CO463: Convex Optimization and Analysis

Felix Zhou [1]

January 5, 2024

[1]From Professor Walaa Moursi's Lectures at the University of Waterloo in Winter 2021

Contents

Chapter 1

Convex Sets

1.1 Introduction

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Consider the problem

$$\begin{aligned} &\min f(x) \\ &x \in C \subseteq \mathbb{R}^n \end{aligned} \qquad (P)$$

In the case when $C = \mathbb{R}^n$, the minimizers of f will occur at the critical points of f. Namely, at $x \in \mathbb{R}^n$ when $\nabla f(x) = 0$. This is known as "Fermat's Rule".

In this course, we seek to approach (P) when f is not differentiable but f is convex and when $\varnothing \neq C \subsetneq \mathbb{R}^n$ is a convex set.

1.2 Affine Sets & Subspaces

Definition 1.2.1 (Affine Set)
$S \subseteq \mathbb{R}^n$ is affine if for all $x, y \in S$ and $\lambda \in \mathbb{R}$,

$$\lambda x + (1 - \lambda) y \in S.$$

Definition 1.2.2 (Affine Subspace)
An affine set $\varnothing \neq S \subseteq \mathbb{R}^n$.

Definition 1.2.3 (Affine Hull)
Let $S \subseteq \mathbb{R}^n$. The affine hull of S

$$\operatorname{aff}(S) := \bigcap_{S \subseteq T \subseteq \mathbb{R}^n : T \text{ is affine}} T$$

is the smallest affine set containing S.

Example 1.2.1
Let L be a linear subspace of $\mathbb{R}^n$ and $a \in \mathbb{R}^n$.

Then $L, a + L, \varnothing, \mathbb{R}^n$ are all examples of affine sets.

1.3 Convex Sets

Definition 1.3.1
$C \subseteq \mathbb{R}^n$ is convex if for all $x, y \in C$ and $\lambda \in (0, 1)$,

$$\lambda x = (1 - \lambda) y \in C.$$

Example 1.3.1
$\varnothing$, $\mathbb{R}^n$, balls, affine, and half-sets are all examples of convex sets.

Theorem 1.3.2
The intersection of an arbitrary collection of convex sets is convex.

Proof
Let I be an index set. Let $(C_i)_{i \in I}$ be a collection of convex subsets of $\mathbb{R}^n$.

Put

$$C := \bigcap_{i \in I} C_i.$$

Pick $x, y \in C$. By the definition of set intersection, $x, y \in C_i$ for all $i \in I$. Since each C_i is convex, for all $\lambda \in (0, 1)$,

$$\lambda x + (1 - \lambda) y \in C_i.$$

It follows that C is convex by the arbitrary choice of i.

Corollary 1.3.2.1
Let $b_i \in \mathbb{R}^n$ and $\beta_i \in \mathbb{R}$ for $i \in I$ for some arbitrary index set I.
The set
$$C := \{x \in \mathbb{R}^n : \langle x, b_i \rangle \leq \beta_i, \forall i \in I\}$$
is convex.

1.4 Convex Combinations of Vectors

Definition 1.4.1 (Convex Combinations)
A vector sum
$$\sum_{i=1}^{m} \lambda_i x_i$$
is a convex conbination if $\lambda \geq 0$ and $1^T \lambda = 1$.

Theorem 1.4.1
$C \subseteq \mathbb{R}^n$ is convex if and only if it contains all convex combinations of its elements.

Proof
($\Longleftarrow$) Apply the definition of convex combination with $m = 2$.

($\Longrightarrow$) We argue by induction on m. Observe that by deleting x_i's if necessary, we may assume without loss of generality that $\lambda > 0$.

When $m = 2$, this is simply the definition of convexity.

For $m > 2$, we can write
$$\begin{aligned}
\sum_{i=1}^{m} \lambda_i x_i &= \sum_{i=1}^{m-1} \lambda_i x_i + \lambda_m x_m \\
&= (1 - \lambda_m) \sum_{i=1}^{m-1} \frac{\lambda_i}{1 - \lambda_m} x_i + \lambda_m x_m \\
&= (1 - \lambda_m) x' + \lambda_m x_m. && x' \in C \text{ by induction}
\end{aligned}$$

Hence C indeed contains all convex combinations of its elements.

Definition 1.4.2 (Convex Hull)
The convex hull of $S \subseteq \mathbb{R}^n$

$$\operatorname{conv} S := \bigcap_{S \subseteq T \subseteq \mathbb{R}^n : T \text{ is convex}} T$$

is the smallest convex set containing S.

Theorem 1.4.2
Let $\subseteq \mathbb{R}^n$. $\operatorname{conv} S$ consists of all convex conbinations of elements of S.

Proof
Let D be the set of convex combinations of elements of S.

($\operatorname{conv} S \subseteq D$) D is convex since convex combinations of convex combinations again yields convex combinations. Moreover, $S \subseteq D$ by considering the trivial convex combination. It follows that $\operatorname{conv} S \subseteq D$ by definition.

($D \subseteq \operatorname{conv} S$) By the previous theorem, the convexity of $\operatorname{conv} S$ means that if contains all convex combinations of elements. In particular, it contains all convex conbinations of $S \subseteq \operatorname{conv} S$.

1.5 The Projection Theorem

Definition 1.5.1 (Distance Function)
Fix $S \subseteq \mathbb{R}^n$. The distance to S is the function $d_S : \mathbb{R}^n \to [0, \infty]$ given by

$$x \mapsto \inf_{s \in S} \|x - s\|.$$

Definition 1.5.2 (Projection onto a Set)
Let $\varnothing \neq C \subseteq \mathbb{R}^n, x \in \mathbb{R}^n$ and $p \in C$. p is a projection of x onto C, if

$$d_C(x) = \|x - p\|.$$

If a projection p of x onto C is unique, we denote it by $P_C(x) := p$.

Recall that a *cauchy sequence* $(x_n)_{n\in N}$ in $\mathbb{R}^n$ is a sequence such that

$$\|x_m - x_n\| \to 0$$

as $\min(m,n) \to \infty$.

Since $\mathbb{R}^n$ is a complete metric space under the Euclidean metric, every cauchy sequence converges in $\mathbb{R}^n$.

Moreover, recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous at $\bar{x} \in \mathbb{R}^n$ if and only if for every sequence $x_n \to \bar{x}$, we have

$$f(x_n) \to f(\bar{x}).$$

Fix $y \in \mathbb{R}^n$. The function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$x \mapsto \|x - y\|$$

is continuous.

Lemma 1.5.1
Let $x, y, z \in \mathbb{R}^n$. Then

$$\|x-y\|^2 = 2\|z-x\|^2 + 2\|z-y\|^2 - 4\left\|z - \frac{x+y}{2}\right\|^2.$$

Proof
This is by computation.

$$\begin{aligned}
2\|x-z\|^2 &= 2\langle z-x, z-x\rangle \\
&= 2\|z\|^2 - 4\langle z,x\rangle + 2\|x\|^2 \\
2\|z-y\|^2 &= 2\|z\|^2 - 4\langle z,y\rangle + 2\|y\|^2 \\
4\left\|z - \frac{x+y}{2}\right\|^2 &= 4\left[\|z\|^2 + \frac{1}{4}\|x+y\|^2 - \langle z, x+y\rangle\right] \\
&= 4\|z^2\| + \|x+y\|^2 - 4\langle z,x\rangle - 4\langle z,y\rangle.
\end{aligned}$$

Putting everything together yields

$$\begin{aligned} 2\|z-x\|^2 + 2\|z-y\|^2 - 4\left\|z - \frac{x+y}{2}\right\|^2 &= 2\|x\|^2 + 2\|y\|^2 - \|x+y\|^2 \\ &= \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle \\ &= \|x-y\|^2. \end{aligned}$$

Lemma 1.5.2
Let $x, y \in \mathbb{R}^n$. Then

$$\langle x, y\rangle \leq 0 \iff \forall \lambda \in [0,1], \|x\| \leq \|x - \lambda y\|.$$

Proof
($\implies$) Suppose $\langle x, y\rangle \leq 0$. Then

$$\begin{aligned} \|x-\lambda y\|^2 - \|x\|^2 &= \lambda\left(\lambda\|y\|^2 - 2\langle x, y\rangle\right) \\ &\geq 0. \end{aligned}$$

($\impliedby$) Conversely, we have $\lambda\|y\|^2 - 2\langle x, y\rangle \geq 0$. This implies

$$\begin{aligned} \langle x, y\rangle &\leq \frac{\lambda}{2}\|y\|^2 \\ &\to 0. & \lambda \to 0 \end{aligned}$$

Theorem 1.5.3 (Projection)
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ be closed and convex. Then the following hold:

i) For all $x \in \mathbb{R}^n$, $P_C(x)$ exists and is unique.

ii) For every $x \in \mathbb{R}^n$ and $p \in \mathbb{R}^n$, $p = P_C(x) \iff p \in C \wedge \forall y \in C, \langle y-p, x-p\rangle \leq 0$.

Proof (i)
Recall that

$$d_C(x) := \inf_{c \in C} \|x - c\|.$$

Hence there is a sequence $(c_n)_{n\in\mathbb{N}}$ in C such that

$$d_C(x) = \lim_{n\to\infty} \|c_n - x\|.$$

Let $m, n \in \mathbb{N}$. By the convexity of C, $\frac{1}{2}c_m + \frac{1}{2}c_n \in C$. But then

$$d_C(x) = \inf_{c \in C} \|x - c\| \leq \left\| x - \frac{1}{2}(c_m + c_n) \right\|.$$

Apply our first lemma with c_m, c_n, x to see that

$$\begin{aligned} \|c_n - c_m\|^2 &= 2\|c_n - x\|^2 + 2\|c_m - x\|^2 - 4\left\| x - \frac{c_n + c_m}{2} \right\|^2 \\ &\leq 2\|c_n - x\|^2 + 2\|c_m - x\|^2 - 4d_C(x)^2. \end{aligned}$$

As $m, n \to \infty$,

$$0 \leq \|c_n - c_m\|^2 \to 4d_C(x)^2 - 4d_C(x)^2 = 0$$

and (c_n) is a Cauchy sequence. But then there is some $c \in p$ such that $c_n \to p$ by the closedness (completeness) of C.

By the continuity of $\|x - \cdot\|$, $c_n \to p$ implies

$$\|x - c_n\| \to d_C(x) = \|x - p\|.$$

This demonstrates the existence of p.

Suppose there is some $q \in C$ such that $d_C(x) = \|q - x\|$. By convexity, $\frac{1}{2}(p + q) \in C$. Using the first lemma again, we have

$$\begin{aligned} 0 &\leq \|p - q\|^2 \\ &= 2\|p - x\|^2 + 2\|q - x\|^2 - 4\left\| x - \frac{p + q}{2} \right\|^2 \\ &\leq 2d_C(x)^2 + 2d_C(x)^2 - 4d_C(x)^2 \\ &\leq 0. \end{aligned}$$

So $\|p - q\| = 0 \implies p = q$.

This shows uniqueness.

Proof (ii)
Observe that $p = P_C(x)$ if and only if $p \in C$ and

$$\|x - p\|^2 = d_C(x)^2.$$

Since C is convex,

$$\forall \alpha \in [0,1], y_\alpha := \alpha y + (a - \alpha)p \in C.$$

Thus

$$\begin{aligned} &\|x-p\|^2 = d_C(x)^2 \\ &\iff \forall y \in C, \alpha \in [0,1], \|x-p\|^2 \leq \|x - y_\alpha\|^2 \\ &\iff \forall y \in C, \alpha \in [0,1], \|x-p\|^2 \leq \|x - p - \alpha(y-p)\|^2 \\ &\iff \forall y \in C, \langle x-p, y-p \rangle \leq 0 && \text{auxiliary lemma 2.} \end{aligned}$$

In the absence of closedness, $P_C(x)$ does not in general exist unless $x \in C$. In the absence of convexity, uniqueness does not in general hold.

Example 1.5.4
Fix $\epsilon > 0$ and $C = B(0;\epsilon)$ be the closed ball around 0 or radius ϵ.

For all $x \in \mathbb{R}^n$, either $P_C(x) = x$ when $x \in C$ or $P_C(x)$ is $\frac{\epsilon}{\|x\|}x$, the vector obtained from x by scaling its norm to ϵ.

In other words,

$$P_C(x) = \frac{\epsilon}{\max(\|x\|, \epsilon)} x.$$

1.6 Convex Set Operations

Definition 1.6.1 (Minkowski Sum)
Let $C, D \subseteq \mathbb{R}^n$. The Minkowski Sum of C, D is

$$C + D := \{c + d : c \in C, d \in D\}.$$

Theorem 1.6.1 (Minkowski)
Let $C_1, C_2 \subseteq \mathbb{R}^n$ be convex. Then $C_1 + C_2$ is convex.

Proof
If either C_1, C_2 is empty, then $C_1 + C_2 = \varnothing$ by definition.

Otherwise, $C_1 + C_2 \neq \varnothing$. Fix $x_1 + x_2, y_1 + y_2 \in C_1 + C_2$ and $\lambda \in (0,1)$. By the convexity of C_1, C_2,

$$\lambda(x_1 + x_2) + (1-\lambda)(y_1 + y_2) = \lambda x_1 + (1-\lambda)y_1 + \lambda x_2 + (1-\lambda)y_2$$
$$\in C_1 + C_2$$

as required.

Proposition 1.6.2
Let $\varnothing \neq C, D \subseteq \mathbb{R}^n$ be closed and convex. Moreover, suppose that D is bounded. Then $C + D \neq \varnothing$ is closed and convex.

Proof
We have already shown non-emptiness and convexity in the previous theorem.

Let $(x_n + y_n)_{n\in N}$ be a convergent sequence in $C + D$. Say that $x_n + y_n \to z$.

Since D is bounded, there is a subsequence $(y_{k_n})_{n\in N}$ such that $y_{k_n} \to y \in D$. It follows that

$$x_{k_n} = z - y_{k_n} \to z - y \in C$$

by the closedness of C.

It follows that $z \in C + y \subseteq C + D$ as desired.

If we drop the assumption that D is bounded, the result no longer holds in general. Indeed, consider $C = \{2, 3, 4, \dots\}$ and $D := \{-n + \frac{1}{n} : n = 2, 3, 4, \dots\}$. $\left(\frac{1}{n}\right)_{n\geq 2}$ is the sum but 0 is not!

Theorem 1.6.3
Let $C \subseteq \mathbb{R}^n$ be convex and $\lambda_1, \lambda_2 \geq 0$. Then

$$(\lambda_1 + \lambda_2)C = \lambda_1 C + \lambda_2 C.$$

Proof
<u>($\subseteq$)</u> This is always true, even if C is not convex.

<u>($\supseteq$)</u> Without loss of generality, we may assume that $\lambda_1 + \lambda_2 > 0$. By convexity, we have

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} C + \frac{\lambda_2}{\lambda_1 + \lambda_2} C \subseteq C.$$

In other words, $\lambda_1 C + \lambda_2 C \subseteq (\lambda_1 + \lambda_2)C$.

1.7 Topological Properties

We will write

$$B(x;\epsilon) := \{y \in \mathbb{R}^n : \|y - x\| \le \epsilon\}$$

to denote the closed ball of radius ϵ about x. In particular, we write

$$B := B(0;1)$$

to denote the closed unit ball.

Definition 1.7.1 (Interior)
The interior of $C \subseteq \mathbb{R}^n$ is

$$\operatorname{int} C := \{x : \exists \epsilon > 0, x + \epsilon B \subseteq C\}.$$

Definition 1.7.2 (Closure)
The closure of $C \subseteq \mathbb{R}^n$ is

$$\bar{C} := \bigcap_{\epsilon > 0} C + \epsilon B.$$

Definition 1.7.3 (Relative Interior)
The relative interior of a convex $C \subseteq \mathbb{R}^n$ is

$$\operatorname{ri} C := \{x \in \operatorname{aff} C : \exists \epsilon > 0, (x + \epsilon B) \cap \operatorname{aff} C \subseteq C\}.$$

Proposition 1.7.1
Let $C \subseteq \mathbb{R}^n$. Suppose that $\operatorname{int} C \neq \varnothing$. Then $\operatorname{int} C = \operatorname{ri} C$.

Proof
Let $x \in \operatorname{int} C$. There is some $\epsilon > 0$ such that $B(x;\epsilon) \subseteq C$. Hence

$$\begin{aligned} \mathbb{R}^n &= \operatorname{aff}(B(x;\epsilon)) \\ &\subseteq \operatorname{aff} C \\ &\subseteq \mathbb{R}^n. \end{aligned}$$

But then $\operatorname{aff} C = \mathbb{R}^n$ and the result follows from definition.

Let $A \subseteq \mathbb{R}^n$ be affine. Every affine set has a corresponding linear subspace

$$L := A - A.$$

This is a linear subspace as it is affine and contains 0.

Definition 1.7.4 (Dimension)
Let $\varnothing \neq A \subseteq \mathbb{R}^n$ be affine. The dimension of A is the dimension of the corresponding linear subspace

$$\dim A := \dim(A - A).$$

It may be useful to consider

$$A - A = \bigcup_{a \in A} (A - a)$$

as the union of translations.

Definition 1.7.5 (Dimension)
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ be convex. The dimension of C, denoted $\dim C$, is the dimension of $\operatorname{aff} C$.

Proposition 1.7.2
Let $C \subseteq \mathbb{R}^n$ be convex. For all $x \in \operatorname{int} C$ and $y \in \bar{C}$,

$$[x, y) \subseteq \operatorname{int} C.$$

Proof
Let $\lambda \in [0, 1)$. We argue that $(1-\lambda)x + \lambda y \in \operatorname{int} C$. It suffices to show that

$$(1-\lambda)x + \lambda y + \epsilon B \subseteq C$$

for some $\epsilon > 0$.

As $y \in \bar{C}$, we have that $\forall \epsilon > 0, y \in C + \epsilon B$. Thus for all $\epsilon > 0$,

$$\begin{aligned}
(1-\lambda)x + \lambda y + \epsilon B &\subseteq (1-\lambda)x + \lambda(C + \epsilon B) + \epsilon B \\
&= (1-\lambda)x + (1+\lambda)\epsilon B + \lambda C && \text{previous theorem} \\
&= (1-\lambda)\left[x + \frac{1+\lambda}{1-\lambda}\epsilon B\right] + \lambda C \\
&\subseteq (1-\lambda)C + \lambda C && \text{sufficiently small } \epsilon, x \in \operatorname{int} C \\
&= C. && \text{previous theorem again}
\end{aligned}$$

Theorem 1.7.3
Let $C \subseteq \mathbb{R}^n$ be convex. Then for all $x \in \operatorname{ri} C$ and $y \in \bar{C}$,

$$[x, y) \subseteq \operatorname{ri} C.$$

Proof
Case I: $\operatorname{int} C \neq \varnothing$ This follows by the observation that $\operatorname{ri} C = \operatorname{int} C$.

Case II: $\operatorname{int} C = \varnothing$ We must have $\dim C = m < n$. Let $L := \operatorname{aff} C - \operatorname{aff} C$ be the corresponding linear subspace of dimension m.

Through translation by $-c \in C$ if necessary, we may assume without loss of generality that $C \subseteq L \cong \mathbb{R}^m$.

But then the interior of C with respect to $\mathbb{R}^m$ is $\operatorname{ri} C$ in $\mathbb{R}^n$. An application of Case I with $C \subseteq \mathbb{R}^m$ yields the result.

Theorem 1.7.4
Let $C \subseteq \mathbb{R}^n$ be convex. The following hold:
(i) $\bar{C}$ is convex.
(ii) $\operatorname{int} C$ is convex.
(iii) If $\operatorname{int} C \neq \varnothing$, then $\operatorname{int} C = \operatorname{int} \bar{C}$ and $\bar{C} = \overline{\operatorname{int} C}$.

Proof (i)
Let $x, y \in \bar{C}$ and $\lambda \in (0, 1)$. There are sequences $x_n, y_n \in C$ such that

$$x_n \to x, y_n \to y.$$

It follows by convexity that

$$C \ni \lambda x_n + (1 - \lambda) y \to \lambda x + (1 - \lambda y) \in \bar{C}.$$

By definition, $\bar{C}$ is convex.

Proof (ii)
If $\operatorname{int} C = \varnothing$, the conclusion is clear.

Otherwise, use the previous proposition with $y \in C \subseteq \bar{C}$ to see that

$$\begin{aligned}[x, y] &= [x, y) \cup \{y\} \\ &\subseteq \operatorname{int} C \cup \operatorname{int} C \\ &= \operatorname{int} C.\end{aligned}$$

Proof (iii)
Since $C \subseteq \bar{C}$, it must hold that $\operatorname{int} C \subseteq \operatorname{int} \bar{C}$.

Conversely, let $y \in \operatorname{int} \bar{C}$. If $y \in \operatorname{int} C$, then we are done. Thus suppose otherwise.

There is some $\epsilon > 0$ such that $B(y; \epsilon) \subseteq \bar{C}$. We may thus choose some $\operatorname{int} C \not\ni y \neq x \in \operatorname{int} C \neq \varnothing$ and $\lambda > 0$ sufficiently small such that

$$y + \lambda(y - x) \in B(y; \epsilon) \subseteq \bar{C}.$$

By a previous proposition applied with $y + \lambda(y - x)$, we have that

$$[x, y + \lambda(y - x)) \subseteq \operatorname{int} C.$$

We now claim that $y \in [x, y + \lambda(y - x))$. Indeed, set $\alpha := \frac{1}{1+\lambda} \in (0, 1)$. We have

$$\begin{aligned}(1 - \alpha)x + \alpha(y + \lambda(y - x)) &= (1 - \alpha(1 + \lambda))x + \alpha(1 + \lambda)y \\ &= y.\end{aligned}$$

It follows by the arbitrary choice of y that $\operatorname{int} \bar{C} \subseteq \operatorname{int} C$. We now turn to the second identity.

Since $\operatorname{int} C \subseteq C$, we must have $\overline{\operatorname{int} C} \subseteq \bar{C}$. Conversely, let $y \in \bar{C}$ and $x \in \operatorname{int} C$. For $\lambda \in [0, 1)$, define

$$y_\lambda = (1 - \lambda)x + \lambda y.$$

The previous proposition agains tells us that

$$y_\lambda \in [x, y) \subseteq \operatorname{int} C.$$

But then $y = \lim_{\lambda \to 0} y_\lambda \in \overline{\operatorname{int} C}$ and $\bar{C} \subseteq \overline{\operatorname{int} C}$.

