our first observation is that we can tie the structure of convex functions to the structure of convex sets. Here we formally define convex sets and analyze this connection.

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 the line between any two points in the set is contained in the set. This notion is closely related to the convexity of functions. Formally, this mapping is obtained by associating functions with sets through *epigraphs*.

Definition 3 (Epigraph)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ its *epigraph* $\mathrm{epi}(f) \subseteq \mathbb{R}^{n+1}$ is define as

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

Now we prove that the mapping between convexity of functions and epigraphs is quite type, a function is convex if and only if its epigraph is.

Lemma 4. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if $\mathrm{epi}(f)$ is a convex set.*

Proof. If f is convex, then if $(x, v_x) \in \mathrm{epi}(f)$ and $(y, v_y) \in \mathrm{epi}(f)$ 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 $t \cdot (x, v_x) + (1-t) \cdot (y, v_y) \in \mathrm{epi}(f)$.
On the other hand if $\mathrm{epi}(f)$ is convex, then for all $x, y \in \mathbb{R}^n$ we have $(x, f(x)), (y, f(y) \in \mathrm{epi}(f)$ and therefore, for all $t \in [0,1]$ we have $t \cdot (x, f(x)) + (1-t) \cdot (y, f(y)) \in \mathrm{epi}(f)$ 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. Another way to see this connection is to consider sublevel sets of convex functions.

Definition 5 (sublevel set)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ and $v \in \mathbb{R}$ we define the *sublevel* set $\mathrm{level}_f(v) \overset{\text{def}}{=} \{x \in \mathbb{R}^n : f(x) \leq v\}$.

Note that if we have a function $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$ then $\mathrm{level}_f(f(x))$ is the set of all points whose value is at most that x. In other-words, if we were going to make an algorithm that attempts to look for points of smaller value, this is the set that should be consider. In other-words, directions moving into this set are *descent directions* we would want to consider in an algorithm.

Lemma 6. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then for all $v \in \mathbb{R}$ the set $\mathrm{level}_f(v)$ is convex. Furthermore, $D \overset{\text{def}}{=} \{x \in \mathbb{R}^n \, : f(x) < v\}$ is convex.*

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

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

This yields that $\mathrm{level}_f(v)$ is convex. The convexity of D follows from the fact that if $f(x) < v$ and $f(y) < v$ then the inequality above is strict as well. □

Note that the converse of this lemma does not hold and we prove this in the homework. However, this lemma does show that if we can understand the structure of convex sets well enough we can perhaps build better algorithms for minimizing them.

3 Basic Properties of Convex Functions

Now the study of convex sets is a wide area of mathematics and optimization and there are courses denoted entirely too it. However, here we will just give a few basic properties that we may use repeatedly in our analysis.

Lemma 7 (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 8 (Closure of Covnex 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,

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

□

Lemma 9 (Halfspaces are Convex). *For all $a \in \mathbb{R}^n$ with $a \neq 0$ and $b \in \mathbb{R}$ the halfspace* $\mathrm{half}(a,b)$ *is convex.*

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

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

□

Note that from the lemmas we have seen this shows that any intersection of a (possibly infinite) number of halfspaces 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 halfspaces. Furthermore, the study of halfspaces associated with convex sets will be one of the primary tools we use to design algorithms.

However, before investigating this characterization further we conclude this section with one more interesting and useful lemma about the boundary of convex sets.

Definition 10 (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.

Lemma 11 (Boundary of Closed Convex Set). *For a convex set $S \subseteq \mathbb{R}^n$ if $x \in \partial S$ then there is a vector $d \in \mathbb{R}^n$ such that for all $\alpha > 0$ we have $x + \alpha d \notin S$.*

Proof. First suppose $x \in \partial S$. Let $x_i \in \mathbb{R}^n$ be infinite sequences such that $\lim_{t \to \infty} x_i = x$ and $x_i \notin S$ for all $i \in \mathbb{Z}_{>0}$. Now, let $d_i = \frac{x_i - x}{\|x_i - x\|_2}$. Since $\|d_i\|_2 = 1$ for all $i \in \mathbb{Z}_{>0}$ there exists a subsequence where d_i converges. Thus we assume without loss of generality that $\lim_{t \to \infty} d_i = d$ for some $d \in \mathbb{R}^n$.