This concludes the argument.

Theorem 1.7.5
Let $C \subseteq \mathbb{R}^n$ be convex. Then $\operatorname{ri} C, \bar{C}$ are convex.
Moreover,
$$C \neq \varnothing \iff \operatorname{ri} C \neq \varnothing.$$

1.8 Separation Theorems

Definition 1.8.1 (Separated)
Let $C_1, C_2 \subseteq \mathbb{R}^n$. We say C_1, C_2 are separated if there is some $b \in \mathbb{R}^n \setminus \{0\}$ such that

$$\sup_{c_1 \in C_1} \langle c_1, b \rangle \boxed{\leq} \inf_{c_2 \in C_2} \langle c_2, b \rangle.$$

If
$$\sup_{c_1 \in C_1} \langle c_1, b \rangle \boxed{<} \inf_{c_2 \in C_2} \langle c_2, b \rangle,$$
then we say C_1, C_2 are *strongly separated*.

Theorem 1.8.1
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ be closed and convex and suppose $x \notin C$. Then x is strongly separated from C.

Proof
The goal is to find some $b \neq 0$ such that
$$\begin{aligned} \sup \langle c, b \rangle &< \langle x, b \rangle \\ \sup \langle c - x, b \rangle &< 0. \end{aligned}$$

Set $p := P_C(X)$ and $b := x - p \neq \varnothing$. Let $y \in C$. By the projection theorem,
$$\begin{aligned} \langle y - p, x - p \rangle &\leq 0 && \forall y \in C \\ \langle y - (x - b), x - (x - b) \rangle &\leq 0 && p = x - b \\ \langle y - x, b \rangle &\leq -\langle b, b \rangle \\ &= -\|b\|^2 \\ \sup_{y \in C} \langle y, b \rangle - \langle x, b \rangle &\leq -\|b\|^2 \\ &< 0 \end{aligned}$$
as desired.

Corollary 1.8.1.1
Let $C_1 \cap C_2 = \varnothing$ be nonempty subsets of $\mathbb{R}^n$ such that $C_1 - C_2$ is closed and convex. Then C_1, C_2 are strongly separated.

Proof
By definition, C_1, C_2 are strongly separated if and only if there is $b \neq 0$ such that

$$\begin{aligned}
\sup_{c_1 \in C_1} \langle c_1, b \rangle &< \inf_{c_2 \in C_2} \langle c_2, b \rangle \\
\sup_{c_1 \in C_1} \langle c_1, b \rangle &< - \sup_{c_2 \in C_2} \langle c_2, b \rangle \\
\sup_{c_1 \in C_1} \langle c_1, b \rangle + \sup_{c_2 \in C_2} \langle c_2, b \rangle &< 0 \\
\sup_{c_1 \in C_1, c_2 \in C_2} \langle c_1 - c_2, b \rangle &< 0.
\end{aligned}$$

Since $C_1 \cap C_2 = \varnothing$, we know that $0 \notin C_1 - C_2$. Hence $C_1 - C_2$ is strongly separated from 0 and the conclusion follows.

Corollary 1.8.1.2
Let $\varnothing \neq C_1, C_2 \subseteq \mathbb{R}^n$ be closed and convex such that $C_1 \cap C_2 = \varnothing$ and C_2 is bounded. Then C_1, C_2 are strongly separted.

Proof
$C_1 \cap C_2 = \varnothing \implies 0 \notin C_1 - C_2$. In addition, $-C_2$ is also closed and convex. It follows by a previous theorem that $C_1 + (-C_2)$ is nonempty, closed, and convex.

Theorem 1.8.2
Let $\varnothing \neq C_1, C_2 \subseteq \mathbb{R}^n$ be closed and convex such that $C_1 \cap C_2 = \varnothing$. Then C_1, C_2 are separated.

Proof
For each $n \in \mathbb{N}$, set

$$D_n := C_2 \cap B(0; n).$$

Observe that $C_1 \cap D_n = \varnothing$ for all n. Moreover, D_n is bounded by construction.

It follows that there is a hyperplane u_n that separates C_1, D_n for all n. Specifically, $\|u_n\| = 1$ and

$$\sup \langle C_1, u_n \rangle < \inf \langle D_n, u_n \rangle.$$

But the sequence u_n is bounded, hence there is a convergent subsequence u_{k_n}. where $u_{k_n} \to u$ with $\|u\| = 1$.

Let $x \in C_1, y \in C_2$. For sufficiently large n, $y \in B(0; k_n)$ and

$$\langle x, u_{k_n} \rangle < \langle y, u_{k_n} \rangle.$$

Taking the limit as $k \to \infty$ yields

$$\langle x, u \rangle \leq \langle y, u \rangle.$$

This completes the proof.

1.9 More Convex Sets

Definition 1.9.1 (Cone)
$C \subseteq \mathbb{R}^n$ is a cone if

$$C = \mathbb{R}_{++}C.$$

Definition 1.9.2 (Conical Hull)
$\operatorname{cone} C$ is the intersection of all cones containing C.

Definition 1.9.3 (Closed Conical Hull)
$\overline{\operatorname{cone}}(C)$ is the smallest closed cone containing C.

Proposition 1.9.1
Let $C \subseteq \mathbb{R}^n$. The following hold:

(i) $\operatorname{cone} C = \mathbb{R}_{++}C$

(ii) $\overline{\operatorname{cone} C} = \overline{\operatorname{cone}}(C)$

(iii) $\operatorname{cone}(\operatorname{conv} C) = \operatorname{conv}(\operatorname{cone} C)$

(iv) $\overline{\operatorname{cone}}(\operatorname{conv} C) = \overline{\operatorname{conv}}(\operatorname{cone} C)$

The proofs of all these are trivial if $C = \varnothing$. Thus in our proofs, we assume that C is nonempty.

Proof (i)
Set $D := \mathbb{R}_{++}C$. It is clear that $C \subseteq D$ with D being a cone. Hence $\operatorname{cone} C \subseteq D$.

Conversely, for $y \in D$, there is some $\lambda > 0, c \in C$ for which $y = \lambda c$. Then $y \in \operatorname{cone} C$ and $D \subseteq \operatorname{cone} C$.

Proof (ii)
$\overline{\text{cone}}(C)$ is a closed cone with $C \subseteq \overline{\text{cone}}(C)$. Hence

$$\overline{\operatorname{cone} C} \subseteq \overline{\overline{\text{cone}}(C)} = \overline{\text{cone}}(C).$$

Conversely, since $\operatorname{cone} C$ is a cone,

$$\overline{\text{cone}}(C) \subseteq \overline{\operatorname{cone} C}.$$

Proof (iii)
$\underline{(\subseteq)}$ Let $x \in \operatorname{cone}(\operatorname{conv} C)$. By i, there is $\lambda > 0, y \in \operatorname{conv} C$ such that $x = \lambda y$. Since $y \in \operatorname{conv} C$, we can express is as a convex combination

$$\begin{aligned} x &= \lambda y \\ &= \lambda \sum_{i=1}^m \lambda_i x_i \\ &= \sum_{i=1}^m \lambda_i \lambda x_i \\ &\in \operatorname{conv}(\operatorname{cone} C). \end{aligned}$$

$\underline{(\supseteq)}$ Let $x \in \operatorname{conv}(\operatorname{cone} C)$. We can write x as convex combinations of scalar multiples of C.

$$\begin{aligned} x &= \sum_{i=1}^m \mu_i \lambda_i x_i \\ &= \left(\sum_{i=1}^m \lambda_i \mu_i \right) \left(\sum_{i=1}^m \frac{\lambda_i \mu_i}{\sum \lambda_i \mu_i} x_i \right) \\ &= \alpha \sum_{i=1}^m \beta_i x_i. \end{aligned}$$

This is a scalar multiple of a convex combination of C and thus $x \in \operatorname{cone}(\operatorname{conv} C)$ as desired.

Proof (iv)
This is a direct consequence of iii.

Lemma 1.9.2
Let $0 \in C \subseteq \mathbb{R}^n$ be convex with $\operatorname{int} C \neq \varnothing$. The following are equivalent:

(i) $0 \in \operatorname{int} C$

(ii) $\operatorname{cone} C = \mathbb{R}^n$

(iii) $\overline{\operatorname{cone}} C = \mathbb{R}^n$

It is a fact that for $0 \in C \subseteq \mathbb{R}^n$ convex with $\operatorname{int} C \neq \varnothing$,

$$\operatorname{int}(\operatorname{cone} C) = \operatorname{cone}(\operatorname{int} C).$$

Proof
(i) $\implies$ (ii) Suppose $0 \in \operatorname{int} C$. Then $B(0;\epsilon) \subseteq C$ for some $\epsilon > 0$. But then

$$\begin{aligned} \mathbb{R}^n &= \operatorname{cone}(B(0;\epsilon)) \\ &\subseteq \operatorname{cone} C \\ &\subseteq \mathbb{R}^n \end{aligned}$$

and we have equality.

(ii) $\implies$ (iii) Recall that $\overline{\operatorname{cone} C} = \overline{\operatorname{cone}} C$. But then

$$\mathbb{R}^n = \operatorname{cone} C \subseteq \overline{\operatorname{cone}} C.$$

(iii) $\implies$ (i) Recall that $\operatorname{cone}(\operatorname{conv} C) = \operatorname{conv}(\operatorname{cone} C)$. Thus

$$\operatorname{conv}(\operatorname{cone} C) = \operatorname{cone} C$$

and $\operatorname{cone} C$ is convex.

By assumption,

$$\varnothing \neq \operatorname{int} C \subseteq \operatorname{int}(\operatorname{cone} C)$$

and $\operatorname{cone} C$ has nonempty interior.

Recall that

$$\operatorname{int}(\operatorname{cone} C) = \operatorname{int}(\overline{\operatorname{cone}} C)$$

as $\operatorname{cone} C$ is convex.

Hence

$$\begin{aligned}\mathbb{R}^n &= \operatorname{int}\mathbb{R}^n \\ &= \operatorname{int}(\overline{\operatorname{cone}}C) \\ &= \operatorname{int}(\operatorname{cone}C) \\ &= \operatorname{cone}(\operatorname{int}C).\end{aligned}$$

Thus $0 \in \lambda \operatorname{int} C$ for some $\lambda > 0$. It must be then that $0 \in C$ as desired.

Definition 1.9.4 (Tangent Cone)
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ with $x \in \mathbb{R}^n$. The tangent cone to C at x is

$$T_C(x) = \begin{cases} \overline{\operatorname{cone}}(C - x) = \overline{\bigcup_{\lambda \in \mathbb{R}_{++}} \lambda(C - x)}, & x \in C \\ \varnothing, & x \notin C \end{cases}$$

Definition 1.9.5 (Normal Cone)
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ with $x \in \mathbb{R}^n$. The normal cone to C at x is

$$N_C(x) = \begin{cases} \{u \in \mathbb{R}^n : \sup_{c \in C} \langle c - x, u \rangle \leq 0\}, & x \in C \\ \varnothing, & x \notin C \end{cases}$$

Theorem 1.9.3
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ be closed and convex. Let $X \in \mathbb{R}^n$.
Both $N_C(x), T_C(x)$ are closed convex cones.

Lemma 1.9.4
Let $\varnothing \neq C \subseteq \mathbb{R}^n$ be closed and convex with $x \in C$.

$$n \in N_C(x) \iff \forall t \in T_C(x), \langle n, t \rangle \leq 0.$$

Proof
$(\implies)$ Let $n \in N_C(x)$ and $t \in T_C(x)$. Recall that $T_C(x) = \overline{\operatorname{cone}}(C - x)$. Thus there is some $\lambda_k > 0$ and $t_k \in \mathbb{R}^n$ such that

$$x + \lambda_k t_k \in C$$

and $t_k \to t$.

Since $n \in N_C(x)$ and $x + \lambda_k t_k \in C$, it follows that for all k, $\langle n, \lambda_k t_k \rangle \leq 0$. But then as

$k \to \infty$ we see that

$$\langle n, t\rangle \leq 0.$$

($\impliedby$) Suppose that $\forall t \in T_C(x)$, we have $\langle n, t\rangle \leq 0$. Pick $y \in C$ and observe that

$$\begin{aligned} y - x &\in C - x \\ &\subseteq \operatorname{cone}(C - x) \\ &\subseteq \overline{\operatorname{cone}}(C - x) \\ &=: T_C(x). \end{aligned}$$

It follows that $\langle n, y - x\rangle \leq 0$ and $n \in N_C(x)$.

Theorem 1.9.5
Let $C \subseteq \mathbb{R}^n$ be convex such that $\operatorname{int} C \neq \varnothing$. Let $x \in C$. The following are equivalent.

(1) $x \in \operatorname{int} C$

(2) $T_C(x) = \mathbb{R}^n$

(3) $N_C(x) = \{0\}$

Proof
(1) $\iff$ (2) Observe that $x \in \operatorname{int} C$ if and only if $0 \in \operatorname{int}(C - x)$ if and only if there is some $\epsilon > 0$ with

$$B(0; \epsilon) \subseteq C - x.$$

Now,

$$\begin{aligned} \mathbb{R}^n &= \operatorname{cone}(B(0; \epsilon)) \\ &\subseteq \operatorname{cone}(C - x) \\ &\subseteq \overline{\operatorname{cone}(C - x)} \\ &= \overline{\operatorname{cone}}(C - x) \\ &= T_C(x) \\ &\subseteq \mathbb{R}^n. \end{aligned}$$

(2) $\iff$ (3) Our previous lemma combined with (1) yields

$$\begin{aligned} n \in N_C(x) &\iff \forall t \in T_C(x) = \mathbb{R}^n, \langle n, t\rangle \leq 0 \\ &\iff n = 0. \end{aligned}$$

Hence $N_C(x) = \{0\}$.

Conversely, suppose $N_C(x) = \{0\}$. It is clear that $0 \in T_C(x)$. Pick $y \in \mathbb{R}^n$. We claim that $y \in T_C(x)$. To see this recall that $T_C(x)$ is a closed convex cone, hence $p = P_{T_C(x)}(y)$ exists and is unique. Moreover, it suffices to show that $y = p \in T_C(x)$.

Indeed, by the projection theorem

$$\langle y - p, t - p \rangle \leq 0$$

for all $t \in T_C(x)$. In particular, it holds for $t = p, 2p \in T_C(x)$ ($T_C(x)$ is a cone). So

$$\langle y - p, \pm p \rangle \leq 0 \implies \langle y - p, p \rangle = 0.$$

But then $\langle y - p, t \rangle \leq 0$ for all $t \in T_C(x)$, which implies that $y - p \in N_C(x) = \{0\}$ and

$$y = p \in T_C(x)$$

as desired.

Chapter 2

Convex Functions

2.1 Definitions & Basic Results

Definition 2.1.1 (Epigraph)
Let $f : \mathbb{R}^n \to [-\infty, \infty]$. The epigraph of f is

$$\operatorname{epi} f := \{(x, \alpha) : f(x) \leq \alpha\} \subseteq \mathbb{R}^n \times \mathbb{R}.$$

Definition 2.1.2 (Domain)
For $f : \mathbb{R}^n \to [-\infty, \infty]$,

$$\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < \infty\}.$$

Definition 2.1.3 (Proper Function)
We say that f is *proper* if $\operatorname{dom} f \neq \varnothing$ and $f(\mathbb{R}^n) > -\infty$.

Definition 2.1.4 (Indicator Function)
Let $C \subseteq \mathbb{R}^n$. The indicator function of C is given by

$$\delta_C(x) := \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases}$$

Definition 2.1.5 (Lower Semicontinuous)
f is lower semicontinuous (l.s.c.) if $\text{epi}(f)$ is closed.

Definition 2.1.6 (Convex Function)
f is convex if $\text{epi} f$ is convex.

Proposition 2.1.1
Let $f : \mathbb{R}^n \to [-\infty, \infty]$ be convex. Then $\text{dom} f$ is convex.

Recall that linear transformations $A : \mathbb{R}^n \to \mathbb{R}^m$ preserve set convexity ($C \subseteq \mathbb{R}^n$ convex implies that $A(C)$ is convex).

Proof
Consider the linear transformation $L : \mathbb{R}^{n+1} \to \mathbb{R}^n$ given by

$$(x, \alpha) \mapsto x.$$

Then $\text{dom} f = L(\text{epi} f)$ is convex.

Theorem 2.1.2
Let $f : \mathbb{R}^m \to [-\infty, \infty]$. Then f is convex if and only if for all $x, y \in \text{dom} f$ and $\lambda \in (0, 1)$,

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

Proof
If $f = \infty \iff \text{epi} f = \varnothing \iff \text{dom} f = \varnothing$, then result is trivial. Hence let us suppose that $f \neq \infty \iff \text{dom} f \neq \varnothing$.

<u>$(\implies)$</u> Pick $x, y \in \text{dom} f$ and $\lambda \in (0, 1)$. Observe that $(x, f(x)), (y, f(y)) \in \text{epi} f$. By convexity,

$$\begin{aligned} \lambda(x, f(x)) + (1 - \lambda)(y, f(y)) = (\lambda x + (1 - \lambda)y, \lambda f(x) - (1 - \lambda)f(y)) \qquad &\in \text{epi}(f) \\ f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \end{aligned}$$

<u>$(\impliedby)$</u> Conversely, suppose the function inequality holds. Pick $(x, \alpha), (y, \beta) \in \text{epi} f$ as well as $\lambda \in (0, 1)$. Now,

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &\leq \lambda f(x) + (1 - \lambda)f(y) \\ &\leq \lambda\alpha + (1 - \lambda)\beta \end{aligned}$$

and

$$(\lambda x + (1-\lambda)y, \lambda\alpha, (1-\lambda)\beta) \in \operatorname{epi} f$$

as desired.

It follows that $\operatorname{epi} f$ is convex and so is f.

2.2 Lower Semicontinuity

Definition 2.2.1 (Lower Semicontinuity; Alternative)
Let $f : \mathbb{R}^n \to [-\infty, \infty]$ and $x \in \mathbb{R}^n$. f is lower semicontinuous (l.s.c) at x if for every sequence $(x_n)_{n\geq 1} \in \mathbb{R}^n$ such that $x_n \to x$,

$$f(x) \leq \liminf f(x_n).$$

We say f is l.s.c. if f is l.s.c. at every point in $\mathbb{R}^n$.

Remark that continuity implies lower semicontinuity. One can show that the two definitions of l.s.c. are equivalent, but we omit the proof.

Theorem 2.2.1
Let $C \subseteq \mathbb{R}^m$. Then the following hold:

(i) $C \neq \varnothing$ if and only if δ_C is proper

(ii) C is convex if and only if δ_C is convex

(iii) C is closed if and only if δ_C is l.s.c.

We prove (i) and (ii) in A2.

Proof ((iii))
Observe that $C = \varnothing \iff \operatorname{epi} \delta_C = \varnothing$, which is certainly closed. Thus we proceed assuming $C \neq \varnothing$.

$(\implies)$ Suppose C is closed. We want to show that $\operatorname{epi} \delta_C$ is closed.

Pick a converging sequence sequence $(x_n, \alpha_n) \to (x, \alpha)$ with every element in $\operatorname{epi} \delta_C$. Observe that x_n is a sequence in C, hence $x \in C$. Moreover, $\alpha_n \in [0, \infty)$ and $\alpha \geq 0$.

It follows that $(x, \alpha) \in \operatorname{epi} \delta_C$ as required.

$(\impliedby)$ Conversely, suppose that δ_C is l.s.c. Let $(x_n)_{n\geq 1}$ be a sequence in C with $x_n \to x$.

By the definition of δ_C, it suffices to show that $\delta_C(x) = 0$.

By lower semicontinuity,

$$\begin{aligned} 0 &\le \delta_C(x) \\ &\le \liminf \delta_C(x_n) \\ &= 0 \end{aligned}$$

and we have equality throughout.

Proposition 2.2.2
Let I be an indexing set and let $(f_i)_{i\in I}$ be a family of l.s.c. convex functions on $\mathbb{R}^n$. Then

$$F := \sup_{i\in I} f_i$$

is convex and l.s.c.

Proof
We claim that $\operatorname{epi} F = \bigcap_{i\in I} \operatorname{epi} f$. Indeed,

$$\begin{aligned} (x,\alpha) \in \operatorname{epi} F &\iff \sup_{i\in I} f_i(x) \le \alpha \\ &\iff \forall i \in I, f_i(x) \le \alpha \\ &\iff \forall i \in I, (x,\alpha) \in \operatorname{epi} f_i \\ &\iff \forall i \in I (x,\alpha) \in \operatorname{epi} f_i. \end{aligned}$$

The result follows by the definition of convex functions and lower semicontinuity as intersections preserve both set convexity and closedness.

2.3 The Support Function

Definition 2.3.1 (Support Function)
Let $C \subseteq \mathbb{R}^m$. The support function $\sigma_C : \mathbb{R}^m \to [-\infty, \infty]$ of C is

$$u \mapsto \sup_{c\in C} \langle c, u\rangle.$$

Proposition 2.3.1
Let $\varnothing \neq C \subseteq \mathbb{R}^n$. Then σ_C is convex, l.s.c., and proper.

Proof
For each $c \in C$, define

$$f_C(x) := \langle x, c \rangle.$$

Then f_c is linear and hence proper, l.s.c., and convex. Moreover,

$$\sigma_C = \sup_{c \in C} f_c.$$

Combined with our previous proposition, we learn that σ_C is convex and l.s.c.

Observe that since $C \neq \varnothing$,

$$\sigma_C(0) = \sup_{c \in C} \langle 0, c \rangle = 0 < \infty.$$

Hence $\operatorname{dom} \sigma_C \neq \varnothing$. In addition, fix $\bar{c} \in C$. Then for all $u \in \mathbb{R}^m$,

$$\begin{aligned} \sigma_C(u) &= \sup_{c \in C} \langle u, c \rangle \\ &\geq \langle u, \bar{c} \rangle \\ &> -\infty. \end{aligned}$$

Hence σ_C is proper as well.

2.4 Further Notions of Convexity

Let $f : \mathbb{R}^m \to [-\infty, \infty]$ be proper. Then f is *strictly convex* if for every $x \neq y \in \operatorname{dom} f$ and $\lambda \in (0, 1)$,

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda) f(y).$$

Moreover, f is *strongly convex* with constant $\beta > 0$ if for every $x, y \in \operatorname{dom} f, \lambda \in (0, 1)$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) - \frac{\beta}{2} \lambda (1 - \lambda) \|x - y\|^2.$$

Clearly, strong convexity implies strict convexity, which in turn implies convexity.

2.5 Operations Preserving Convexity

Proposition 2.5.1
Let I be a finite indexing set and $(f_i)_{i\in I}$ a family of convex functions $\mathbb{R}^m \to [-\infty, \infty]$. Then

$$\sum_{i\in I} f_i$$

is convex.

Proposition 2.5.2
Let f be convex and l.s.c. and pick $\lambda > 0$. Then

$$\lambda f$$

is convex and l.s.c.

2.6 Minimizers

Definition 2.6.1 (Global Minimizer)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and $x \in \mathbb{R}^m$. Then x is a (global) minimizer of f if

$$f(x) = \min f(\mathbb{R}^m).$$

We will use $\operatorname{argmin} f$ to denote the set of minimizers of f.

Definition 2.6.2 (Local Minimum)
Let $f : \mathbb{R}^m \to]-\infty, \infty]$ be be proper and $\bar{x} \in \mathbb{R}^m$. Then $\bar{x}$ is a local minimum of f if there is $\delta > 0$ such that

$$\|x - \bar{x}\| < \delta \implies f(\bar{x}) \leq f(x).$$

We way that $\bar{x}$ is a *global minimum* of f if for all $x \in \operatorname{dom} f$,

$$f(\bar{x}) \leq f(x).$$

Analogously, we define the *local maximum* and *global maximum*.

Why are convex functions so special?

Proposition 2.6.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and convex. Then every local minimizer of f is a global minimizer.

Proof
Let x be a local minimizer of f. There is some $\rho > 0$ such that

$$f(x) = \min f(B(x; \rho)).$$

Pick some $y \in \operatorname{dom} f \setminus B(x; \rho)$. Notice that

$$\lambda := 1 - \frac{\rho}{\|x - y\|} \in (0, 1).$$

Set

$$z := \lambda x + (1 - \lambda) y \in \operatorname{dom} f.$$

We know this is in the domain as $\operatorname{dom} f$ is convex by our prior work.

We have

$$\begin{aligned} z - x &= (1 - \lambda) y - (1 - \lambda) x \\ &= (1 - \lambda)(y - x) \\ \|z - x\| &= \|(1 - \lambda)(y - x)\| \\ &= \frac{\rho}{\|y - x\|} \|y - x\| \\ &= \rho. \end{aligned}$$

This shows that $z \in B(x; \rho)$.

By the convexity of f,

$$\begin{aligned} f(x) &\le f(z) \\ &\le \lambda f(x) + (1 - \lambda) f(y) \\ (1 - \lambda) f(x) &\le (1 - \lambda) f(y) \\ f(x) &\le f(y). \end{aligned}$$

Proposition 2.6.2
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and convex. Let $C \subseteq \mathbb{R}^m$. Suppose that x is a minimizer of f over C such that $x \in \operatorname{int} C$. Then x is a minimizer of f.

Proof
There is some $\epsilon > 0$ such that x minimizes f over $B(x;\epsilon) \subseteq \operatorname{int} C$. Since x is a local minimizer, it is a global minimizer as well.

2.7 Conjugates

Definition 2.7.1 (Fenchel-Legendre/Convex Conjugate)
Let $f : \mathbb{R}^m \to [-\infty, \infty]$. Then Fenchel-Legendre/Convex Conjugate of f, denoted $f^* : \mathbb{R}^m \to [-\infty, \infty]$ is given by

$$u \mapsto \sup_{x \in \mathbb{R}^m} \langle x, u\rangle - f(x).$$

Recall that a closed convex set is the intersection of all supporting hyperplanes. The idea is that the epigraph of a convex, l.s.c. function f can be recovered by the supremum of affine functions majorized by f.

Given a slope $x \in \mathbb{R}^m$, we want the best translation α which supports f.

$$\begin{aligned} f(x) &\geq \langle u, x\rangle - \alpha && \forall x \in \mathbb{R}^n \\ \alpha &\geq \langle u, x\rangle - f(x) && \forall x \in \mathbb{R}^n. \end{aligned}$$

Thus $f^*(u) := \sup_{x\in\mathbb{R}^n} \langle u, x\rangle - f(x)$ is the best translation such that $\langle u, x\rangle - f^*(u)$ is majorized by f.

Proposition 2.7.1
Let $f : \mathbb{R}^m \to [-\infty, \infty]$. Then f^* is convex and l.s.c.

Proof
Observe that $f \equiv \infty \iff \operatorname{dom} f = \varnothing$. Hence if $f \equiv \infty$, for all $u \in \mathbb{R}^m$

$$\begin{aligned} f^*(u) &= \sup_{x\in\mathbb{R}^m} \langle x, u\rangle - f(x) \\ &= \sup_{x \in \operatorname{dom} f} \langle x, u\rangle - f(x) \\ &= -\infty. \end{aligned}$$

This is trivially convex and l.s.c.

Now suppose that $f \not\equiv \infty$. We claim that $f^*(u) = \sup_{(x,\alpha)\in \operatorname{epi} f} \langle x, u\rangle - \alpha$. Observe that

$f_{(x,\alpha)} := \langle x, \cdot \rangle - \alpha$ is an affine function. By definition,

$$\sup_{x \in \operatorname{dom} f} \langle x, u \rangle - f(x) \geq \sup_{(x,\alpha) \in \operatorname{epi} f} \langle x, u \rangle - \alpha$$

as $f(x) \leq \alpha$ by the definition of the epigraph. On the other hand,

$$\sup_{(x,f(x)) : x \in \operatorname{dom} f} \langle x, u \rangle - f(x) \leq \sup_{(x,\alpha) \in \operatorname{epi} f} \langle x, u \rangle - \alpha$$

as each $(x, f(x)) \in \operatorname{epi} f$.

But then

$$f^*(u) = \sup_{(x,\alpha) \in \operatorname{epi} f} f_{(x,\alpha)}(u)$$

is a supremum of convex and l.s.c. (affine) functions which is convex and l.s.c. by our earlier work.

Example 2.7.2
Let $1 < p, q$ such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then for $f(x) := \frac{|x|^p}{p}$,

$$f * (x) = \frac{|u|^q}{q}.$$

This can be shown by differentiating to find maximums.

Example 2.7.3
Let $f(x) := e^x$. Then

$$f^*(u) = \begin{cases} u \ln u - u, & u > 0 \\ 0, & u = 0 \\ \infty, & u < 0 \end{cases}$$

Example 2.7.4
Let $C \subseteq \mathbb{R}^m$, then

$$\delta_C^* = \sigma_C.$$

By definition,

$$\delta_C^*(y) := \sup_{y \in \operatorname{dom} \delta_C} \langle x, y \rangle - \delta_C(y)$$
$$= \sup_{y \in C} \langle x, y \rangle.$$

2.8 The Subdifferential Operator

Definition 2.8.1 (Subdifferential)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper. The subdifferential of f is the set-valued operator $\partial f : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ given by

$$x \mapsto \{u \in \mathbb{R}^m : \forall y \in \mathbb{R}^m, f(y) \geq f(x) + \langle u, y - x \rangle\}.$$

We say f is *subdifferentiable* at x if $\partial f(x) \neq \varnothing$.

The elements of $\partial f(x)$ are called the *subgradient* of f at x.

The idea is that for a differentiable convex function, the derivative at $x \in \mathbb{R}^n$ is the slope for a line tangent to x which lies strictly below f. If f is not differentiable at x, we can still ask for slopes of line segments tangent to x which lie below x.

Theorem 2.8.1 (Fermat)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper. Then

$$\operatorname{argmin} f = \{x \in \mathbb{R}^m : 0 \in \partial f(x)\} =: \operatorname{zer} \partial f.$$

Proof
Let $x \in \mathbb{R}^m$.

$$\begin{aligned} x \in \operatorname{argmin} f &\iff \forall y \in \mathbb{R}^m, f(x) \leq f(y) \\ &\iff \forall y \in \mathbb{R}^m, \langle 0, y - x \rangle + f(x) \leq f(y) \\ &\iff 0 \in \partial f(x). \end{aligned}$$

Example 2.8.2
Consider $f(x) = |x|$. Then

$$\partial f(x) = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$$

Lemma 2.8.3
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper. Then

$$\operatorname{dom} \partial f \subseteq \operatorname{dom} f.$$

Proof
We argue by the contrapositive, suppose $x \notin \operatorname{dom} f$. Then $f(x) = \infty$ and $\partial f(x) = \varnothing$.

Proposition 2.8.4
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ be closed and convex. Then

$$\partial \delta_C(x) = N_C(x).$$

Proof
Let $u \in \mathbb{R}^m$ and $x \in C = \operatorname{dom} \delta_C$. Then

$$\begin{aligned} u \in \partial \delta_C(x) &\iff \forall y \in \mathbb{R}^m, \delta_C(y) \geq \delta_C(x) + \langle u, y - x \rangle \\ &\iff \forall y \in C, \delta_C(y) \geq \delta_C(x) + \langle u, y - x \rangle \\ &\iff \forall y \in C, 0 \geq \langle u, y - x \rangle \\ &\iff u \in N_C(x). \end{aligned}$$

Consider the constrained optimization problem $\min f(x), x \in C$, where f is proper, convex, l.s.c. and $C \neq \varnothing$ is closed and convex. We can rephrase this as $\min f(x) + \delta_C(x)$.

In some cases, $\partial(f + \delta_C) = \partial f + \partial \delta_C = \partial f + N_C(x)$. Thus by Fermat's theorem, we look for some x where

$$0 \in \partial f(x) + N_C(x).$$

2.9 Calculus of Subdifferentials

The main question we are concerned with is whether the subdifferential operator is additive.

Proposition 2.9.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Then

$$\varnothing \neq \operatorname{ri}\operatorname{dom} f \subseteq \operatorname{dom} \partial f.$$

In particular,

$$\operatorname{ri}\operatorname{dom} f = \operatorname{ri}\operatorname{dom} \partial f$$
$$\overline{\operatorname{dom} f} = \overline{\operatorname{dom} \partial f}.$$

Definition 2.9.1 (Properly Separated)
Let $\varnothing \neq C_1, C_2 \subseteq \mathbb{R}^m$. Then C_1, C_2 are properly separated if there is some $b \neq 0$ such that

$$\sup_{c_1 \in C} \langle b, c_1 \rangle \leq \inf_{c_2 \in C} \langle b, c_2 \rangle$$

(separated) AND such that

$$\inf_{c_1 \in C_2} \langle b, c_1 \rangle < \sup_{c_2 \in C_2} \langle b, c_2 \rangle.$$

A problem with the definition of separated is that a set can be separated from itself. Indeed, the x-axis is separated from itself with itself as a separating hyperplane. To be properly separated, there must be some $c_1 \in C_1, c_2 \in C_2$ such that

$$\langle b, c_1 \rangle < \langle b, c_2 \rangle.$$

In otherwords, $C_1 \cup C_2$ is not fully contained in the hyperplane.

Proposition 2.9.2
Let $\varnothing \neq C_1, C_2 \subseteq \mathbb{R}^m$ be convex. Then C_1, C_2 are properly separated if and only if

$$\operatorname{ri} C_1 \cap \operatorname{ri} C_2 = \varnothing.$$

Proposition 2.9.3
Let $C_1, C_2 \subseteq \mathbb{R}^m$ be convex. Then

$$\operatorname{ri}(C_1 + C_2) = \operatorname{ri} C_1 + \operatorname{ri} C_2.$$

Moreover,

$$\operatorname{ri}(\lambda C) = \lambda(\operatorname{ri} C)$$

for all $\lambda \in \mathbb{R}$.

Proposition 2.9.4
Let $C_1 \subseteq \mathbb{R}^m$ and $C_2 \subseteq \mathbb{R}^p$ be convex. Then

$$\operatorname{ri}(C_1 \oplus C_2) = \operatorname{ri} C_1 \oplus \operatorname{ri} C_2.$$

Theorem 2.9.5
Let $C_1, C_2 \subseteq \mathbb{R}^m$ be convex such that $\operatorname{ri} C_1 \cap \operatorname{ri} C_2 \neq \varnothing$. For each $x \in C_1 \cap C_2$,

$$N_{C_1 \cap C_2}(x) = N_{C_1}(x) + N_{C_2}(x).$$

Proof
The reverse inclusion is not hard. Hence we check the inclusion only.

Let $x \in C_1 \cap C_2$ and $n \in N_{C_1 \cap C_2}(x)$. Then for each $u \in C_1 \cap C_2$,

$$\langle n, y - x \rangle \leq 0.$$

Set $E_1 := \operatorname{epi} \delta_{C_1} = C_1 \times [0, \infty) \subseteq \mathbb{R}^m \times \mathbb{R}$. Moreover, put

$$E_2 := \{(y, \alpha) : y \in C_2, \alpha \leq \langle n, y - x \rangle\} \subseteq \mathbb{R}^m \times \mathbb{R}.$$

By a previous fact,

$$\operatorname{ri} E_1 = \operatorname{ri} C_1 \times (0, \infty).$$

Similarly,

$$\operatorname{ri} E_2 = \{(y, \alpha), \alpha < \langle n, y - x \rangle\}.$$

We claim that $\operatorname{ri} E_1 \cap \operatorname{ri} E_2 = \varnothing$. Indeed, suppose towards a contradiction that there is some $(z, \alpha) \in \operatorname{ri} E_1 \cap \operatorname{ri} E_2$. Then

$$0 < \alpha < \langle n, z - x \rangle \leq 0$$

which is impossible.

It follows by a previous fact that E_1, E_2 are properly separated. Namely, there is $(b, \gamma) \in \mathbb{R}^m \times \mathbb{R} \setminus \{0\}$ such that

$$\begin{aligned} &\langle x, b \rangle + \alpha\gamma \leq \langle y, b \rangle + \beta\gamma && \forall (x, \alpha) \in E_1, (y, \beta) \in E_2 \\ &\langle \bar{x}, b \rangle + \bar{\alpha}\gamma < \langle \bar{y}, b \rangle + \bar{\beta}\gamma && \exists (\bar{x}, \bar{\alpha}) \in E_1, (\bar{y}, \bar{\beta}) \in E_2 \end{aligned}$$

We claim that $\gamma < 0$. Indeed, $(x, 1) \in E$ and $(x, 0) \in E_2$. So

$$\langle x, b \rangle + \gamma \leq \langle x, b \rangle \implies \gamma \leq 0.$$

Next we claim that $\gamma \neq 0$. Suppose to the contrary that $\gamma = 0$. But then

$$\begin{aligned} \langle x, b\rangle &\le \langle y, b\rangle && \forall (x,\alpha) \in E_1, (y,\beta) \in E_2 \\ \langle \bar{x}, b\rangle &< \langle \bar{y}, b\rangle && \exists (\bar{x},\bar{\alpha}) \in E_1, (\bar{y},\bar{\beta}) \in E_2 \end{aligned}$$

and C_1, C_2 are properly separated.

From our earlier fact, this contradicts the assumption that $\operatorname{ri} C_1 \cap \operatorname{ri} C_2 \neq \varnothing$. Altogether, $\gamma < 0$.

Our goal is to show that

$$n = \underbrace{-\frac{b}{\gamma}}_{\in N_{C_1}(x)} + \underbrace{n + \frac{b}{\gamma}}_{\in N_{C_2}(x)}.$$

First, we claim that $b \in N_{C_1}(x)$. This happens if and only if for all $y \in C_1$,

$$\langle y - x, b\rangle \le 0 \iff \langle b, y\rangle \le \langle b, x\rangle.$$

Indeed, we know that $(y, 0) \in E_1$. Moreover, $(x, 0) \in E_2$ by construction. Hence

$$\langle y, b\rangle + 0 \cdot \gamma \le \langle x, b\rangle + 0 \cdot \gamma.$$

Thus $b \in N_{C_1}(x) \implies -\frac{1}{\gamma} b \in N_{C_1}(x)$.

Now, for all $y \in C_2$, $(y, \langle n, y - x\rangle) \in E_2$ by construction, Hence for all $y \in C_2$,

$$\langle b, x\rangle + 0 \cdot \gamma \le \langle b, y\rangle + \gamma\langle n, y - x\rangle.$$

Equivalently,

$$\left\langle \frac{b}{\gamma} + n, y - x \right\rangle \le 0.$$

This shows that

$$\frac{b}{\gamma} + n \in N_{C_2}(x).$$

Thus $n \in N_{C_1}(x) + N_{C_2}(x)$ and we are done.

Proposition 2.9.6
Let $f : \mathbb{R}^m \to (-\infty, \infty)$ be convex, l.s.c. and proper. Let $x, u \in \mathbb{R}^m$. Then

$$u \in \partial f(x) \iff (u, -1) \in N_{\operatorname{epi} f}(x, f(x)).$$

Proof
Observe that $\operatorname{epi} f \neq \varnothing$ and is convex since f is proper and convex. Now let $u \in \mathbb{R}^m$. Then

$$
\begin{aligned}
&(u,-1) \in N_{\operatorname{epi} f}(x, f(x)) \\
&\iff x \in \operatorname{dom} f \wedge \forall (y,\beta) \in \operatorname{epi} f, \langle (y,\beta) - (x, f(x)), (u,-1)\rangle \leq 0 \\
&\iff x \in \operatorname{dom} f \wedge \forall (y,\beta) \in \operatorname{epi} f, \langle (y-x), \beta - f(x), (u,-1)\rangle \leq 0 \\
&\iff \forall (y,\beta) \in \operatorname{epi} f, \langle y-x, u\rangle + f(x) \leq \beta \\
&\iff \forall y \in \operatorname{dom} f, \langle y-x, u\rangle + f(x) \leq f(y) \\
&\iff u \in \partial f(x).
\end{aligned}
$$

Theorem 2.9.7
Let $f, g : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Suppose that $\operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g \neq \varnothing$. Then for all $x \in \mathbb{R}^m$,

$$\partial f(x) + \partial g(x) = \partial(f+g)(x).$$

Proof
Let $x \in \mathbb{R}^m$. If $x \notin \operatorname{dom}(f+g) = \operatorname{dom} f \cap \operatorname{dom} g$, then $\partial f(x) + \partial g(x) = \varnothing$. Also, $\partial(f+g)(x) = \varnothing$.

Suppose now that $x \in \operatorname{dom} f \cap \operatorname{dom} g = \operatorname{dom}(f+g)$. It is easy to check that

$$\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x).$$

We verify the reverse inclusion.

Pick any $u \in \partial(f+g)(x)$. By definition, for all $y \in \mathbb{R}^m$,

$$(f+g)(y) \geq (f+g)(x) + \langle u, y-x\rangle.$$

Consider the closed convex sets

$$
\begin{aligned}
E_1 &= \{(x,\alpha,\beta) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} : f(x) \leq \alpha\} = \operatorname{epi} f \times \mathbb{R} \\
E_2 &= \{(x,\alpha,\beta) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} : g(x) \leq \beta\} \cong \operatorname{epi} g \times \mathbb{R}.
\end{aligned}
$$

We claim that

$$(u,-1,-1) \in N_{E_1 \cap E_2}(x, f(x), g(x)).$$

Indeed, let $(y,\alpha,\beta) \in E_1, E_2$. We have by construction $f(y) - \alpha, g(y) - \beta \leq 0$.

Now,

$$\begin{aligned}
&\langle (u,-1,-1),(y,\alpha,\beta)-(x,f(x),g(x))\rangle \\
&= \langle u, y-x\rangle - (\alpha - f(x)) - (\beta - g(x)) \\
&= \langle u, y-x\rangle + (f+g)(x) - (\alpha+\beta) \\
&\le (f+g)(y) - \alpha - \beta && u \in \partial(f+g)(x) \\
&\le 0.
\end{aligned}$$

Next, we claim that $\operatorname{ri} E_i \cap \operatorname{ri} E_2 \neq \varnothing$. Indeed, by a previous fact,

$$\begin{aligned}
\operatorname{ri} E_1 &= \operatorname{ri}(\operatorname{epi} f \times \mathbb{R}) \\
&= \operatorname{ri} \operatorname{epi} f \times \mathbb{R}.
\end{aligned}$$

Similarly,

$$\operatorname{ri} E_2 = \{(x,\alpha,\beta) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} : g(x) < \beta\}.$$

Pick $z \in \operatorname{ri}\operatorname{dom} f \cap \operatorname{ri}\operatorname{dom} g$. Then $(z, f(z)+1, g(z)+1) \in \operatorname{ri} E_1, \operatorname{ri} E_2$. Hence, $(z, f(z)+1, g(z)+1) \in \operatorname{ri} E_1 \cap \operatorname{ri} E_2 \neq \varnothing$.

All in all, $E_1, E_2 \neq \varnothing$ are closed, convex, with $\operatorname{ri} E_1 \cap \operatorname{ri} E_2 \neq \varnothing$. Hence by the previous theorem,

$$N_{E_1 \cap E_2}(x, f(x), g(x)) = N_{E_1}(x, f(x), g(x)) + N_{E_2}(x, f(x), g(x)).$$

Now, it can be shown that $N_{\operatorname{epi} f \times \mathbb{R}} = N_{\operatorname{epi} f} \times N_{\mathbb{R}}$ and similarly for E_2. Therefore, there is some $u_1, u_2 \in \mathbb{R}^m, \alpha, \beta \in \mathbb{R}$ for which

$$(u,-1,-1) = (u_1, -\alpha, 0) + (u_2, 0, -\beta).$$

Thus $u = u_1 + u_2$ and $\alpha = \beta = 1$. It follows that

$$\begin{aligned}
(u_1, -1) &\in N_{\operatorname{epi} f}(x, f(x)) \\
(u_2, -1) &\in N_{\operatorname{epi} g}(x, g(x)).
\end{aligned}$$

From a previous proposition, we conclude that $u_1 \in \partial f(x)$ and $u_2 \in \partial g(x)$. Hence

$$u = u_1 + u_2 \in \partial f(x) + \partial g(x),$$

completing the proof.

Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Suppose $\phi \neq C \subseteq \mathbb{R}^m$ is closed and convex. Furthermore, suppose $\operatorname{ri} C \cap \operatorname{ri} \operatorname{dom} f \neq \varnothing$. Consider the problem

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

Then $\bar{x} \in \mathbb{R}^m$ solves (P) if and only if

$$(\partial f(\bar{x})) \cap (-N_C(\bar{x})) \neq \varnothing.$$

Indeed, we convert this to the unconstrained minimization problem $\min f + \delta_C$. This function is convex, l.s.c., and proper. By Fermat's theorem, $\bar{x}$ solves P if and only if

$$0 \in \partial(f + \delta_C)(\bar{x}).$$

Now, $\operatorname{ri} \operatorname{dom} f \cap \operatorname{ri} \operatorname{dom} \delta_C \neq \varnothing$. Hence by the previous theorem, $\bar{x}$ solves (P) if and only if

$$\begin{aligned} 0 \in \partial(f + \delta_C)(\bar{x}) = \partial f(\bar{x}) + N_C(\bar{x}) &\iff \exists u \in \partial f(\bar{x}), -u \in N_C(\bar{x}) \\ &\iff \partial f(\bar{x}) \cap (-N_C(\bar{x})) \neq \varnothing. \end{aligned}$$

Example 2.9.8
Let $d \in \mathbb{R}^m$ and $\varnothing \neq C \subseteq \mathbb{R}^m$ be convex and closed. Consider

$$\min_{x \in C} \langle d, x \rangle \qquad (P)$$

Let $\bar{x} \in \mathbb{R}^m$. Then $\bar{x}$ solves (P) if and only if

$$-d \in N_C(\bar{x}).$$

2.10 Differentiability

Definition 2.10.1 (Directional Derivative)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and $x \in \operatorname{dom} f$. The directional derivative of f at x in the direction of d is

$$f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

Definition 2.10.2 (Differentiable)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and $x \in \operatorname{dom} f$. f is differentiable at x if there is a linear operator $\nabla f(x) : \mathbb{R}^m \to \mathbb{R}^m$, called the derivative (gradient) of f at x, that satisfies

$$\lim_{0 \neq \|y\| \to 0} \frac{\|f(x+y) - f(x) - \nabla f(x) \cdot y\|}{\|y\|} = 0.$$

If f is differentiable at x, then the directional derivative of f at x in the direction of d is

$$f'(x; d) = \langle \nabla f(x), d \rangle.$$

Theorem 2.10.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex. Suppose $f(x) < \infty$. For each y, the quotient in the definition of $f'(x; y)$ is a non-decreasing function of $\lambda > 0$. So $f'(x; y)$ exists and

$$f'(x; y) = \inf_{\lambda > 0} \frac{f(x + \lambda y) - f(x)}{\lambda}.$$

Theorem 2.10.2
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex and proper. Let $x \in \operatorname{dom} f$ and $u \in \mathbb{R}^m$. Then u is a subgradient of f at x if and only if

$$\forall y \in \mathbb{R}^m, f'(x; y) \geq \langle u, y \rangle.$$

Proof
By definition,

$$\begin{aligned}
u \in \partial f(x) &\iff \forall y \in \mathbb{R}^m, \lambda > 0, f(x + \lambda y) \geq f(x) + \langle u, \lambda y \rangle \\
&\iff \forall y \in \mathbb{R}^m, \lambda > 0, \frac{f(x + \lambda y) - f(x)}{\lambda} \geq \langle u, y \rangle \\
&\iff \forall y \in \mathbb{R}^m, \inf_{\lambda > 0} \frac{f(x + \lambda y) - f(x)}{\lambda} \geq \langle u, y \rangle \\
&\iff \forall y \in \mathbb{R}^m, f'(x; y) \geq \langle u, y \rangle.
\end{aligned}$$

Theorem 2.10.3
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex and proper. Suppose $x \in \operatorname{dom} f$. If f is differentiable at x, then $\nabla f(x)$ is the unique subgradient of f at x.

Proof
Recall that for each $y \in \mathbb{R}^m$,

$$f'(x; y) = \langle \nabla f(x), y \rangle.$$

Let $u \in \mathbb{R}^m$. By the previous theorem,

$$\begin{aligned} u \in \partial f(x) &\iff \forall y \in \mathbb{R}^m, f'(x; y) \geq \langle u, y \rangle \\ &\iff \forall y \in \mathbb{R}^m, \langle \nabla f(x), y \rangle \geq \langle u, y \rangle. \end{aligned}$$

It is clear that $\nabla f(x) \in \partial f(x)$. Conversely, by setting $y := u - \nabla f(x)$. We see that

$$\begin{aligned} \langle \nabla f(x), u - \nabla f(x) \rangle \geq \langle u, u - \nabla f(x) \rangle &\iff \langle u - \nabla f(x), u - \nabla f(x) \rangle \leq 0 \\ &\iff u = \nabla f(x). \end{aligned}$$

Lemma 2.10.4
Let $\varphi : \mathbb{R} \to (-\infty, \infty]$ be a proper function that is differentiable on an interval $\varnothing \neq I \subseteq \operatorname{dom} \varphi$. If φ' is increasing on I, then φ is convex on I.

Proof
Fix $x, y \in I$ and $\lambda \in (0, 1)$. Let $\psi : \mathbb{R} \to (-\infty, \infty]$ be given by

$$z \mapsto \lambda\varphi(x) + (1 - \lambda)\varphi(z) - \varphi(\lambda x + (1 - \lambda)z).$$

Then

$$\psi'(z) = (1 - \lambda)\phi'(z) - (1 - \lambda)\phi'(\lambda x + (1 - \lambda)z)$$

and $\psi'(x) = 0 = \psi(x)$.

Since ϕ' is increasing, $\psi'(z) \leq 0$ when $z < x$ and $\psi'(z) > 0$ whenever $z > x$. It follows that ψ achieves its infimum on I at x.

That is, for all $y \in I$, $\psi(y) \geq \psi(x) = 0$. But then

$$\lambda\phi(x) + (1 - \lambda)\phi(y) \geq \phi(\lambda x + (1 - \lambda)y)$$

as desired.

Proposition 2.10.5
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper. Suppose that $\operatorname{dom} f$ is open and convex, and that f is differentiable on $\operatorname{dom} f$. The following are equivalent.

(i) f is convex

(ii) $\forall x, y \in \operatorname{dom} f, \langle x - y, \nabla f(y)\rangle + f(y) \leq f(x)$

(iii) $\forall x, y \in \operatorname{dom} f, \langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq 0$

Proof
(i) $\implies$ (ii) $\nabla f(y)$ is the unique subgradient of f at y. Hence for all $x \in \mathbb{R}^m$ and $y \in \operatorname{dom} f$,

$$f(x) \geq \langle x - y, \nabla f(y)\rangle + f(y).$$

(ii) $\implies$ (iii) We prove this in assignment 2.

(iii) $\implies$ (i) Fix $x, y \in \operatorname{dom} f$ and $z \in \mathbb{R}^m$. By assumption, $\operatorname{dom} f$ is open. Thus there is some $\epsilon > 0$ such that

$$\begin{aligned} y + (1+\epsilon)(x - y) = x + \epsilon(x - y) \in \operatorname{dom} f \\ y - \epsilon(x - y) = y + \epsilon(y - x) \in \operatorname{dom} f. \end{aligned}$$

By the convexity of $\operatorname{dom} f$, for every $\alpha \in (-\epsilon, 1+\epsilon)$, $y + \alpha(x - y) \in \operatorname{dom} f$.

Set $C = (-\epsilon, 1+\epsilon) \subseteq \mathbb{R}$ and $\phi : \mathbb{R} \to (-\infty, \infty]$ be given by

$$\phi(\alpha) := f(y + \alpha(x - y)) + \delta_C(\alpha).$$

By construction, ϕ is differentiable on C and for each $\alpha \in C$,

$$\phi'(\alpha) = \langle \nabla f(y + \alpha(x - y)), x - y\rangle.$$

Now, take $\alpha < \beta \in C$. Set

$$\begin{aligned} y_\alpha &:= y + \alpha(x - y) \\ y_\beta &:= y + \beta(x - y) \\ y_\beta - y_\alpha &= (\beta - \alpha)(x - y). \end{aligned}$$

Then by assumption,

$$\begin{aligned} \varphi'(\beta) - \varphi'(\alpha) &= \langle \nabla f(y + \beta(x - y)), x - y\rangle - \langle \nabla f(y + \alpha(x - y)), x - y\rangle \\ &= \langle \nabla f(y_\beta) - \nabla f(y_\alpha), x - y\rangle \\ &= \frac{1}{\beta - \alpha}\langle \nabla f(y_\beta) - \nabla f(y_\alpha), y_\beta - y_\alpha\rangle \\ &\geq 0. \end{aligned}$$

That is, φ' is increasing on C and φ is convex on C. But then

$$\begin{aligned} f(\alpha x+(1-\alpha)y) &= \varphi(\alpha) \\ &\leq \alpha\varphi(1)+(1-\alpha)\varphi(0) \\ &= \alpha f(x)+(1-\alpha)f(y). \end{aligned}$$

Example 2.10.6
Let A be a $m \times m$ matrix, and set $f : \mathbb{R}^m \to \mathbb{R}$ be given by

$$f(x) = \langle x, Ax \rangle.$$

Then $\boldsymbol{\nabla} f(x) = A + A^T$ and f is convex if and only if $A + A^T$ is posiitve semidefinite.

2.11 Conjugacy

Proposition 2.11.1
Let f, g be functions from $\mathbb{R}^m \to [-\infty, \infty]$. Then
(1) $f^{**} := (f^*)^* \leq f$
(2) $f \leq g \implies f^* \geq g^*, f^{**} \leq g^{**}$

Proposition 2.11.2 (Fenchel-Young Inequality)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper. Then for all $x, u \in \mathbb{R}^m$,

$$f(x) + f^*(u) \geq \langle x, u \rangle.$$

Proof
By definition, $f^*(x) = -\infty \iff f \equiv \infty$. Hence by assumption $f^*(\mathbb{R}^m) > 0$.

Now, let $x, u \in \mathbb{R}^m$. If $f(x) = \infty$, the inequality trivially holds. Otherwise,

$$f^*(u) := \sup_{y \in \mathbb{R}^m} \langle y, u \rangle - f(u) \geq \langle y, x \rangle - f(x)$$

as desired.

Proposition 2.11.3
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex and proper. For $x, u \in \mathbb{R}^m$,

$$u \in \partial f(x) \iff f(x) + f^*(x) = \langle x, u \rangle.$$

Proof
We have

$$
\begin{aligned}
& u \in \partial f(x) \\
\iff & \forall y \in \operatorname{dom} f, \langle y - x, u\rangle + f(x) \leq f(y) \\
\iff & \forall y \in \operatorname{dom} f, \langle y, u\rangle - f(y) \leq \langle x, u\rangle - f(x) \\
\iff & f^*(u) = \sup_{y \in \mathbb{R}^m} \langle y, u\rangle - f(y) \leq \langle x, u\rangle - f(x) \\
\iff & f^*(u) = \langle x, u\rangle - f(x). \qquad \langle x, u\rangle - f(x) \leq f^*(u)
\end{aligned}
$$

Proposition 2.11.4
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex and proper. Pick $x \in \mathbb{R}^n$ such that $\partial f(x) \neq \varnothing$. Then

$$f^{**}(x) = f(x).$$

Proof
Let $u \in \partial f(x)$. By the previous proposition,

$$\langle u, x\rangle = f(x) + f^*(u).$$

Consequently,

$$
\begin{aligned}
f^{**}(x) &:= \sup_{y \in \mathbb{R}^m} \langle x, y\rangle - f^*(y) \\
&\geq \langle x, u\rangle - f^*(u) \\
&= f(x).
\end{aligned}
$$

Conversely,

$$
\begin{aligned}
f^{**}(x) &= \sup_{y \in \mathbb{R}^m} \langle y, x\rangle - f^*(y) \\
&= \sup_{y \in \mathbb{R}^m} \langle y, x\rangle - \sup_{z \in \mathbb{R}^m} (\langle z, y\rangle - f(z)) \\
&= \sup_{y \in \mathbb{R}^m} \langle y, x\rangle + \inf_{z \in \mathbb{R}^m} (f(z) - \langle y, z\rangle) \\
&= \sup_{y \in \mathbb{R}^m} \inf_{z \in \mathbb{R}^m} (f(z) + \langle y, x - z\rangle) \\
&\leq \sup_{y \in \mathbb{R}^m} f(x) + \langle y, x - x\rangle \\
&= \sup_{y \in \mathbb{R}^m} f(x) \\
&= f(x).
\end{aligned}
$$

Proposition 2.11.5
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper. Then f is convex and l.s.c. if and only if

$$f = f^{**}.$$

In this case, f^* is also proper.

Corollary 2.11.5.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c. and proper. Then
(i) f^* is convex, l.s.c., and proper
(ii) $f^{**} = f$

Proof
To see (i), combine the previous proposition and the fact that f^* is always convex and l.s.c.

(ii) follows from the previous proposition.

Proposition 2.11.6
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Then

$$u \in \partial f(x) \iff x \in \partial f^*(u).$$

Proof
Recall that
$$u \in \partial f(x) \iff f(x) + f^*(u) = \langle x, u \rangle.$$
By a previous proposition, $g := f^*$ satifies $g^* = f$. Moreover, g is convex, l.s.c., and proper.

Hence,

$$\begin{aligned} u \in \partial f(x) &\iff f(x) + f^*(u) = \langle x, u \rangle \\ &\iff g^*(x) + g(u) = \langle x, u \rangle \\ &\iff x \in \partial g(u) = \partial f^*(u) \end{aligned}$$

as desired.

2.12 Coercive Functions

Theorem 2.12.1
Let $f : \mathbb{R}^m \to \mathbb{R}$ be proper, l.s.c. and compact $C \subseteq \mathbb{R}^m$ such that

$$C \cap \operatorname{dom} f \neq \varnothing.$$

Then the following hold:
(i) f is bounded below over C
(ii) f attains its minimal value over C

Proof
(i): Suppose towards a contradiction that f is not bounded below over C. There is a sequence x_n in C such that

$$\lim_n f(x_n) = -\infty.$$

Since C is (sequentially) compact, there there is a convergent subsequence $x_{k_n} \to \bar{x} \in C$. But f is l.s.c., hence

$$f(\bar{x}) \leq \liminf_n f(x_{k_n}) = -\infty$$

which contradicts the properness of f.

(ii): Since f is bounded below,

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

exists. There is a sequence x_n in C such that $f(x_n) \to f_{\min}$.

Again, there is a convergent subsequence $x_{k_n} \to \bar{x} \in C$. Then

$$f(\bar{x}) \leq \liminf_n f(x_{k_n}) = f_{\min}.$$

Thus $\bar{x}$ is a minimizer of f over C.

Definition 2.12.1 (Coercive Function)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$. Then f is coercive if

$$\lim_{\|x\| \to \infty} f(x) = \infty.$$

Definition 2.12.2 (Super Coercive)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$. Then f is super coercive if

$$\lim_{\|x\| \to \infty} \frac{f(x)}{\|x\|} = \infty.$$

Theorem 2.12.2
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper, l.s.c., and coercive. Let $C \subseteq \mathbb{R}^m$ be a closed subset of $\mathbb{R}^m$ satisfying

$$C \cap \operatorname{dom} f \neq \varnothing.$$

Then f attains its minimal value over C.

Proof
Let $x \in C \cap \operatorname{dom} f$. Since f is coercive, there is some M such that

$$\forall y, \|y\| > M \implies f(y) > f(x).$$

But then the set of minimizers of f over C is the same as the set of minimizers of f over $C \cap B(0; M)$. This set is compact. Hence by the previous theorem, f attains its minimal value over C.

2.13 Strong Convexity

Definition 2.13.1 (Lipschitz Function)
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ and $L \geq 0$. Then T is L-Lipschitz if for all $x, y \in \mathbb{R}^m$,

$$\|Tx - Ty\| \leq L\|x - y\|.$$

Example 2.13.1
Let $f : \mathbb{R}^m \to \mathbb{R}$ be given by

$$x \mapsto \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + x$$

where $A \succeq 0$ is positive semi-definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Then

(i) $\nabla f(x) = Ax$ for all $x \in \mathbb{R}^m$

(ii) ∇f is Lipschitz with constant $\|A\|$, the operator norm of A

Example 2.13.2
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ be closed and convex. Then P_C is Lipschitz continuous with constant 1.

Lemma 2.13.3 (Descent)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be differentiable on $\varnothing \neq D \subseteq \operatorname{int} \operatorname{dom} f$ such that ∇f is L-Lipschitz. Moreover, suppose that D is convex.
Then for all $x, y \in D$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2.$$

Proof
Recall that the fundamental theorem of calculus implies that

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle dt \\ &= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle dt. \end{aligned}$$

Hence

$$\begin{aligned} &|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \\ &= \left| \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle dt \right| \\ &\leq \int_0^1 |\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle| dt \\ &\leq \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\| \cdot \|y - x\| dt \\ &\leq \int_0^1 L\|x + t(y-x) - x\| \cdot \|y - x\| dt && f \text{ is } L\text{-Lipschitz} \\ &= \int_0^1 tL\|x - y\|^2 dt \\ &= \frac{L}{2}\|x - y\|^2. \end{aligned}$$

It follows that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2.$$

Theorem 2.13.4
Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and differentiable and $L > 0$. The following are equivalent:

(i) ∇f is L-Lipschitz

(ii) for all $x, y \in \mathbb{R}^m$, $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$

(iii) for all $x, y \in \mathbb{R}^m$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$

(iv) for all $x, y \in \mathbb{R}^m$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$

Proof
(i) $\implies$ (ii): This is the descent lemma.

(ii) $\implies$ (iii): If $\nabla f(x) = \nabla f(y)$, the this follows immediately from the subgradient inequality and the fact that $\partial f(x) = \{\nabla f(x)\}$.

Fix $x \in \mathbb{R}^m$ and define

$$h_x(y) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

Observe that h_x is convex, differentiable, with

$$\nabla h_x(y) = \nabla f(y) - \nabla f(x).$$

We claim that for all $y, z \in \mathbb{R}^m$,

$$h_x(z) \leq h_x(y) + \langle \nabla h_x(y), z - y \rangle + \frac{L}{2}\|z - y\|^2.$$

Indeed,

$$\begin{aligned}
h_x(z) &= f(z) - f(x) - \langle \nabla f(x), z - x \rangle \\
&\leq f(y) + \langle \nabla f(y), z - y \rangle + \frac{L}{2}\|z - y\|^2 - f(x) - \langle \nabla f(x), z - x \rangle \\
&= f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \langle \nabla f(x), z - y \rangle + \langle \nabla f(y), z - y \rangle + \frac{L}{2}\|z - y\|^2 \\
&= f(y) - f(x) - \langle \nabla f(x), y - x \rangle + \langle \nabla f(y) - \nabla f(x), z - y \rangle + \frac{L}{2}\|z - y\|^2 \\
&= h_x(y) + \langle \nabla h_x(y), z - y \rangle + \frac{L}{2}\|z - y\|^2.
\end{aligned}$$

By construction, $\nabla h_x(x) = 0$. But the convexity of h_x then asserts that x is a global minimizer of h_x. That is, for all $z \in \mathbb{R}^n$,

$$h_x(x) \leq h_x(z).$$

Pick $y, v \in \mathbb{R}^m$ be such that $\|v\| = 1$ and $\langle \nabla h_x(y), v \rangle = \|\nabla h_x(y)\|$. Set

$$z = y - \frac{\|\nabla h_x(y)\|}{L} v.$$

From the fact that x is a global minimizer, we have

$$\begin{aligned} 0 &= h_x(x) \\ &\leq h_x\left(y - \frac{\|\nabla h_x(y)\|}{L} v\right). \end{aligned}$$

On the other hand, the earlier inequality yields

$$\begin{aligned} 0 &= h_x(x) \\ &\leq h_x(y) - \frac{\|\nabla h_x(y)\|}{L} \langle \nabla h_x(y), v \rangle + \frac{1}{2L} \|\nabla h_x(y)\|^2 \|v\|^2 \\ &= h_x(y) - \frac{\|\nabla h_x(y)\|^2}{L} + \frac{1}{2L} \|\nabla h_x(y)\|^2 \\ &= h_x(y) - \frac{1}{2L} \|\nabla h_x(y)\|^2 \\ &= f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2L} \|\nabla f(x) - \nabla g(y)\|^2. \end{aligned}$$

<u>(iii) $\implies$ (iv):</u> Using (iii),

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \\ f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

<u>(iv) $\implies$ (i):</u> If $\nabla f(x) = \nabla f(y)$, the implication is trivial. We proceed assuming otherwise.

We have

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\|^2 &\leq L \langle \nabla f(x) - \nabla f(y), x - y \rangle \\ &\leq L \|\nabla f(x) - \nabla f(y)\| \cdot \|x - y\| \\ \|\nabla f(x) - \nabla f(y)\| &\leq L \|x - y\|. \end{aligned}$$

Example 2.13.5 (Firm Nonexpansiveness)
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ be closed and convex. Then for each $x, y \in \mathbb{R}^m$,

$$\|P_C(x) - P_c(y)\|^2 \leq \langle P_C(x) - P_C(y), x - y \rangle.$$

Example 2.13.6
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ be closed and convex. Let $f : \mathbb{R}^m \to \mathbb{R}$ be given by

$$f(x) = \frac{1}{2} d_C^2(x).$$

Then the following holds

(i) f is differentiable over $\mathbb{R}^m$ with $\nabla f(x) = x - P_C(x)$ for all $x \in \mathbb{R}^m$

(ii) ∇f is 1-Lipschitz

Indeed, for $x \in \mathbb{R}^m$, define

$$h_x(y) := f(x+y) - f(x) - \langle y, x - P_C(x) \rangle.$$

It can be shown that

$$\frac{|h_x(y)|}{\|y\|} \to 0$$

as $y \to 0$ by bounding $|h_x(y)| \leq \frac{1}{2}\|y\|^2$.

To see the 1-Lipschitz continuity of ∇f, we would apply the non-expansiveness of projections onto closed convex sets.

Theorem 2.13.7 (Second Order Characterization)
Let $f : \mathbb{R}^m \to \mathbb{R}$ be twice continuously differentiable over $\mathbb{R}^m$ and let $L \geq 0$. The following are equivalent.

(i) ∇f is L-Lipschitz

(ii) for all $x \in \mathbb{R}^m$, $\|\nabla^2 f(x)\| \leq L$ (operator norm)

Proof
(i) $\implies$ (ii) Suppose that ∇f is L-Lipschitz continuous. For any $y \in \mathbb{R}^m$ and $\alpha > 0$,

$$\|\nabla f(x + \alpha y) - \nabla f(x)\| \leq L\|x + \alpha y - x\| = \alpha L \|y\|.$$

That is,

$$\begin{aligned}\|\nabla^2 f(x)(y)\| &= \lim_{\alpha\downarrow 0} \frac{\|\nabla f(x+\alpha y) - \nabla f(x)\|}{\alpha} \\ &\leq \lim_{\alpha\downarrow 0} \frac{L\|x+\alpha y - x\|}{\alpha} \\ &= \lim_{\alpha\downarrow 0} L\|y\| \\ &= L\|y\|.\end{aligned}$$

Equivalently,

$$\|\nabla^2 f(x)\| \leq L$$

as desired. Note that we used the fact that $\nabla^2 f(x)(y) = (\nabla f)'(x; y)$.