Now proceed by contradiction and suppose that $x + \alpha \cdot d \in S$ for some $\alpha \in \mathbb{R}$ □

4 Separating Convex Sets

So what structure of convex sets should we exploit to design fast optimization algorithms?

There are two ways to motivate the structure we will consider. The first stems from our previous discussion of the intersection of halfspaces being convex sets. We could try to make a converse statement and argue that all convex sets are intersections of halfspaces and therefore we can understand convex sets by finding the halfspaces that border them. This motivates the notion of separating hyperplanes that we will ultimately consider in these notes

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 $f(y) \geq f(x) + \bigtriangledown f(x)^\top (y - x)$. Consequently, if $f(y) \leq f(x)$ then it must be the case that $\bigtriangledown f(x)^\top y \leq \bigtriangledown f(x)^\top x$. Furthermore, this implies that $\mathrm{level}_f(f(x)) \subseteq \mathrm{half}(-\bigtriangledown f(x), -\bigtriangledown f(x)^\top x)$. Consequently, we have that the gradient always induces a halfspace that is on the boundary of the level set for where the gradient is computed. Consequently, 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.

In the case of non-differentiable f this natural generalization of this concept is a *subgradient.*

Definition 12 (Subgradient)**.** For $f : \mathbb{R}^n \to \mathbb{R}$ we say that $g \in \mathbb{R}^n$ is a subgradient of f at $x \in \mathbb{R}^n$ if for all $y \in \mathbb{R}^n$ it is the case that $f(y) \geq f(x) + g^\top (y - x)$.

Ultimately, we will show how subgradient oracles, that is algorithms which compute subgradients, suffice to perform convex optimization. Moreover, we will eventually prove that subgradients exists rather generically fo convex functions.

Here, we show that the existence of objects like sub-gradients stems from a more general property of convex sets, namely that of separating hyperplanes.

Definition 13 (Separating Hyperplane)**.** For a set $S \subseteq \mathbb{R}^n$ and $x_0 \in \mathbb{R}$ we say that for $g \in \mathbb{R}^n$ and $c \in \mathbb{R}$ the hyperplane $H(g, c) \stackrel{\mathrm{def}}{=} \{x \in \mathbb{R}^n : g^\top x = c\}$ is separating if for all $x \in S$ it is the case that $g^\top x_0 \leq c \leq g^\top x$. We call the separation strict if the inequalities are all strict and we call the hyperplane supporting if $c = g^\top x_0$.

Note that computing separating hyperplanes for many convex sets can be quite easy. For example, for any convex set S given as an intersection of half-spaces, e.g. a polytope, $S = \{x \in \mathbb{R}^n : \mathbf{A}x \geq b\}$ if we are given a point $y \notin S$, we can find a separating hyperplane by simply finding one of the halfspaces, $H = \mathrm{half}(a, b)$ in the intersection such that $y \notin H$ or $a^\top y = b$ and the return $H(a, b)$.

In the remainder of this chapter we prove that separating hyperplanes between convex sets and points on the boundary of the set or outside the set always exist. This is part of a richer theory about duality between convex sets and halfspaces that induce them which we will only touch upon. Rather, in the lectures to come, our emphasis will be to design effecient algorithms and we will prove further structure mostly as we need it, e.g. to prove that subgradients often exist and design efficient algorithms.

5 Separating Hyperplane Theorem

Here we prove that convex sets and always have separating hyperplanes them from points on their boundary or outside them all together.

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 give 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 14 (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 15. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and μ-strongly convex with respect to $\|\cdot\|$ for some norm for $\mu > 0$. Furthermore, let $S \subseteq \mathbb{R}^n$ be a non-empty closed set. The 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 $f(y) \geq f(x_0) + \bigtriangledown f(x_0)^\top (y - x_0) + \frac{\mu}{2}\|y - x_0\|^2$. Consequently, if $\|y - x_0\| > \frac{1}{\mu}\| \bigtriangledown f(x_0)\|_*$by Cauchy Schwarz

$$f(y) \geq f(x_0) - \| \bigtriangledown f(x_0)\|_* \cdot \|y - x_0\| + \mu\|y - x_0\|^2 > f(x_0)$$

and therefore if we let $B \stackrel{\text{def}}{=} \{y : \|y - x_0\| \leq \frac{1}{\mu}\| \bigtriangledown f(x_0)\|_*\}$ 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. Consequently by the Multivariable Extreme Value Theorem (Lemma 14) the result follows. □