(ii) $\implies$ (i) Suppose that $\|\nabla^2 f(x)\| \leq L$ and fix $x, y \in \mathbb{R}^m$. By the fundamental theorem of calculus,

$$\begin{aligned}\nabla f(x) &= \nabla f(y) + \int_0^1 \nabla^2 f(y + \alpha(x-y))(x-y) d\alpha \\ &= \nabla f(y) + \left[\int_0^1 \nabla^2 f(y+\alpha(x-y)) d\alpha\right](x-y)\end{aligned}$$

Hence

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\| &\leq \left\|\int_0^1 \nabla^2 f(x+\alpha(x-y)) d\alpha\right\| \cdot \|x-y\| \\ &\leq \int_0^1 \|\nabla^2 f(x+\alpha(x-y))\| d\alpha \|x-y\| \\ &\leq L\|x-y\|.\end{aligned}$$

Proposition 2.13.8
For a symmetric $A \in \mathbb{R}^{m\times m}$,

$$\sup_{\|x\|=1} \|Ax\| = \max_{1\leq i\leq m} |\lambda_i|$$

where λ_i are the eigenvalues of A.

Proof
Write x as a linear combination of some orthonormal eigenvector basis of A.

Proposition 2.13.9
A twice continuously differentiable function $f : \mathbb{R}^m \to \mathbb{R}$ is convex if and only if $\nabla^2 f(x)$ is positive semi-definite.

Proof
See A3.

Corollary 2.13.9.1
Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and twice continuously differentiable. Suppose $L \geq 0$. Then ∇f is L-Lipschitz if and only if for all $x \in \mathbb{R}^m$,

$$\lambda_{\max}(\nabla^2 f(x)) \leq L.$$

Proof
Since f is convex and twice continuously differentiable, $\nabla^2 f(x)$ is positive semidefinite everwhere. Combined with the earlier result,

$$\begin{aligned} L &\geq \|\nabla^2 f(x)\| \\ &= |\lambda_{\max}(\nabla^2 f(x))| \\ &= \lambda_{\max}(\nabla^2 f(x)). \end{aligned}$$

Example 2.13.10
Let $f : \mathbb{R}^m \to \mathbb{R}$ be given by

$$x \mapsto \sqrt{1 + \|x\|^2}.$$

Then

(i) f is convex

(ii) ∇f is 1-Lipschitz

Proposition 2.13.11
Let $\beta > 0$. $f : \mathbb{R}^m \to (-\infty, \infty]$ is β-strongly convex if and only if

$$f - \frac{\beta}{2}\|\cdot\|^2$$

is convex.

Proof
See A3.

Proposition 2.13.12
Let $f, g : \mathbb{R}^m \to (-\infty, \infty]$ and $\beta > 0$. Suppose that f is β-strongly convex and that g is convex. Then $f + g$ is β-strongly convex.

Proof
Define

$$h := \left(f - \frac{\beta}{2}\|\cdot\|^2\right) + g.$$

Then h is convex as it is the sum of two convex functions. Thus applying the previous proposition yields the result.

Proposition 2.13.13
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be strongly convex, l.s.c., and proper. Then f has a unique minimizer.

2.14 The Proximal Operator

Definition 2.14.1 (Proximal Point Mapping)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$. The proximal point mapping of f is the operator $\text{Prox}_f : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ given by

$$\text{Prox}_f(x) := \text{argmin}_{u \in \mathbb{R}^m}\{f(u) + \frac{1}{2}\|u - x\|^2\}.$$

Theorem 2.14.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Then for every $x \in \mathbb{R}^m$, $\text{Prox}_f(x)$ is a singleton.

Proof
For a fixed $x \in \mathbb{R}^m$,

$$h_x := \frac{1}{2}\|\cdot - x\|^2$$

is β-strongly convex for all $\beta < 1$. Therefore,

$$g_x := f + h_x$$

is strongly convex for every $x \in \mathbb{R}^m$.

We know that g_x is l.s.c. as f, h_x are l.s.c. Moreover, g_x is proper as f, g is proper with $\operatorname{dom} f \cap \operatorname{dom} g_x = \operatorname{dom} f$. Thus from the previous proposition,

$$\operatorname{argmin}_{u \in \mathbb{R}^m} g_x =: \operatorname{Prox}_f(x)$$

exists and is unique.

Example 2.14.2
For $\varnothing \neq C \subseteq \mathbb{R}^m$ closed and convex,

$$\operatorname{Prox}_{\delta_C} = P_C.$$

Proposition 2.14.3
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Let $x, p \in \mathbb{R}^m$. Then $p = \operatorname{Prox}_f(x)$ if and only if for all $y \in \mathbb{R}^m$,

$$\langle y - p, x - p \rangle + f(p) \leq f(y).$$

Proof
$(\implies)$ Suppose that $p = \operatorname{Prox}_f(x)$. For each $\lambda \in (0, 1)$, set

$$p_\lambda := \lambda y + (1 - \lambda) p.$$

Thus

$$\begin{aligned}
f(p) &\leq f(p_\lambda) + \frac{1}{2}\|x - p_\lambda\|^2 - \frac{1}{2}\|x - p\|^2 \\
&\leq f(p_\lambda) + \frac{1}{2}\|x - \lambda y - (1 - \lambda)p\|^2 - \frac{1}{2}\|x - p\|^2 \\
&= f(p_\lambda) + \frac{1}{2}\langle x - p - \lambda(y - p) - (x - p), x - p - \lambda(y - p) + (x - p)\rangle \\
&= f(p_\lambda) + \frac{1}{2}\langle -\lambda(y - p), 2(x - p) - \lambda(y - p)\rangle \\
&= f(p_\lambda) + \frac{\lambda}{2}\|y - p\|^2 - \lambda\langle x - p, y - p\rangle \\
&= f(\lambda y + (1 - \lambda)p) + \frac{\lambda^2}{2}\|y - p\|^2 - \lambda\langle x - p, y - p\rangle \\
f(p) &\leq \lambda f(y) + (1 - \lambda) f(p) + \frac{\lambda^2}{2}\|y - p\|^2 - \lambda\langle x - p, y - p\rangle
\end{aligned}$$

$$\lambda\langle x - p, y - p\rangle + \lambda f(p) \leq \lambda f(y) + \frac{\lambda^2}{2}\|y - p\|^2.$$

Division by λ and taking the limit as $\lambda \to 0$ yields the result.

$(\Longleftarrow)$ Suppose that

$$\langle y-p, x-p\rangle + f(p) \leq f(y).$$

Then

$$f(p) \leq f(y) - \langle y-p, x-p\rangle = f(y) + \langle x-p, p-y\rangle.$$

It follows that

$$\begin{aligned} f(p) + \frac{1}{2}\|x-p\|^2 &\leq f(y) + \langle x-p, p-y\rangle + \frac{1}{2}\|x-p\|^2 \\ &\leq f(y) + \langle x-p, p-y\rangle + \frac{1}{2}\|x-p\|^2 + \frac{1}{2}\|p-y\|^2 \\ &\leq f(y) + \|x-p+p-y\|^2 \\ &= f(y) + \|x-y\|^2. \end{aligned}$$

Example 2.14.4
Let $f : \mathbb{R}^m \to \mathbb{R}$ be given by

$$x \mapsto |x|.$$

Then

$$\operatorname{Prox}_f(x) := \begin{cases} x-1, & x > 1 \\ 0, & x \in [-1, 1] \\ x+1, & x < -1 \end{cases}$$

We need only apply the previous proposition and consider 3 cases.

Proposition 2.14.5
Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex, l.s.c., and proper. Then x minimizes f over $\mathbb{R}^m$ if and only if

$$x = \operatorname{Prox}_f(x).$$

Proof
By the previous proposition,

$$\begin{aligned} x = \operatorname{Prox}_f(x) &\iff \forall y \in \mathbb{R}^m, \langle y-x, x-x\rangle + f(x) \leq f(y) \\ &\iff \forall y \in \mathbb{R}^m, f(x) \leq f(y). \end{aligned}$$

Convexity is crucial for the proximal operator to be well-defined.

Example 2.14.6
Let $g, h : \mathbb{R} \to \mathbb{R}$ be given by

$$g(x) := \begin{cases} 0, & x \neq 0 \\ \lambda, & x = 0 \end{cases}$$

$$h(x) := \begin{cases} 0, & x \neq 0 \\ -\lambda, & x = 0 \end{cases}$$

for some $\lambda > 0$.

Then

$$\text{Prox}_h(x) = \begin{cases} \{x\}, & |x| > \sqrt{2\lambda} \\ \{0, x\}, & |x| = \sqrt{2\lambda} \\ \{0\}, & |x| < \sqrt{2\lambda} \end{cases}$$

$$\text{Prox}_h(x) = \begin{cases} \{x\}, & x \neq 0 \\ \varnothing, & x = 0 \end{cases}$$

Example 2.14.7 (Soft Threshold)
Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$x \mapsto \lambda|x|$$

for some $\lambda \geq 0$.

For all $x \in \mathbb{R}$,

$$\text{Prox}_f(x) = \begin{cases} x - \lambda, & x > \lambda \\ 0, & x \in [-\lambda, \lambda] \\ x + \lambda, & x < -\lambda \end{cases}$$

Note that the above formula can be written as

$$\text{Prox}_f(x) = \text{sign}(x)(|x| - \lambda)_+$$

where $\text{sign}(y)$ is $1, -1$ depending on the sign of y and $[-1, 1]$ if $y = 0$. Moreover, $(y)_+ = y$ if $y \geq 0$ and is 0 otherwise.

Theorem 2.14.8
Suppose $f : \mathbb{R}^m \to (-\infty, \infty]$ is given by

$$f(x) := \sum_{i=1}^{m} f_i(x_i)$$

for $f_i\mathbb{R} \to (-\infty, \infty]$ convex, l.s.c,, and proper.
Then for all $x \in \mathbb{R}^m$,

$$\text{Prox}_f(x) = (\text{Prox}_{f_i}(x_i))_{i=1}^m.$$

Proof
From A2, f is convex, l.s.c., and proper. We know that

$$\begin{aligned} p = \text{Prox}_f(x) &\iff \forall y \in \mathbb{R}^m, f(y) \geq f(p) + \langle y - p, x - p \rangle \\ &\iff \forall y \in \mathbb{R}^m, \sum_{i=1}^{m} f_i(y_i) \geq \sum_{i=1}^{m} f_i(p_i) + \sum_{i=1}^{m} (y_i - p_i)(x_i - p_i). \end{aligned}$$

In particular, for some $j \in [m]$, let $y_j \in \mathbb{R}$ and $y_i = 0$ for all $i \neq j$. Then

$$f_i(y_i) \geq f_i(p_i) + (y_i - p_i)(x_i - p_i)$$

which happens if and only if $p_i = \text{Prox}_{f_i}(x_i)$.

Conversely, if $f_i(y_i) \geq f_i(p_i) + (y_i - p_i)(x_i - p_i)$ for each $i \in [m]$, then clearly $p = \text{Prox}_f(x)$.

Example 2.14.9
Let $g : \mathbb{R}^m \to (-\infty, \infty]$ be given by

$$x \mapsto \begin{cases} -\alpha \sum_{i=1}^m \log x_i, & x > 0 \\ \infty, & \text{else} \end{cases}$$

where $\alpha > 1$.

Then

$$\text{Prox}_g(x) = \left(\frac{x_i + \sqrt{x_i^2 + 4\alpha}}{2} \right)_{i=1}^m$$

since

$$\text{Prox}_{g_i}(x_i) = \frac{x_i + \sqrt{x_i^2 + 4\alpha}}{2}.$$

This can be proven by differentiating to find the minimizer of $h_i(y_i) := g_i(y_i) + \frac{1}{2}(y_i - x_i)^2$.

Theorem 2.14.10
Let $g : \mathbb{R}^m \to (-\infty, \infty]$ be proper and $c > 0$. Let $a \in \mathbb{R}^m, \gamma \in \mathbb{R}$. For each $x \in \mathbb{R}^m$, define

$$f(x) = g(x) + \frac{c}{2}\|x\|^2 + \langle a, x\rangle + \gamma.$$

Then for all $x \in \mathbb{R}^m$,

$$\operatorname{Prox}_f(x) = \operatorname{Prox}_{\frac{1}{c+1}g}\left(\frac{x-a}{c+1}\right).$$

Proof
Indeed, recall that

$$\begin{aligned}\operatorname{Prox}_f(x) &:= \operatorname{argmin}_{u\in\mathbb{R}^m} f(u) + \frac{1}{2}\|u-x\|^2 \\ &= \operatorname{argmin}_{u\in\mathbb{R}^m} g(u) + \frac{c}{2}\|u\|^2 + \langle a, u\rangle + \gamma + \frac{1}{2}\|u-x\|^2.\end{aligned}$$

Now,

$$\begin{aligned}\frac{c}{2}\|u\|^2 + \langle a, u\rangle + \frac{1}{2}\|u-x\|^2 &= \frac{c}{2}\|u\|^2 + \langle a, u\rangle + \frac{1}{2}\|u\|^2 - \langle u, x\rangle + \frac{1}{2}\|x\|^2 \\ &= \frac{c+1}{2}\|u\|^2 - \langle u, x-a\rangle + \frac{1}{2}\|x\|^2 \\ &= \frac{c+1}{2}\left[\|u\|^2 - 2\left\langle u, \frac{x-a}{c+1}\right\rangle + \frac{1}{c+1}\|x\|^2\right] \\ &= \frac{c+1}{2}\left[\left\|u - \frac{x-a}{c+1}\right\|^2 - \frac{\|x-a\|^2}{c+1} + \frac{1}{c+1}\|x\|^2\right] \\ &= \frac{c+1}{2}\left\|u - \frac{x-a}{c+1}\right\|^2 - \frac{\|x-a\|^2}{2} + \frac{1}{2}\|x\|^2.\end{aligned}$$

Finally, since minimizers are preserved under positive scalar multiplication and translation,

$$\begin{aligned}\operatorname{Prox}_f(x) &= \operatorname{argmin}_{u\in\mathbb{R}^m} g(u) + \frac{c+1}{2}\left\|u - \frac{x+a}{c+1}\right\|^2 + \gamma - \frac{\|x-a\|^2}{2} + \frac{1}{2}\|x\|^2 \\ &= \operatorname{argmin}_{u\in\mathbb{R}^m} g(u) + \frac{c+1}{2}\left\|u - \frac{x+a}{c+1}\right\|^2 \\ &= \operatorname{argmin}_{u\in\mathbb{R}^m} \frac{1}{c+1}g(u) + \frac{1}{2}\left\|u - \frac{x-a}{c+1}\right\|^2 \\ &=: \operatorname{Prox}_{\frac{1}{c+1}g}\left(\frac{x+a}{c+1}\right).\end{aligned}$$

Example 2.14.11
Let $\mu \in \mathbb{R}$ and $\alpha \geq 0$. Consider $f : \mathbb{R} \to (-\infty, \infty]$ given by

$$f(x) := \begin{cases} \mu x, & x \in [0, \alpha] \\ \infty, & \text{else} \end{cases}$$

For each $x \in \mathbb{R}$,

$$f(x) = \mu x + \delta_{[0,\alpha]}(x).$$

Moreover,

$$\operatorname{Prox}_f(x) = \min(\max(x - \mu, 0), \alpha).$$

Indeed, apply the previous theorem with $g = \delta_{[0,\alpha]}$ and $c = \gamma = 0$. Then

$$\operatorname{Prox}_f(x) = \operatorname{Prox}_g(x - \mu) = P_C(x - \mu).$$

Theorem 2.14.12
Let $g : \mathbb{R} \to (-\infty, \infty]$ be convex, l.s.c. and proper such that $\operatorname{dom} g \subseteq [0, \infty)$ and let $f : \mathbb{R}^m \to \mathbb{R}$ be given by

$$f(x) = g(\|x\|).$$

Then

$$\operatorname{Prox}_f(x) = \begin{cases} \operatorname{Prox}_g(\|x\|) \frac{x}{\|x\|}, & x \neq 0 \\ \{u \in \mathbb{R}^m : \|u\| = \operatorname{Prox}_g(x)\}, & x = 0 \end{cases}$$

Proof
Case I: $x = 0$ By definition,

$$\operatorname{Prox}_f(x) = \operatorname{argmin}_{u \in \mathbb{R}^m} f(u) + \frac{1}{2}\|u\|^2.$$

By the change of variable $w = \|u\|$, then above set of minimizers is the same as

$$\operatorname{argmin}_{w \in \mathbb{R}^m} g(w) + \frac{1}{2} w^2 =: \operatorname{Prox}_g(0).$$

Case II: $x \neq 0$ By definition, $\text{Prox}_f(x)$ is the set of solutions to the minimization problem

$$\begin{aligned}
&\min_{u\in\mathbb{R}^m} g(\|u\|) + \frac{1}{2}\|u-x\|^2 \\
&= \min_{u\in\mathbb{R}^m} g(\|u\|) + \frac{1}{2}\|u\|^2 - \langle u, x\rangle + \frac{1}{2}\|x\|^2 \\
&= \min_{\alpha\geq 0} \min_{u\in\mathbb{R}^m:\|u\|=\alpha} g(\alpha) + \frac{1}{2}\alpha^2 - \langle u, x\rangle + \frac{1}{2}\|x\|^2
\end{aligned}$$

Now, $\langle u, x\rangle \leq \|u\|\cdot\|x\|$ by the Cauchy-Schwartz inequality with equality when $u = \lambda x$ for some $\lambda \geq 0$. Thus

$$\left\{\alpha\frac{x}{\|x\|}\right\} = \min_{u\in\mathbb{R}^m:\|u\|=\alpha} g(\alpha) + \frac{1}{2}\alpha^2 - \langle u, x\rangle + \frac{1}{2}\|x\|^2.$$

The values of α which minimize $\alpha\frac{x}{\|x\|}$ are then given by

$$\begin{aligned}
&\min_{\alpha\geq 0} g(\alpha) + \frac{1}{2}\alpha^2 - \alpha\|x\| + \frac{1}{2}\|x\|^2 \\
&= \min_{\alpha\geq 0} g(\alpha) + \frac{1}{2}(\alpha - \|x\|)^2.
\end{aligned}$$

This is precisely $\text{Prox}_g(\|x\|)$.

Hence

$$\text{Prox}_f(x) = \text{Prox}_g(\|x\|)\frac{x}{\|x\|}$$

as desired.

Example 2.14.13
Let $\alpha > 0, \lambda \geq 0$, and $f : \mathbb{R}^{\rightarrow}(-\infty, \infty]$ be given by

$$f(x) = \begin{cases} \lambda|x|, & |x| \leq \alpha \\ \infty, & |x| > \alpha \end{cases}$$

Then f is convex, l.s.c. and proper (see A3).

Define

$$g(x) = \begin{cases} \lambda x, & x \in [0, \alpha] \\ \infty, & x \notin [0, \alpha] \end{cases}$$

so that $f(x) = g(|x|)$. By the previous theorem,

$$\operatorname{Prox}_f(x) = \begin{cases} \operatorname{Prox}_g(|x|)\operatorname{sgn}(x), & x \neq 0 \\ 0, & x = 0 \end{cases}$$
$$= \min(\max(|x| - \lambda, 0), \alpha)\operatorname{sgn}(x).$$

Example 2.14.14
Let $w, \alpha \in \mathbb{R}^m_+$ and $f : \mathbb{R}^m \to (-\infty, \infty]$ given by

$$f(x) = \begin{cases} \sum_{i=1}^m w_i|x_i|, & -\alpha \leq x \leq \alpha \\ \infty, & \text{else} \end{cases}$$

Then $\operatorname{Prox}_f(x) = (\min(\max(|x_i| - w_i, 0), \alpha_i)\operatorname{sgn}(x_i))_{i=1}^m$ (see A3).

Moreover, consider the problem

$$\min \sum_{i=1}^m w_i|x_i| \qquad (P)$$
$$|x_i| \leq \alpha_i, \qquad \forall i \in [m]$$

Let the sequence $(x_n)_{n\geq 0}$ be recursively defined by $x_0 \in \mathbb{R}^m$ and $x_{n+1} = \operatorname{Prox}_f(x_n)$. Then $x_n \to \bar{x}$ where $\bar{x}$ is a minimizer of (P).

2.15 Nonexpansive & Averaged Operators

We use $\operatorname{Id} : \mathbb{R}^m \to \mathbb{R}^m$ to denote the $m \times m$ identity matrix.

Definition 2.15.1 (Nonexpansive)
Let $T : \mathbb{R}^m \to \mathbb{R}^m$. Then T is nonexpansive if for all $x, y \in \mathbb{R}^m$,

$$\|Tx - Ty\| \leq \|x - y\|$$

Definition 2.15.2 (Firmly Nonexpansive)
Let $T : \mathbb{R}^m \to \mathbb{R}^m$. Then T is firmly nonexpansive (f.n.e.) if for all $x, y \in \mathbb{R}^m$,

$$\|Tx - Ty\|^2 + \|(\operatorname{Id} - T)x - (\operatorname{Id} - T)y\|^2 \leq \|x - y\|^2$$

Definition 2.15.3 (Averaged)
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ and $\alpha \in (0,1)$. Then T is α-averaged if there is some $N : \mathbb{R}^m \to \mathbb{R}^m$ such that N is nonexpansive and

$$T = (1-\alpha)\operatorname{Id} + \alpha N.$$

Proposition 2.15.1
$T : \mathbb{R}^m \to \mathbb{R}^m$. The following are equivalent.

(i) T is f.n.e.

(ii) $\operatorname{Id} - T$ is f.n.e.

(iii) $2T - \operatorname{Id}$ is nonexpansive

(iv) for all $x, y \in \mathbb{R}^m$, $\|Tx - Ty\|^2 \le \langle x - y, Tx - Ty\rangle$.

(v) for all $x, y \in \mathbb{R}^m$, $\langle Tx - Ty, (\operatorname{Id} - T)x - (\operatorname{Id} - T)y\rangle \ge 0$

Proof
(i) $\iff$ (ii): This is clear from the definition.

(i) $\iff$ (iii) $\iff$ (iv) $\iff$ (v): See A3.

We can refine the previous result when T is linear.

Proposition 2.15.2
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be linear. Then the following are equivalent.

(i) T is f.n.e.

(ii) $\|2T - \operatorname{Id}\| \le 1$

(iii) for all $x \in \mathbb{R}^m$, $\|Tx\|^2 \le \langle x, Tx\rangle$

(iv) for all $x \in \mathbb{R}^m$, $\langle Tx, x - Tx\rangle \ge 0$

Proof
(i) $\iff$ (ii) We know that T is f.n.e. if and only if $2T - \operatorname{Id}$ is nonexpansive. This happens if and only if for all $x \neq y$,

$$\begin{aligned} \|(2T - \operatorname{Id})(x-y)\| &= \|(2T - \operatorname{Id})x - (2T - \operatorname{Id})y\| \\ &\le \|x - y\| \\ &\iff \\ &\|2T - \operatorname{Id}\| \le 1. \end{aligned}$$

(i) $\iff$ (iii) This is easily seen by the previous proposition and the fact that $Tx - Ty = T(x-y)$.

(i) $\iff$ (iv) This is seen by applying the previous proposition and observing that $Tx - Ty = T(x-y)$ as well as

$$(\operatorname{Id} - T)x - (\operatorname{Id} - T)y = x - y - T(x-y).$$

Observe that T is f.n.e. if and only if $N := 2T - \operatorname{Id}$ is nonexpansive if and only if $2T = \operatorname{Id} + N$ for N nonexpansive if and only if $T = \frac{1}{2}\operatorname{Id} + \frac{1}{2}N$ for N nonexpansive if and only if T is $\frac{1}{2}$-averaged.

Example 2.15.3
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ be convex and closed. Then $P_C(x)$ is f.n.e. Indeed, for all $x, y \in \mathbb{R}^m$,

$$\|P_C(x) - P_C(y)\| \leq \langle P_C(x) - P_C(y), x - y\rangle.$$

Example 2.15.4
Suppose that $T = -\frac{1}{2}\operatorname{Id}$. Then T is averaged but NOT f.n.e.

We have

$$T = \frac{1}{4}\operatorname{Id} + \frac{3}{4}(-\operatorname{Id})$$

and so T is $\frac{3}{4}$-averaged.

But T is not f.n.e. as for all $0 \neq x \in \mathbb{R}^m$,

$$\begin{aligned} \|Tx\|^2 + \|x - Tx\|^2 &= \frac{1}{4}\|x\|^2 + \frac{9}{4}\|x\|^2 \\ &= \frac{5}{2}\|x\|^2 \\ &> \|x\|^2. \end{aligned}$$

Example 2.15.5
$T := -\operatorname{Id}$ is nonexpansive but NOT averaged. Indeed suppose there is some nonexpansive $N : \mathbb{R}^m \to \mathbb{R}^m$ and $\alpha \in (0,1)$ such that

$$\begin{aligned} T = (1-\alpha)\operatorname{Id} + \alpha N &\iff -\operatorname{Id} = (1-\alpha)\operatorname{Id} + \alpha N \\ &\iff (-1+\alpha)\operatorname{Id} = \alpha N \\ &\iff N = \frac{\alpha - 2}{\alpha}\operatorname{Id}. \end{aligned}$$

But then

$$\begin{aligned} \|N\| = \left|\frac{\alpha - 2}{\alpha}\right| &\leq 1 \\ \iff \frac{2-\alpha}{\alpha} &\leq 1 \\ \iff 2 - \alpha &\leq \alpha \\ \iff \alpha &\geq 1 \end{aligned}$$

which is impossible by the definition of averaged.

Proposition 2.15.6
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive. Then T is continuous.

Proof
Suppose $x_n \to \bar{x}$. Then

$$\|Tx_n - T\bar{x}\| \leq \|x_n - \bar{x}\| \to 0.$$

Definition 2.15.4 (Fixed Point)
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ then

$$\operatorname{Fix} T := \{x \in \mathbb{R}^m : x = Tx\}.$$

2.16 Féjer Monotonocity

Definition 2.16.1 (Féjer Monotone)
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ and $(x_n)_{n\in\mathbb{N}}$ a sequence in $\mathbb{R}^m$. Then (x_n) is a Féjer monotone with respect to C if for all $c \in C, n \in \mathbb{N}$,

$$\|x_{n+1} - c\| \leq \|x_n - c\|.$$

Example 2.16.1
Suppose $\operatorname{Fix} T \neq \varnothing$ for some nonexpansive $T : \mathbb{R}^m \to \mathbb{R}^m$. For any $x_0 \in \mathbb{R}^n$, the sequence defined recursively by

$$x_n := T(x_{n-1})$$

is Féjer monotone with respect to $\operatorname{Fix} T$.

Proposition 2.16.2
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ and $(x_n)_{n\geq 0}$ a Féjer monotone sequence in $\mathbb{R}^m$ with respect to C. The following hold:

(i) (x_n) is bounded

(ii) for every $c \in C$, $(\|x_n - c\|)_{n\geq 0}$ converges

(iii) $(d_C(x_n))_{n\geq 0}$ is decreasing and converges

Proof
Fix $c \in C$. We have

$$\begin{aligned} \|x_n\| &\leq \|c\| + \|x_n - c\| \\ &\leq \|c\| + \|x_0 - c\|. \end{aligned}$$

Hence (x_n) is a bounded sequence.

Now, $\|x_n - c\|$ is bounded below by 0 and monotonic, hence necessarily converges to the infimum.

Observe that for each $n \in \mathbb{N}, c \in C$,

$$\|x_{n+1} - c\| \leq \|x_n - c\|.$$

Taking infimums on both sides preserve this inequality.

Recall the following analysis fact.

Proposition 2.16.3
A bounded sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^m$ converges if and only if it has a unique cluster point.

Proof
The forward direction is clear. Suppose now that $(x_n)_{n\in\mathbb{N}}$ has a unique cluster point $\bar{x}$.

Suppose that $x_n \not\to \bar{x}$. Then there is some $\epsilon_0 > 0$ and subsequence x_{k_n} such that for all n,

$$\|x_{k_n} - \bar{x}\| \geq \epsilon_0.$$

But then $(x_{k_n})_{n\in\mathbb{N}}$ is bounded and hence contains a convergent subsequence. This is still a subsequence of $(x_n)_{n\in\mathbb{N}}$ but cannot converge to $\bar{x}$.

It follows that $(x_n)_{n\in\mathbb{N}}$ has more than one cluster point. By contradiction, $x_n \to \bar{x}$.

Lemma 2.16.4
Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$ and $\varnothing \neq C \subseteq \mathbb{R}^m$ be such that for all $c \in C$, $(\|x_n - c\|)_{n\in\mathbb{N}}$ converges and every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in C.
Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in C.

Proof
(x_n) is necessarily bounded since $\|x_n\| \leq \|c\| + \|x_n - c\|$ is bounded. It suffices by the previous proposition to show that $(x_n)_{n\in\mathbb{N}}$ has a unique cluster point.

Let x, y be two cluster points of $(x_n)_{n\in\mathbb{N}}$. That is, there are subsequences

$$x_{k_n} \to x, x_{\ell_n} \to y.$$

By assumption, $x, y \in C$. Hence $\|x_n - x\|, \|x_n - y\|$ converges.

Observe that

$$\begin{aligned}
&2\langle x_n, x - y\rangle \\
&= \|x_n\|^2 + \|y\|^2 - 2\langle x_n, y\rangle - \|x_n\|^2 - \|x\|^2 + 2\langle x_n, x\rangle + \|x\|^2 - \|y\|^2 \\
&= \|x_n - y\| - \|x_n - x\|^2 + \|x\|^2 - \|y\|^2 \\
&\to L \in \mathbb{R}^m.
\end{aligned}$$

But then taking the limit along k_n, ℓ_n,

$$\begin{aligned}
\langle x, x - y\rangle &= \langle y, x - y\rangle \\
\|x - y\|^2 &= 0 \\
x &= y.
\end{aligned}$$

Theorem 2.16.5
Let $\varnothing \neq C \subseteq \mathbb{R}^m$ and $(x_n)_{n\in\mathbb{N}}$ a sequence in $\mathbb{R}^m$. Suppose that $(x_n)_{n\in\mathbb{N}}$ is Féjer monotone with respect to C, and that every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in C. Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in C.

Proof
We know that for all $c \in C$,

$$\|x_n - c\|$$

converges. Hence the result follows from the previous lemma.

Let $x, y \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$. By computation,

$$\|\alpha x + (1-\alpha)y\|^2 + \alpha(1-\alpha)\|x-y\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2.$$

Theorem 2.16.6
Let $\alpha \in (0,1]$ and $T : \mathbb{R}^m \to \mathbb{R}^m$ be α-averaged such that $\operatorname{Fix} T \neq \varnothing$. Let $x_0 \in \mathbb{R}^m$. Define

$$x_{n+1} := Tx_n.$$

The following hold:

(i) $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\operatorname{Fix} T$.

(ii) $Tx_n - x_n \to 0$.

(iii) $(x_n)_{n\in\mathbb{N}}$ converges to a point in $\operatorname{Fix} T$.

Proof
Now, T being averaged implies that it is nonexpansive. The example earlier shows that $(x_n)_{n\in\mathbb{N}}$ is Féjer monotone.

By the definition of averaged, there is some nonexpansive $N : \mathbb{R}^m \to \mathbb{R}^m$ such that

$$T = (1-\alpha)\operatorname{Id} + \alpha N.$$

Hence for each $n \in \mathbb{N}$,

$$x_{n+1} = (1-\alpha)x_n + \alpha N(x_n).$$

Let $f \in \operatorname{Fix} T$.

$$\begin{aligned}
\|x_{n+1} - f\|^2 &= \|(1-\alpha)(x_n - f) + \alpha(N(x_n) - f)\|^2 \\
&= (1-\alpha)\|x_n - f\|^2 + \alpha\|N(x_n) - N(f)\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
&\leq (1-\alpha)\|x_n - f\|^2 + \alpha\|x_n - f\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
&= \|x_n - f\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
\alpha(1-\alpha)\|N(x_n) - x_n\|^2 &\leq \|x_n - f\|^2 - \|x_{n+1} - f\|^2.
\end{aligned}$$

By a telescoping sum argument,

$$\begin{aligned}
\sum_{n=0}^{k} \alpha(1-\alpha)\|N(x_0) - x_n\|^2 &= \|x_0 - f\|^2 - \|x_{k+1} - f\|^2 \\
&\leq \|x_0 - f\|^2.
\end{aligned}$$

By our work with non-negative monotone series, it must be that $\|N(x_n) - x_n\| \to 0$.

In particular,

$$\|Tx_n - x_n\| = \|(1-\alpha)x_n + \alpha N(x_n) - x_n\| \qquad = \alpha\|N(x_n) - x_n\|$$
$$\to 0.$$

Now, $(x_n)_{n\in\mathbb{N}}$ is Féjer monotone with respect to $\operatorname{Fix} T = \operatorname{Fix} N$. Let $\bar{x}$ be a cluster point of $(x_n)_{n\in\mathbb{N}}$, say $x_{k_n} \to \bar{x}$. Observe that N being nonexpansive implies that N is continuous.

Since $Nx_n - x_n \to 0$, we must also have $Nx_{k_n} - x_{k_n} \to 0$. Thus

$$Nx_{k_n} = (Nx_{k_n} - x_{k_n}) + x_{k_n} \to 0 + \bar{x}.$$

By continuity,

$$N\bar{x} = \lim_n Nx_{k_n} = \bar{x}.$$

That is, every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in $\operatorname{Fix} N = \operatorname{Fix} T$. Combined with a previous theorem, this yield the proof.

Corollary 2.16.6.1
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be f.n.e. and suppose that $\operatorname{Fix} T \neq \varnothing$. Put $x_0 \in \mathbb{R}^m$. Recursively define

$$x_{n+1} := Tx_n.$$

There is some $\bar{x} \in \operatorname{Fix} T$ such that

$$x_n \to \bar{x}.$$

Proof
Since T is f.n.e., T is also averaged. The result follows then by the previous theorem.

Proposition 2.16.7
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Then Prox_f is f.n.e.

Proof
Let $x, y \in \mathbb{R}^m$. Set $p := \operatorname{Prox}_f(x)$ and $q := \operatorname{Prox}_f(y)$.

By our work with the proximal operator, p, q are characterized as $\forall z \in \mathbb{R}^m$,

$$\langle z - p, x - p\rangle + f(p) \leq f(z)$$
$$\langle z - q, y - q\rangle + f(q) \leq f(z).$$

By choosing $z = p, q$,

$$\begin{aligned} \langle q-p, x-p\rangle + f(p) &\leq f(q) \\ \langle p-q, y-q\rangle + f(q) &\leq f(p) \\ \langle q-p, (x-p)-(y-q)\rangle &\leq 0 \\ \langle p-q, (x-p)-(y-q)\rangle &\geq 0. \end{aligned}$$

But then by our characterization of f.n.e. operators, Prox_f is f.n.e.

Corollary 2.16.7.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper such that $\operatorname{argmin} f \neq \varnothing$. Let $x_0 \in \mathbb{R}^m$ and updated via

$$x_{n+1} = \operatorname{Prox}_f(x_n).$$

There is some $\bar{x} \in \operatorname{argmin} f$ such that $x_n \to \bar{x}$.

Proof
Recall that

$$x \in \operatorname{argmin} f \iff x = \operatorname{Prox}_f(x) \iff x \in \operatorname{Fix} \operatorname{Prox}_f .$$

Thus $\operatorname{argmin} f = \operatorname{Fix} \operatorname{Prox}_f \neq \varnothing$.

By the previous proposition, Prox_f is f.n.e. Thus the result follows from a previous theorem.

2.17 Composition of Averaged Operators

Consider the following identity for all $x, y \in \mathbb{R}^m, \alpha \in \mathbb{R} \setminus \{0\}$:

$$\alpha^2 \left(\|x\|^2 - \left\| \left(1 - \frac{1}{\alpha}\right) x + \frac{1}{\alpha} y \right\|^2 \right) = \alpha \left(\|x\|^2 - \frac{1-\alpha}{\alpha} \|x-y\|^2 - \|y\|^2 \right)$$

Proposition 2.17.1
Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive and $\alpha \in (0, 1)$. The following are equivalent:

1. T is α-averaged
2. $\left(1 - \frac{1}{\alpha}\right) \operatorname{Id} + \frac{1}{\alpha} T$ is nonexpansive
3. For each $x, y \in \mathbb{R}^m$, $\|Tx - Ty\|^2 \leq \|x-y\|^2 - \frac{1-\alpha}{\alpha} \|(\operatorname{Id} - T)x - (\operatorname{Id} - T)y\|^2$

Proof
(i) $\iff$ (ii): We have T is α-averaged if and only if there is some $N : \mathbb{R}^m \to \mathbb{R}^m$ nonexpansive such that

$$\begin{aligned} T = (1-\alpha)\operatorname{Id} + \alpha N \\ &\iff N = \frac{1}{\alpha}(T - (1-\alpha)\operatorname{Id}) \\ &\iff N = \left(1 - \frac{1}{\alpha}\right)\operatorname{Id} + \frac{1}{\alpha}T \end{aligned}$$

if and only if $\left(1-\frac{1}{\alpha}\right)\operatorname{Id} + \frac{1}{\alpha}T$ is nonexpansive.

(ii) $\iff$ (iii) By definition $\left(1-\frac{1}{\alpha}\right)\operatorname{Id} + \frac{1}{\alpha}T$ is nonexpansive if and only if for all $x, y \in \mathbb{R}^m$,

$$\begin{aligned} &\|x-y\|^2 \\ &\geq \left\| \left(1-\frac{1}{\alpha}\right)x + \frac{1}{\alpha}Tx - \left(1-\frac{1}{\alpha}\right)y - \frac{1}{\alpha}Ty \right\|^2 \\ &= \left\| \left(1-\frac{1}{\alpha}\right)(x-y) + \frac{1}{\alpha}(Tx-Ty) \right\|^2 \\ &= \|x-y\|^2 - \frac{1}{\alpha}\left(\|x-y\|^2 - \frac{1-\alpha}{\alpha}\|(x-Tx)-(y-Ty)\|^2 - \|Tx-Ty\|^2 \right) \quad \text{identity} \\ 0 &\geq -\frac{1}{\alpha}\left(\|x-y\|^2 - \frac{1-\alpha}{\alpha}\|(x-Tx)-(y-Ty)\|^2 - \|Tx-Ty\|^2 \right) \\ 0 &\leq \|x-y\|^2 + \frac{1-\alpha}{\alpha}\|(x-Tx)-(y-Ty)\|^2 - \|Tx-Ty\|^2 \qquad \alpha > 0. \end{aligned}$$

Theorem 2.17.2
Let $\alpha_1, \alpha_2 \in (0,1)$ and $T_i : \mathbb{R}^m \to \mathbb{R}^m$ be α_i-averaged. Define

$$\begin{aligned} T &:= T_1 T_2 \\ \alpha &:= \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2}. \end{aligned}$$

Then T is α-averaged.

Proof
First observe that by computation,

$$\alpha \in (0,1) \iff \alpha_1(1-\alpha_2) < 1 - \alpha_2$$

which is a tautology.

By the previous proposition, for each $x, y \in \mathbb{R}^m$,

$$\begin{aligned}
&\|Tx - Ty\|^2 \\
&= \|T_1T_2x - T_1T_2y\|^2 \\
&\leq \|T_2x - T_2y\|^2 - \frac{1-\alpha_1}{\alpha_1}\|(\mathrm{Id} - T_1)T_2x - (\mathrm{Id} - T_1)T_2y\|^2 \\
&\leq \|x - y\|^2 - \frac{1-\alpha_2}{\alpha_2}\|(\mathrm{Id} - T_2)x - (\mathrm{Id} - T_2)y\|^2 - \frac{1-\alpha_1}{\alpha_1}\|(\mathrm{Id} - T_1)T_2x - (\mathrm{Id} - T_1)T_2y\|^2 \\
&= \|x - y\|^2 - V_1 - V_2.
\end{aligned}$$

Set

$$\beta := \frac{1-\alpha_1}{\alpha_1} + \frac{1-\alpha_2}{\alpha_2} > 0.$$

By computation,

$$V_1 + V_2 \geq \frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2.$$

Consequently,

$$\begin{aligned}
\|Tx - Ty\|^2 &\leq \|x - y\|^2 - \frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 \\
&= \|x - y\|^2 - \frac{1-\alpha}{\alpha}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2.
\end{aligned}$$

By the previous proposition, we are done.

Chapter 3

Constrained Convex Optimization

We now consider the problem

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

where $f : \mathbb{R}^m \to (-\infty, \infty]$ is convex, l.s.c., and proper with $C \neq \varnothing$ being convex and closed.

3.1 Optimality Conditions

Recall that if $\operatorname{ri} C \cap \operatorname{ri} \operatorname{dom} f \neq \varnothing$, then $\bar{x} \in \mathbb{R}^m$ solves (P) if and only if

$$(\partial f(\bar{x})) \cap (-N_C(\bar{x})) \neq \varnothing.$$

We now explore weaker results in the absence of convexity.

Theorem 3.1.1
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and $g : \mathbb{R}^m \to (-\infty, \infty]$ convex, l.s.c., proper with $\operatorname{dom} g \subseteq \operatorname{int}(\operatorname{dom} f)$. Consider the problem

$$\min_{x \in \mathbb{R}^m} f(x) + g(x). \qquad (P)$$

(i) If f is differentiable at $x^* \in \operatorname{dom} g$ and x^* is a local minima of (P), then $-\boldsymbol{\nabla} f(x^*) \in \partial g(x^*)$

(ii) If f is convex and differentiable at $x^* \in \operatorname{dom} g$ then x^* is a global minimizer of (P) if and only if $-\boldsymbol{\nabla} f(x^*) \in \partial g(x^*)$

Proof (i)
Let $y \in \operatorname{dom} g$. Since g is convex, we know that $\operatorname{dom} g$ is convex. Hence for any $\lambda \in (0, 1)$,

$$\begin{aligned} x^* + \lambda(y - x^*) &= (1 - \lambda)x^* + \lambda y \\ &=: x_\lambda \\ &\in \operatorname{dom} g. \end{aligned}$$

Hence for sufficiently small λ,

$$\begin{aligned} f(x_\lambda) + g(x_\lambda) &\geq f(x^*) + g(x^*) \\ f(x_\lambda) + (1 - \lambda)g(x^*) + \lambda g(y) &\geq f(x^*) + g(x^*) \\ \lambda g(x^*) - \lambda g(y) &\leq f(x_\lambda) - f(x^*) \\ g(x^*) - g(y) &\leq \frac{f(x_\lambda) - f(x^*)}{\lambda} \\ &\to f'(x^*; y - x^*) && \lambda \to 0^+ \\ &= \langle \boldsymbol{\nabla} f(x^*), y - x^* \rangle. \end{aligned}$$

In other words, for all $y \in \operatorname{dom} g$,

$$\begin{gathered} g(y) \geq g(x^*) + \langle \boldsymbol{\nabla} f(x^*), y - x^* \rangle \\ \implies \\ -\boldsymbol{\nabla} f(x^*) \in \partial g(x^*) \end{gathered}$$

Proof (ii)
Suppose that f is convex and observe that (i) proves the forward direction.

Now suppose $-\nabla f(x^*) \in \partial g(x^*)$. By definition, for each $y \in \operatorname{dom} g$,

$$g(y) \geq g(x^*) + \langle -\nabla f(x^*), y - x \rangle.$$

Moreover, since f is differentiable at x^* one of our characterizations of the convexity of f is that for any $y \in \operatorname{dom} g \subseteq \operatorname{int} \operatorname{dom} f$,

$$f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle.$$

Adding the inequalities yield that for all $y \in \operatorname{dom} g$,

$$f(y) + g(y) \geq f(x^*) + g(x^*)$$

and x^* solves (P).

3.1.1 The Karush-Kuhn-Tucker Conditions

In the following, we assume that

$$f, g_1, \dots, g_n : \mathbb{R}^m \to \mathbb{R}$$

are of full domain.

Consider the problem

$$\begin{aligned} \min f(x) && (P) \\ g_i(x) \leq && \forall i \in [n] \end{aligned}$$

We assume that (P) has at least one solution and that

$$\mu := \min\{f(x) : \forall i \in I, f(x) \leq 0\} \in \mathbb{R}$$

is the optimal value. Put

$$F(x) := \max\{\underbrace{f(x) - \mu}_{=:g_0(x)}, g_1(x), \dots, g_n(x)\}.$$

Lemma 3.1.2
For all $x \in \mathbb{R}^m$, $F(x) \geq 0$. Moreover, the solution of (P) are precisely the minimizers of

$$F := \{x : F(x) = 0\}.$$

Proof
Let $x \in \mathbb{R}^n$.