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

Lemma 16. *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 $\bigtriangledown f(x_*)^\top (x - x_*) \geq 0$ for all $x \in S$.*

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

Consequently, it just remains to show that when $\bigtriangledown f(x_*)^\top (x - x_*) < 0$ 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

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

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 $f(x_0) \leq f(x)$ for all $x \in S$. Then from the above lemmas we see that there is some x_*such that $\bigtriangledown f(x_*)^\top (x - x_*) \geq 0$ for all $x \in S$. However, by convexity we know that $f(x_0) \geq f(x_*)^\top + \bigtriangledown f(x_*)^\top (x_0 - x_*)$ and therefore as $f(x_0) \leq f(x_*)$ we have $\bigtriangledown f(x_*)^\top x_0 \leq \bigtriangledown f(x_*)^\top x_*$. Consequently, we have that $H(\bigtriangledown f(x_*), \bigtriangledown f(x_*)^\top x_*)$ would be a separating hyperplane.

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 $f(x) \stackrel{\text{def}}{=} g(x) - \left[g(x_0) + \bigtriangledown g(x_0)^\top (x - x_0)\right]$ is a μ-strongly convex function that obtains its minimum at

x_0, since $\bigtriangledown f(x_0) = \bigtriangledown g(x_0) - \bigtriangledown g(x_0) = \vec{0}$. 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 17. 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

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

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

We remark that projection obeys a type of Pythagorean Theorem.

Lemma 18. *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

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

Where we used Lemma 16 to conclude the final line. □

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

Theorem 19. *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 $\bigtriangledown\left(\frac{1}{2}\|x - x_0\|_2^2\right) = x - x_0$ the first claim follows from Lemma 16. Since

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

the remainder of the theorem follows. □

To show that supporting hyperplanes exist we take a limit of these strict supporting hyperplanes.

We can now prove that supporting hyperplanes exist. We just need to formally define the boundary of a set.

Definition 20 (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.

Lemma 21. *For any convex set S and x_0 on the boundary of S there is a supporting hyperplane from x_0 to S.*

Proof. First replace S with its convex closure. Note that its closure is still convex and x_0 is still on the boundary thus if we can prove the claim for the closure, we obtain the desired result. Now let $x_1, x_2 ....$ be a infinite sequence of points such that $\lim_{k \to \infty} x_k$ and all the $x_i \notin S$ for all i. Let

$$g_i \stackrel{\text{def}}{=} \frac{\pi_S(x_i) - x_i}{\|\pi_S(x_i) - x_i\|_2} \text{ and } c_i \stackrel{\text{def}}{=} g_i^\top x_i$$

By the separating hyperplane theorem we have that for all i

$$(\pi_S(x_i) - x_i)^\top x \geq (\pi_S(x_i) - x_i)^\top x_0 + \|\pi_S(x_i) - x_i\|_2^2$$

and consequently dividing both sides by $\|\pi_S(x_i) - x_i\|_2$ we have that

$$g_i^\top x \geq g_i^\top x_i + \|\pi_S(x_i) - x_i\|_2$$

for all i. Now since $\|g_i\|_2 \leq 1$ for all $i \in [n]$ we have that there is a convergent subsequence where $\lim_{k \to \infty} g_k$ converges and therefore we assume this happens without loss of generality. Now let $g = \lim_{k \to \infty} g_k$. We have that

$$\begin{aligned} g^\top (x - x_0) &= \lim_{k \to \infty} g_k^\top (x - x_0) \geq \lim_{k \to \infty} g_k^\top (x_i - x_0) + \|\pi_S(x_i) - x_i\|_2 \\ &\geq \lim_{k \to \infty} -\|g_k\|_2 \cdot \|x_i - x_0\|_2 = 0. \end{aligned}$$

□

Note that this lemma gives another characterization of convex functions. It says that for any $x \in \partial S$ for convex S it is the case that there is a direction $d \neq 0$ such that $x + \alpha d \notin S$ for all $\alpha > 0$.