Case Ia: x is infeasible Then there is some $j \in [n]$ such that $g_j(x) > 0$. Hence $F(x) \geq g_i(x) > 0$.

Case Ib: x is not optimal Then $g_i(x) \leq 0$ but $f(x) > \mu$. Thus $F(x) \geq g_0(x) > 0$.

Case II: x solves (P) Then x is feasible and $f(x) = \mu$. Hence $F(x) = 0$.

Proposition 3.1.3 (Max Rule for Subdifferential Calculus)
Let $g_1, \dots, g_n : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Define

$$\begin{aligned} g(x) &= \max\{g_i(x), \dots, g_n(x)\} \\ A(x) &= \{i \in [n] : g_i(x) = g(x)\}. \end{aligned}$$

Now, let

$$x \in \bigcap_{n=1}^{n} \operatorname{int} \operatorname{dom} g_i.$$

We have

$$\partial g(x) = \operatorname{conv}\left(\bigcup_{i \in A(x)} \partial g_i(x) \right).$$

Theorem 3.1.4 (Fritz-John Optimality Conditions)
Suppose that $f, g_1, \dots, g_n$ are convex and x^* solves (P). There exists

$$\alpha_0, \dots, \alpha \geq 0$$

not all 0 for which

$$0 \in \alpha_0 \partial f(x^*) + \sum_{i=1}^{n} \alpha_i \partial g_i(x^*)$$

$$\alpha_i g_i(x^*) = 0 \qquad \forall i \in [n]$$

(complementary slackness)

Proof
Recall that $F(x) := \max\{f(x) - \mu, g_i(x), \dots, g_n(x)\}$. By the previous lemma,

$$F(x^*) = 0 = \min F(\mathbb{R}^n).$$

Hence

$$0 \in \partial F(x^*) = \operatorname{conv}_{i \in A(x^*)} \partial g_i(x^*).$$

where $A(x^*) := \{0 \leq i \leq n : g_i(x^*) = 0\}$.

Note that $0 \in \partial f(x^*)$ since $f_0(x^*) = f(x^*) - \mu = 0$. So

$$0 \in \partial g_0 = \partial f.$$

By our work with convex hulls, there is some $\alpha_0, \dots, \alpha_n$ such that $\sum_{i \in A(x^*)} \alpha_i = 1$ (so $\alpha_j = 0$ if $j \notin A(x^*)$) and that

$$\begin{aligned} 0 &\in \sum_{i \in A(x^*)} \alpha_i \partial g_i(x^*) \\ &= \alpha_0 \partial g_0(x^*) + \sum_{i \in A(x^*) \setminus \{0\}} \alpha_i \partial g_i(x^*) \\ &= \alpha_0 \partial g_0(x^*) + \sum_{i=1}^{n} \alpha_i \partial g_i(x^*). \end{aligned}$$

Now to see complementary slackness: If $i \in A(x^*) \cap [n]$, then $g_i(x^*) = 0$. Else if $i \in [n] \setminus A^*(x)$, then $\alpha_i = 0$. In all cases,

$$\alpha_i g_i(x^*) = 0$$

for all $i \in [n]$.

Theorem 3.1.5 (Karush-Kuhn-Tucker; Necessary Conditions)
Suppose $f, g_1, \dots, g_n$ are convex, and x^* solves (P). Suppose that Slater's condition holds, ie there is some $s \in \mathbb{R}^m$ such that for all $i \in [n]$,

$$g_i(s) < 0.$$

Then there exists $\lambda_1, \dots, \lambda_m \geq 0$ such that the KKT conditions hold: (stationarity condition)

$$0 \in \partial f(x^*) + \sum_{i \in I} \lambda_i \partial g_i(x^*)$$

and (complementary slackness condition) for each $i \in [n]$,

$$\lambda_i g_i(x^*) = 0.$$

Proof
By the Fritz-John necessary conditions, there are $\alpha_0, \alpha_1, \dots, \alpha_n \geq 0$ not all 0 such that

$$0 \in \alpha_0 \partial f(x^*) + \sum_{i=1}^n \alpha_i \partial g_i(x^*).$$

and for all $i \in [n]$,

$$\alpha_i g_i(x^*) = 0.$$

We claim that $\alpha_0 \neq 0$. Then it is necessary that

$$0 \in \partial f(x^*) + \sum_{i=1}^n \frac{\alpha_i}{\alpha_0} \partial g_i(x^*).$$

Suppose towards a contradiction that $\alpha_0 = 0$. There exist $y_i \in \partial g_i(x^*)$ such that

$$\sum_{i=1}^n \alpha_i y_i = 0.$$

By the definition of the subgradient, for all $y \in \mathbb{R}^m$,

$$g_i(x^*) + \langle y_i, y - x^* \rangle \leq g_i(y).$$

Thus for each $i \in [n]$,

$$g_i(x^*) + \langle y_i, s - x^* \rangle \leq g_i(s).$$

Multiplying each inequality by α_i and adding them yields

$$\begin{aligned} 0 &= \sum_{i=1}^n \alpha_i g_i(x^*) + \left\langle \sum_{i=1}^n \alpha_i y_i, s - x^* \right\rangle \\ &\leq \sum_{i=1}^n \alpha_i g_i(s) \\ &< 0 \end{aligned}$$

which is absurd.

By contradiction, $\alpha_0 > 0$ and we are done.

Theorem 3.1.6 (Karush-Kuhn-Tucker; Sufficient Conditions)
Suppose $f, g_1, \dots, g_n$ are convex and $x^* \in \mathbb{R}^m$ satisfies

$$\begin{aligned} \forall i \in [n], g_i(x^*) &\le 0 && \text{primal feasibility} \\ \forall i \in [n], \lambda_i &\ge 0 && \text{dual feasibility} \\ \partial f(x^*) + \sum_{i=1}^n \lambda_i \partial g_i(x^*) &\ni 0 && \text{stationarity} \\ \forall i \in [n], \lambda_i g_i(x^*) &= 0 && \text{complementary slackness} \end{aligned}$$

Then x^* solves (P).

Proof
Define

$$h(x) := f(x) + \sum_{i=1}^n \lambda_i g_i(x).$$

Then h is convex since non-negative multiplication preserves convexity.

Apply the sum rule to obtain that

$$\partial g(x) = \partial f(x) + \sum_{i=1}^n \lambda_i \partial g_i(x).$$

By assumption,

$$0 \in \partial h(x^*) = \partial f(x^*) + \sum_{i=1}^n \lambda_i \partial g_i(x^*).$$

Thus by Fermat's theorem, x^* is a global minimizer of H.

Let x be feasible for (P). Then

$$\begin{aligned} f(x^*) &= f(x^*) + \sum_{i=1}^n \lambda_i g_i(x^*) \\ &= h(x^*) \\ &\le h(x) \\ &= f(x) + \sum_{i=1}^n \lambda_i g_i(x) \\ &\le f(x). \end{aligned}$$

3.2 Gradient Descent

Consider the problem

$$\min_{x \in \mathbb{R}^m} f(x) \tag{P}$$

Definition 3.2.1 (Descent Direction)
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be proper and let $x \in \operatorname{int} \operatorname{dom} f$. $d \in \mathbb{R}^m \setminus \{0\}$ is a descent direction of f at x if the directional derivative satisfies

$$f'(x; d) < 0.$$

Remark that if $0 \neq \boldsymbol{\nabla} f(x)$ exists, then $\nabla f(x)$ is a descent direction. Indeed,

$$f'(x; -\nabla f(x)) = -\|\nabla f(x)\|^2 < 0.$$

Also remark that for convex f and $x \in \operatorname{dom} f$,

$$f'(x, d) = \lim_{\lambda \to 0^+} \frac{f(x + \lambda d) - f(x)}{\lambda}.$$

Thus $f(x, d) < 0$ implies that there is some ϵ such that $\lambda \in (0, \epsilon)$ implies that

$$\frac{f(x + \lambda d) - f(x)}{\lambda} < 0 \iff f(x + \lambda d) < f(x).$$

The gradient/steepest descent method consists of the following:

1. Initialize $x_0 \in \mathbb{R}^m$.
2. For each $n \in \mathbb{N}$:
   (a) Pick $t_n \in \operatorname{argmin}_{t \geq 0} f(x_n - t\boldsymbol{\nabla} f(x_n))$.
   (b) Update $x_{n+1} := x_n - t_n \boldsymbol{\nabla} f(x_n)$

Theorem 3.2.1 (Peressini, Sullivan, Uhl)
If f is strictly convex and coercive, then x_n converges to the unique minimizer of f.

In the lack of smoothness, a lot of pathologies happen.

Example 3.2.2 (L. Vandenberghe)
Negative subgradients are NOT necessarily descent directions. Consider $f : \mathbb{R}^2 \to \mathbb{R}_+$ given by

$$(x_1, x_2) \mapsto |x_1| + 2|x_2|.$$

Then f is convex as it is a direct sum of convex functions.

Since f has full domain and is continuous,

$$\partial f(1,0) = \{1\} \times [-2, 2].$$

Take $d := (-1, -2) \in -\partial f(1,0)$.

d is NOT a descent direction. Moreover,

$$f(1,0) = 1 < f[(1,0) + t(-1,-2)]$$

for all $t > 0$.

Example 3.2.3 (Wolfe)
Let $\gamma > 1$. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$(x_1, x_2) \mapsto \begin{cases} \sqrt{x_1^2 + \gamma x_2^2}, & |x_2| \leq x_1 \\ \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}}, & \text{else} \end{cases}$$

Observe that $\operatorname{argmin}_{x \in \mathbb{R}^m} f = \varnothing$. One can show that $f = \sigma_C$where

$$C = \left\{ x \in \mathbb{R}^2 : x_2^2 + \frac{x_2^2}{\gamma} \leq 1, x_2 \geq \frac{1}{\sqrt{1+\gamma}} \right\}.$$

Thus f is convex. Moreover, f is differentiable on

$$D := \mathbb{R}^2 \setminus ((-\infty, 0] \times \{0\}).$$

Let $x_0 := (\gamma, 1) \in D$.

The steepest descent method will generate a equence

$$x_n := \left(\gamma \left(\frac{\gamma - 1}{\gamma + 1} \right)^n, \left(-\frac{\gamma - 1}{\gamma + 1} \right)^n \right) \to (0,0)$$

which is not a minimizer of f!

3.3 Projected Subgradient Method

Consider

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

where $f : \mathbb{R}^m \to (-\infty, \infty]$ is convex, l.s.c., and proper, $\varnothing \neq C \subseteq \operatorname{int}\operatorname{dom} f$ is convex and closed.

Suppose

$$S := \operatorname{argmin}_{x\in C} f(x) \neq \varnothing$$
$$\mu := \min_{x\in C} f(x).$$

Moreover, there is some $L > 0$ such that

$$\sup\|\partial f(C)\| \leq L < \infty.$$

In other words, for all $c \in C$ and $u \in \partial f(c)$, $\|u\| \leq L$.

1) Get $x_0 \in C$.
2) Given x_n, pick a stepsize $t_n > 0$ and $f'(x_n) \in \partial f(x_n)$
3) Update $x_{n+1} := P_C(x_n - t_n f'(x_n))$.

Recall that $C \subseteq \operatorname{int}\operatorname{dom} f$, hence each $x_n \in \operatorname{int}\operatorname{dom} f$ and $\partial f(x_n) \neq \varnothing$. Thus the algorithm is well-defined.

Lemma 3.3.1
Let $s \in S := \operatorname{argmin}_{x\in C} f(x)$. Then

$$\|x_{n+1} - s\|^2 \leq \|x_n - s\|^2 - 2t_n(f(x_n) - \mu) + t_n^2\|f'(x_n)\|^2.$$

Observe that $S \subseteq C$.

Proof
We have

$$\begin{aligned}\|x_{n+1} - s\|^2 &= \|P_C(x_n - t_n f'(x_n)) - P_C(s)\|^2 \\ &\leq \|x_n - t_n f'(x_n) - s\|^2 \\ &= \|x_n - s\|^2 + t_n^2\|f'(x_n)\|^2 - 2t_n\langle x_n - s, f'(x_n)\rangle.\end{aligned}$$

It suffices to show that

$$
\begin{aligned}
2t_n\langle x_n - s, f'(x_n)\rangle &\le -2t_n(f(x_n)-\mu) \\
\langle x_n - s, f'(x_n)\rangle &\ge f(x_n)-\mu \\
\langle x_n - s, f'(x_n)\rangle &\ge f(x_n)-f(x)
\end{aligned}
$$

which holds by the subgradient inequality.

What is a good step size? We wish to minimize the upper bound derived in the previous lemma.

$$
\begin{aligned}
0 &= \frac{d}{dt_n}(-2t_n(f(x_n)-\mu) + t_n^2\|f'(x_n)\|^2) \\
&= -2(f(x_n)-\mu) + 2t_n\|f'(x_n)\|^2.
\end{aligned}
$$

If x_n is not a global minimizer, then $0 \notin \partial f(x_n)$ and $f'(x_n) \neq 0$. Hence we can take

$$t_n := \frac{f(x_n)-\mu}{\|f'(x_n)\|^2}.$$

Definition 3.3.1 (Polyak's Rule)
The projected subgradient method with step size

$$t_n := \frac{f(x_n)-\mu}{\|f'(x_n)\|^2}.$$

Theorem 3.3.2
We have

(i) For all $s \in S, n \in \mathbb{N}$, $\|x_{n+1}-s\| \le \|x_n - s\|$, ie $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to S

(ii) $f(x_n) \to \mu$

(iii) $\mu_n - \mu \le \frac{L\cdot d_S(x_0)}{\sqrt{n+1}} \in O\left(\frac{1}{\sqrt{n}}\right)$, where $\mu_n := \min_{0\le k\le n} f(x_k)$

(iv) For each $\epsilon > 0$, if $n \ge \frac{L^2 d_S^2(x_0)}{\epsilon^2} - 1$, then $\mu_n \le \mu + \epsilon$

Proof (i)
Let $s \in S, n \in \mathbb{N}$ By computation

$$\begin{aligned}
\|x_{n+1} - s\|^2 &\leq \|x_n - s\|^2 - 2t_n(f(x_n) - \mu) + t_n^2\|f'(x_n)\|^2 \\
&= \|x_n - s\|^2 - 2\frac{f(x_n) - \mu}{\|f'(x_n)\|^2}(f(x_n) - \mu) + \left(\frac{f(x_n) - \mu}{\|f'(x_n)\|^2}\right)^2 \|f'(x_n)\|^2 \\
&= \|x_n - s\|^2 - \frac{(f(x_n) - \mu)^2}{\|f'(x_n)\|^2} \\
&\leq \|x_n - s\|^2 - \frac{(f(x_n) - \mu)^2}{L^2} \\
&\leq \|x_n - s\|^2.
\end{aligned}$$

Proof (ii)
From our work in (i): for all $k \in \mathbb{N}$,

$$\frac{(f(x_k) - \mu)^2}{L^2} \leq \|x_k - s\|^2 - \|x_{k+1} - s\|.$$

Summing the above inequalities over $k = 0, \dots, n$ yields

$$\begin{aligned}
\frac{1}{L^2}\sum_{k=0}^{n}(f(x_k) - \mu^2) &\leq \|x_0 - s\|^2 - \|x_{n+1} - s\|^2 \\
&\leq \|x_0 - s\|^2.
\end{aligned}$$

Letting $n \to \infty$,

$$0 \leq \sum_{k=0}^{\infty}(f(x_k) - \mu)^2 \leq L^2\|x_0 - s\|^2 < \infty$$

and it must be that $f(x_k) \to \mu$.

Proof (iii)
Recall that

$$\mu_n := \min_{0 \leq k \leq n} f(x_k).$$

Let $n \geq 0$. For each $0 \leq k \leq n$,

$$\begin{aligned}(\mu_n - \mu)^2 &\leq (f(x_k) - \mu)^2 \\ (n+1)\frac{(\mu_n - \mu)^2}{L^2} &\leq \frac{1}{L^2}\sum_{k=0}^{n}(f(x_k) - \mu)^2 \\ &\leq \|x_0 - s\|^2.\end{aligned}$$

Minimizing over $s \in S$, we get that

$$(n+1)\frac{(\mu_n - \mu)^2}{L^2} \leq d_S^2(x_0).$$

Proof (iv)
Suppose that

$$\begin{aligned}n &\geq \frac{L^2 d_S^2(x_0)}{\epsilon^2} - 1 \\ &\iff \\ \frac{d_S^2(x_0)L^2}{n+1} &\leq \epsilon^2.\end{aligned}$$

Apply (iii) yields

$$\begin{aligned}(\mu_n - \mu)^2 &\leq \frac{d_S^2(x_0)L^2}{n+1} \\ &\leq \epsilon^2 \\ \mu_n - \mu &\leq \epsilon.\end{aligned}$$

Recall that if $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to some $\varnothing \neq C \subseteq \mathbb{R}^m$, and every cluster point lies in C, then $x_n \to c \in C$.

Theorem 3.3.3 (Convergence of Projected Subgradient)
Suppose x_n is generated as in the projected subgradient method with Polyak's rule. Then $x_n \to s \in S$.

Proof
We have already shown that (x_n) is Fejér monotone with respect to S. Thus the sequence

is also bounded. Also, by the previous theorem,

$$f(x_n) \to \mu = \min_{x \in C} f(x).$$

By Bolzano-Weirestrass, there is some subsequence $x_{k_n} \to \bar{x} \in C$. Now,

$$\begin{aligned} \mu &= \min_{x \in C} f(x) \\ &\leq f(\bar{x}) \\ &\leq \liminf_n f(x_{k_n}) \\ &= \mu && f(x_n) \to \mu. \end{aligned}$$

Hence $\bar{x} \in S$. That is, all cluster points of $(x_n)_{n \in \mathbb{N}}$ lie in S.

It follows that $x_n \to \bar{x} \in S$ by the Fejér monotonicity theorem.

Example 3.3.4
Let $C \subseteq \mathbb{R}^m$ be convex, closed, and non-empty. Fix $x \in \mathbb{R}^m$.

$$\partial d_C(x) = \begin{cases} \frac{x - P_C(x)}{d_C(x)}, & x \notin C \\ N_C(x) \cap B(0; 1), & x \in C \end{cases}$$

Moreover, $\sup \|\partial d_C(x)\| \leq 1$.

Lemma 3.3.5
Let f be convex, l.s.c., and proper. Fix $\lambda > 0$. Then

$$\partial(\lambda f) = \lambda \partial f.$$

3.3.1 The Convex Feasibility Problem

Problem 1
Given k closed convex subsets $S_i \subseteq \mathbb{R}^m$ such that

$$S := \bigcap_{i=1}^{k} S_i \neq \varnothing,$$

find $x \in S$.

We take

$$f(x) := \max\{d_{S_i}(x) : i \in [k]\}.$$

The domain is $C := \mathbb{R}^m$. Observe that $f \geq 0$ with

$$\begin{aligned} f(x) = 0 &\iff \forall i, d_{S_i}(x) = 0 \\ &\iff \forall i, x \in S_i \\ &\iff x \in S. \end{aligned}$$

Recall that the max rule for subdifferentials implies that for all $x \notin S$,

$$\partial f(x) = \operatorname{conv}\{\partial d_{S_i}(x) : d_{S_i}(x) = f(x) > 0\}$$

Thus $\|\partial f(x)\| \leq 1$ as a convex combination preserves the norm bound.

Given x_n, pick an index $\bar{i}$ such that $d_{S_{\bar{i}}}(x_n) = f(x_n) > 0$. Set

$$f'(x_n) := \frac{x_n - P_{S_{\bar{i}}}(x_n)}{d_{S_{\bar{i}}}(x_n)}.$$

Since this is a unit vector, Polyak's step size simplifies to

$$t_n = d_{S_{\bar{i}}}(x_n).$$

The sequence converging to a member of S is thus

$$\begin{aligned} x_{n+1} &:= P_C(x_n - t_n f'(x_n)) \\ &= x_n - t_n f'(x_n) \\ &= x_n - d_{S_{\bar{i}}}(x_n) \frac{x_n - P_{S_{\bar{i}}}(x_n)}{d_{S_{\bar{i}}}(x_n)} \\ &= x_n - (x_n - P_{S_{\bar{i}}}(x_n)) \\ &= P_{S_{\bar{i}}}(x_n). \end{aligned}$$

By the convergence of the projected subgradient method, $x_n \to S$.

Note that in practice, it is possible that $\mu := \min_{x\in C} f(x)$ is NOT known to us. In this case, replace Polyak's stepsize by a sequence $(t_n)_{n\in\mathbb{N}}$ such that

$$\frac{\sum_{k=0}^n t_k^2}{\sum_{k=0}^n t_k} \to 0, n \to \infty.$$

For example, $t_k := \frac{1}{k+1}$. One can show that

$$\mu_n := \min_{k=0}^{n} f(x_k) \to \mu$$

as $n \to \infty$.

3.4 Proximal Gradient Method

Consider the problem

$$\begin{aligned} &\min F(x) := f(x) + g(x) \\ &x \in \mathbb{R}^m \end{aligned} \tag{P}$$

We shall assume that $S := \operatorname{argmin}_{x\in\mathbb{R}^m} F(x) \neq \varnothing$ and define

$$\mu := \min_{x\in\mathbb{R}^m} F(x).$$

f is "nice" in that it is convex, l.s.c., proper, and differentiable on $\operatorname{int}\operatorname{dom} f \neq \varnothing$. Moreover, $\boldsymbol{\nabla} f$ is L-Lipschitz on $\operatorname{int}\operatorname{dom} f$.

g is convex, l.s.c., and proper with $\operatorname{dom} g \subseteq \operatorname{int}\operatorname{dom} f$. In particular,

$$\begin{aligned} \varnothing &\neq \operatorname{ri}\operatorname{dom} g \\ &\subseteq \operatorname{dom} g \\ &\subseteq \operatorname{ri}\operatorname{dom} f \\ &= \operatorname{int}\operatorname{dom} f \\ &\Longrightarrow \\ \operatorname{ri}\operatorname{dom} g \cap \operatorname{ri}\operatorname{dom} f &= \operatorname{ri}\operatorname{dom} g \\ &\neq \varnothing. \end{aligned}$$

Example 3.4.1
We can model contrained optimization functions as

$$\min_{x \in \mathbb{R}^m} f(x) + \delta_C(x)$$

where $\varnothing \neq C \subseteq \mathbb{R}^m$ is convex and closed.

Let $x \in \operatorname{int}\operatorname{dom} f \supseteq \operatorname{dom} g$. Update via

$$\begin{aligned} x_+ &:= \operatorname{Prox}_{\frac{1}{L}g}(x - \frac{1}{L}\boldsymbol{\nabla} f(x)) \\ &= \operatorname{argmin}_{y \in \mathbb{R}^m} \frac{1}{L}g(y) + \frac{1}{2}\left\| y - \left(\frac{1}{L}\boldsymbol{\nabla} f(x)\right)\right\|^2 \\ &\in \operatorname{dom} g \\ &\subseteq \operatorname{int}\operatorname{dom} f \\ &= \operatorname{dom} \boldsymbol{\nabla} f. \end{aligned}$$

Let the update operator be denoted

$$T := \operatorname{Prox}_{\frac{1}{L}g}(\operatorname{Id} - \frac{1}{L}\boldsymbol{\nabla} f).$$

Theorem 3.4.2
Let $x \in \mathbb{R}^m$. Then

$$\begin{aligned} &x \in S \\ &\quad = \operatorname{argmin}_{x \in \mathbb{R}^m} F \\ &\quad = \operatorname{argmin}_{x \in \mathbb{R}^m} (f + g) \\ &\quad \Longleftrightarrow \\ &x = Tx \\ &\quad \Longleftrightarrow \\ &x \in \operatorname{Fix} T. \end{aligned}$$

Proof
By Fermat's theorem,

$$\begin{aligned}
x \in S &\iff 0 \in \partial(f+g)(x) = \nabla f(x) + \partial g(x) \\
&\iff -\nabla f(x) \in \partial g(x) \\
&\iff x - \frac{1}{L}\nabla f(x) \in x + \frac{1}{L}\partial g(x) = \left(\operatorname{Id} + \partial\left(\frac{1}{L}g\right)\right)(x) \\
&\iff x \in \left(\operatorname{Id} + \partial\left(\frac{1}{L}g\right)\right)^{-1}\left(x - \frac{1}{L}\nabla f(x)\right) \\
&\iff x = \operatorname{Prox}_{\frac{1}{L}g}\left(\operatorname{Id} - \frac{1}{L}\nabla f\right)(x) = Tx.
\end{aligned}$$

Proposition 3.4.3
Let $f : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Fix $\beta > 0$. Then f is β-strongly convex if and only if for all $x \in \operatorname{dom} \partial f, u \in \partial f(x)$,

$$f(y) \geq f(x) + \langle u, y - x\rangle + \frac{\beta}{2}\|y - x\|^2.$$

3.4.1 Proximal-Gradient Inequality

Proposition 3.4.4
Let $x \in \mathbb{R}^m, y_+ \in \operatorname{int}\operatorname{dom} f$, and

$$y_+ := \operatorname{Prox}_{\frac{1}{L}g}(y - \nabla f(y)) = Ty.$$

Then

$$F(x) - F(y_+) \geq \frac{L}{2}\|x - y_+\|^2 - \frac{L}{2}\|x - y\|^2 + D_f(x, y).$$

where

$$D_f(x, y) := f(x) - f(y) - \langle \nabla f(y), x - y\rangle.$$

D_f is known as the *Bregman distance*.

Proof
Define

$$h(z) := f(y) + \langle \nabla f(y), z - y\rangle + g(z) + \frac{L}{2}\|z - y\|^2.$$

Then h is L-strongly convex.

We claim that y_+ is the unique minimizer of h. Indeed, for $z \in \mathbb{R}^m$,

$$\begin{aligned}
z \in \operatorname{argmin} h &\iff 0 \in \partial\left(f(y) + \langle \nabla f(y), z - y\rangle + g(z) + \frac{L}{2}\|z-y\|^2\right)\\
&\iff 0 \in \partial\left(\langle \nabla f(y), z - y\rangle + g(z) + \frac{L}{2}\|z-y\|^2\right)\\
&\iff 0 \in \nabla f(y) + \partial g(z) + L(z-y)\\
&\iff 0 \in \frac{1}{L}\nabla f(y) + \partial\left(\frac{1}{L}g\right)(z) + (z-y)\\
&\iff y - \frac{1}{L}\nabla f(y) \in z + \partial\left(\frac{1}{L}g\right)(z)\\
&\iff y - \frac{1}{L}\nabla f(y) \in \left(\operatorname{Id} + \partial\left(\frac{1}{L}g\right)\right)(z)\\
&\iff z \in \left(\operatorname{Id} + \partial\left(\frac{1}{L}g\right)\right)^{-1}\left(y - \frac{1}{L}\nabla f(y)\right)\\
&\iff z = \operatorname{Prox}_{\frac{1}{L}g}\left(y - \frac{1}{L}\nabla f(y)\right)\\
&\iff z = Ty = y_+.
\end{aligned}$$

Applying the previous proposition yields that

$$\begin{aligned}
h(x) &\geq h(y_+) + \langle 0, x - y_+\rangle + \frac{L}{2}\|x - y_+\|^2\\
&= h(y_+) + \frac{L}{2}\|x-y_+\|^2\\
h(x) - h(y_+) &\geq \frac{L}{2}\|x - y_+\|^2.
\end{aligned}$$

Moreover, by the descent lemma,

$$f(y_+) \leq f(y) + \langle \nabla f(y), y_+ - y\rangle + \frac{L}{2}\|y_+ - y\|^2.$$

Hence

$$\begin{aligned}
h(y_+) &:= f(y) + \langle \nabla f(y), y_+ - y\rangle + g(y_+) + \frac{L}{2}\|y_+ - y\|^2\\
&\geq f(y_+) + g(y_+)\\
&= F(y_+).
\end{aligned}$$

.

Combining with our work above,

$$\begin{aligned}
h(x) - F(y_+) &\geq h(x) - h(y_+) \\
&\geq \frac{L}{2}\|x - y_+\|^2 \\
f(y) + \langle \nabla f(y), x - y\rangle + g(x) + \frac{L}{2}\|x-y\|^2 - F(y_+) &\geq \frac{L}{2}\|x - y_+\|^2 \\
f(x) + g(x) - F(y_+) &\geq \frac{L}{2}\|x - y_+\|^2 - \frac{L}{2}\|x-y\|^2 + D_f(x, y).
\end{aligned}$$

Lemma 3.4.5 (Sufficient Decrease)
We have

$$F(y_+) \leq F(y) - \frac{L}{2}\|y - y_+\|^2.$$

Proof
Recall that

$$\begin{aligned}
F(y) - F(y_+) &\geq \frac{L}{2}\|y - y_+\|^2 - \frac{L}{2}\|y - y\|^2 + D_f(y, y) \\
F(y) - F(y_+) &\geq \frac{L}{2}\|y - y_+\|^2 && f \text{ is convex} \\
F(y_+) &\leq F(y) - \frac{L}{2}\|y - y_+\|^2.
\end{aligned}$$

3.4.2 The Algorithm

Given $x_0 \in \operatorname{int}\operatorname{dom} f$, update via

$$x_{n+1} := Tx_n = \operatorname{Prox}_{\frac{1}{L}g}\left(x_n - \frac{1}{L}\nabla f(x_n)\right).$$

Theorem 3.4.6 (Rate of Convergence)
The following hold:

(i) For all $s \in S, n \in \mathbb{N}$, $\|x_{n+1} - s\| \leq \|x_n - s\|$ (ie x_n is Fejér monotone with respect to S).

(ii) $(F(x_n))_{n\in\mathbb{N}}$ satisfies $0 \leq F(x_n) - \mu \leq \frac{Ld_S^2(x_0)}{2n} \in O\left(\frac{1}{n}\right)$. Hence $F(x_n) \to \mu$.

Proof
(i): Recall the previous proposition that

$$
\begin{aligned}
0 &\geq F(s) - F(x_{k+1}) && F(x) = \mu \\
&\geq \frac{L}{2}\|s - x_{k+1}\|^2 - \frac{L}{2}\|s - x_k\|^2.
\end{aligned}
$$

Thus (x_n) is Fejér monotone with respect to S.

(ii): Multiplying this inequality by $\frac{2}{L}$ and adding the resulting inequalities from $k = 0, \dots, n-1$ and telescoping yields

$$
\begin{aligned}
\frac{2}{L}\left(\sum_{k=0}^{n-1}(\mu - F(x_{k+1}))\right) &\geq \|s - x_k\|^2 - \|s - x_0\|^2 \\
&\geq -\|s - x_0\|^2.
\end{aligned}
$$

In particular, by setting $s := P_S(x_0) \in S$, we obtain

$$
\begin{aligned}
d_S^2(x_0) &= \|P_S(x_0) - x_0\|^2 \\
&\geq \frac{2}{L}\sum_{k=0}^{n-1}(F(x_{k+1}) - \mu) \\
&\geq \frac{2}{L}\sum_{k=0}^{n-1}(F(x_n) - \mu) \\
&= \frac{2}{L}n(F(x_n) - \mu).
\end{aligned}
$$

Equivalently,

$$
\begin{aligned}
0 &\leq F(x_n) - \mu \\
&\leq \frac{L d_S^2(x_0)}{2n}
\end{aligned}
$$

and $F(x_n) \to \mu$.

Theorem 3.4.7 (Convergence of Proximal Gradient Method)
x_n converges to some solution in

$$S := \operatorname{argmin}_{x \in \mathbb{R}^m} F(x).$$

Proof
By the previous theorem we know that (x_n) is Fejér monotone with respect to S. Thus it suffices to show that every cluster point of (x_n) lies in S.

Suppose $\bar{x}$ is a cluster point of (x_n), say $x_{k_n} \to \bar{x}$. We argue that $F(\bar{x}) = \mu$. Indeed,

$$\begin{aligned} \mu &\leq F(\bar{x}) \\ &\leq \liminf_n F(x_{k_n}) \\ &= \mu \end{aligned}$$

Hence $F(\bar{x}) = \mu$ and $\bar{x} \in S$.

Proposition 3.4.8
The following hold:
(i) $\frac{1}{L}\boldsymbol{\nabla} f$ is f.n.e.
(ii) $\operatorname{Id} - \frac{1}{L}\boldsymbol{\nabla} f$ is f.n.e.
(iii) $T = \operatorname{Prox}_{\frac{1}{L}g}(\operatorname{Id} - \boldsymbol{\nabla} f)$ is $\frac{2}{3}$-averaged.

Proof
(i), (ii): Recall for real-valued, convex, differentiable functions with L-Lipschitz gradient,

$$\langle \boldsymbol{\nabla} f(x) - \boldsymbol{\nabla} f(y), x - y \rangle \geq \frac{1}{L}\|\boldsymbol{\nabla} f(x) - \boldsymbol{\nabla} f(y)\|^2$$
$$\left\langle \frac{1}{L}\boldsymbol{\nabla} f(x) - \frac{1}{L}\boldsymbol{\nabla} f(y), x - y \right\rangle \geq \left\| \frac{1}{L}\boldsymbol{\nabla} f(x) - \frac{1}{L}\boldsymbol{\nabla} f(y) \right\|^2.$$

The result follows then from the two equivalent characterizations of f.n.e.: $\operatorname{Id} - T$ is non-expansive and

$$\langle Tx - Ty, Tx - Ty \rangle \geq \|Tx - Ty\|^2.$$

(iii): Recall that $\operatorname{Prox}_{\frac{1}{L}g}$ is f.n.e. Hence, $\operatorname{Prox}_{\frac{1}{L}g}$ and $\operatorname{Id} - \frac{1}{L}\boldsymbol{\nabla} f$ are both $\frac{1}{2}$-averaged. Consequently, the composition

$$\operatorname{Prox}_{\frac{1}{L}g}\left(\operatorname{Id} - \frac{1}{L}\boldsymbol{\nabla} f\right)$$

is averaged with constant $\frac{2}{3}$.

Theorem 3.4.9
The PGM iteration satisifes

$$\|x_{n+1} - x_n\| \le \frac{\sqrt{2} d_S(x_0)}{\sqrt{n}} \in O\left(\frac{1}{\sqrt{n}}\right).$$

Proof
Using the previous remark, we have that for all x, y,

$$\frac{1}{2}\|(\operatorname{Id} - T)x - (\operatorname{Id} - T)y\|^2 < \|x - y\|^2 - \|Tx - Ty\|^2.$$

Let $x \in S$ and observe that $s \in \operatorname{Fix} s$ by a previous theorem. Applying the above inequality with $x = x_k, y = s$ yields

$$\begin{aligned}\frac{1}{2}\|(\operatorname{Id} - T)x_k - (\operatorname{Id} - T)s\| &\le \|x_k - s\|^2 - \|Tx_k - Ts\|^2 \\ \frac{1}{2}\|x_k - x_{k+1}\|^2 &\le \|x_k - s\|^2 - \|x_{k+1} - s\|^2.\end{aligned}$$

Now, T is $\frac{2}{3}$ averaged and thus nonexpansive. Therefore,

$$\begin{aligned}\|x_k - x_{k+1}\| = \|Tx_{k-1} - Tx_k\| && \le \|x_{k-1} - x_k\| \\ \le \dots \\ \le \|x_0 - x_1\|.\end{aligned}$$

Summing over $k = 0 \dots, n-1$ yields

$$\begin{aligned}\|x_0 - s\|^2 - \|x_n - s\|^2 &\ge \frac{1}{2}\sum_{k=0}^{n-1}\|x_k - x_{k+1}\|^2 \\ &\ge \frac{1}{2}n\|x_{n-1} - x_n\|^2.\end{aligned}$$

Specifically, for $x := P_S(x_0)$,

$$\begin{aligned}\frac{1}{2}n\|x_{n-1} - x_n\|^2 &\le d_S^2(x_0) \\ \|x_{n-1} - x_n\| &\le \frac{\sqrt{2}}{\sqrt{n}} d_S(x_0).\end{aligned}$$

Corollary 3.4.9.1 (Classical Proximal Point Algorithm)
Let $g : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. Fix $c > 0$. Consider the problem

$$\min_{x \in \mathbb{R}^m} g(x) \qquad (P)$$

Assume that $S := \operatorname{argmin}_{x\in\mathbb{R}^m} g(x) \leq \varnothing$.
Let $x_0 \in \mathbb{R}^m$ and update via

$$x_{n+1} := \operatorname{Prox}_{cg} x_n.$$

Then

(i) $g(x_n) \downarrow \mu := \min g(\mathbb{R}^m)$

(ii) $0 \leq g(x_n) - \mu \leq \frac{d_S^2(x_0)}{2cn}$

(iii) x_n converges to a point within S

(iv) $\|x_{n-1} - x_n\| \leq \frac{\sqrt{2} d_S(x_0)}{\sqrt{n}}$

Proof
Set $f \equiv 0$ and observe that $\nabla f \equiv 0$ and ∇f is L-Lipchitz for any $L > 0$. Specifically, for $L := \frac{1}{c} > 0$.

We can thus write (P) as

$$\min_{x \in \mathbb{R}^m} f(x) + g(x) \qquad (P)$$

Now, $S = \operatorname{argmin} f + g = \operatorname{argmin} g$. Moreover, $\nabla f \equiv 0 \implies \operatorname{Id} - \frac{1}{L}\nabla f = \operatorname{Id}$.

Hence

$$\begin{aligned} T &:= \operatorname{Prox}_{\frac{1}{L}g}(\operatorname{Id} - \frac{1}{L}\nabla f) \\ &= \operatorname{Prox}_{cg} \end{aligned}$$

and we are done by the previous results.

3.5 Fast Iterative Shrinkage Thresholding

Consider the following problem

$$\min_{x \in \mathbb{R}^m} F(x) := f(x) + g(x) \tag{P}$$

We assume (P) has solutions so that

$$S := \operatorname{argmin}_{x\in\mathbb{R}^m} F(x) \neq \varnothing$$

and write $\mu := \min_{x\in\mathbb{R}^m} F(x)$.

We assume f is convex, l.s.c., and proper, as well as being differnentiable on $\mathbb{R}^m$. Moreover, $\boldsymbol{\nabla} f$ is L-Lipschitz on $\mathbb{R}^m$.

We also assume that g is convex, l.s.c., and proper.

3.5.1 The Algorithm

Initially, set $x_0 \in \mathbb{R}^m, t_0 = 1, y_0 = x_0$. We update via

$$\begin{aligned}
t_{n+1} &= \frac{1+\sqrt{1+4t_n^2}}{2} \\
x_{n+1} &= \operatorname{Prox}_{\frac{1}{L}g}\left(\operatorname{Id} - \frac{1}{L}\boldsymbol{\nabla} f\right)(y_n) = Ty_n \\
y_{n+1} &= x_{n+1} + \frac{t_n - 1}{t_{n+1}}(x_{n+1} - x_n) \\
&= \left(1 - \frac{1-t_n}{t_{n+1}}\right) x_{n+1} + \frac{1-t_n}{t_{n+1}} x_n \\
&\in \operatorname{aff}\{x_n, x_{n+1}\}
\end{aligned}$$

Observe that

$$t_{n+1}^2 - t_{n+1} = t_n^2.$$

3.5.2 Correctness

Proposition 3.5.1
The sequence $(t_n)_{n\in\mathbb{N}}$ satisfies
$$t_n \geq \frac{n+2}{2} \geq 1.$$

Proof
Induction.

Theorem 3.5.2 (Quadratic Converge for FISTA)
The sequence (x_n) satisfies
$$\begin{aligned} 0 &\leq F(x_n) - \mu \\ &\leq \frac{2Ld_S^2(x_0)}{(n+1)^2} \\ &\in O\left(\frac{1}{n^2}\right). \end{aligned}$$

Notice that this converges significantly faster than $O\left(\frac{1}{n}\right)$ for PGM.

Proof
Set $s := P_S(x_0)$. By the convexity of F,
$$F\left(\frac{1}{t_n}s + \left(1 - \frac{1}{t_n}\right)x_n\right) \leq \frac{1}{t_n}F(s) + \left(1 - \frac{1}{t_n}\right)F(x_n)$$
For each $n \in \mathbb{N}$, set
$$s_n := F(x_n) - \mu \geq 0.$$

By computation,
$$\left(1 - \frac{1}{t_n}\right)s_n - s_{n+1} \geq F\left(\frac{1}{t_n}s + \left(1 - \frac{1}{t_n}\right)x_n\right) - F(x_{n+1}).$$

Now, applying the proximal gradient inequality with
$$\begin{aligned} x &= \frac{1}{x_n}s + \left(1 - \frac{1}{t_n}\right)x_n \\ y &= y_n \\ y_+ &= Ty_n = x_{n+1} \end{aligned}$$

yields

$$\begin{aligned}
&F\left(\frac{1}{t_n}s + \left(1 - \frac{1}{t_n}x_n\right)\right) - F(x_{n+1}) \\
&\geq \frac{L}{2t_n^2}\|t_n x_{n+1} - (s + (t_n - 1)x_n)\|^2 - \frac{L}{2t_n^2}\|t_n y_n - (s + (t_n - 1)x_n)\|^2
\end{aligned}$$

Simplying yields that

$$\|t_n y_n - (s + (t_n - 1)x_n)\|^2 = \|t_{n-1}x_n - (s + (t_{n-1} - 1))x_{n-1}\|^2.$$

Combined with the fact that $t_{n+1}^2 - t_{n+1} = t_n^2$, we get that

$$\begin{aligned}
t_{n-1}^2 s_n - t_n^2 s_{n+1} &\geq t_n^2\left(F\left(\frac{1}{t_n}s = \left(1 - \frac{1}{t_n}\right)\right)x_n\right) - F(x_{n+1}) \\
&\geq \frac{L}{2}\|t_n x_{n+1} - (s + (t_n - 1))x_n\|^2 - \frac{L}{2}\|t_n y_n - (s + (t_n - 1))x_n\|^2 \\
&= \frac{L}{2}\|t_n x_{n+1} - (s + (t_n - 1))x_n\|^2 - \frac{L}{2}\|t_{n-1}x_n - (s + (t_{n-1} - 1))x_{n-1}\|^2
\end{aligned}$$

Set $u_n := t_{n-1}x_n - (s + (t_{n-1} - 1)x_{n-1})$. Multiplying the inequality above by $\frac{2}{L}$ and rearranging yields

$$\|u_{n+1}\|^2 + \frac{2}{L}t_n^2 s_{n+1} \leq \|u_n\|^2 + \frac{2}{L}t_{n-1}^2 s_n.$$

It follows that

$$\begin{aligned}
\frac{2}{L}t_{n-1}^2 s_n &\leq \|u_n\|^2 + \frac{2}{L}t_n^2 s_{n+1} \\
&\leq \|u_1\|^2 + \frac{2}{L}t_0^2 s_1 \\
&= \|x_1 - s\|^2 + \frac{2}{L}(F(x_1) - \mu) \\
&\leq \|x_0 - s\|^2
\end{aligned}$$

where the last inequality follows from the proximal gradient inequality.

In other words,

$$\begin{aligned}
F(x_n) - \mu = s_n& \\
&\leq \frac{L}{2}\|x_0 - s\|^2 \frac{1}{t_{n-1}^2} \\
&\leq \frac{L}{2}\|x_0 - s\|^2 \frac{4}{(n+1)^2} && t_n \geq \frac{n+2}{2} \\
&= \frac{2Ld_S^2(x_0)}{(n+1)^2}.
\end{aligned}$$

3.6 Iterative Shrinkage Thresholding Algorithm

This is a special case of PGM with $g(x) = \lambda\|x\|, \lambda > 0$. Hence

$$\frac{1}{L}g(x) = \frac{\lambda}{L}\|x\|_1$$

and

$$\begin{aligned}\operatorname{Prox}_{\frac{1}{L}g}(x) &= \left(\operatorname{Prox}_{\frac{\lambda}{L}\|\cdot\|_1}(x)\right)_{i=1}^n \\ &= \left(\operatorname{sign}(x_i)\max\{0, |x_i| - \frac{\lambda}{L}\}\right)_{i=1}^n\end{aligned}$$

FISTA is the accelerated version of ISTA.

3.6.1 Norm Comparison

Consider the problems

$$\begin{aligned}&\min \|x\|_2 &&(P_1)\\ &Ax = b\\ &\min \|x\|_1 &&(P_2)\\ &Ax = b\end{aligned}$$

Example 3.6.1
Consider the problem

$$\min_{x\in\mathbb{R}^m} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \tag{P}$$

where $\lambda > 0$ and $A \in \mathbb{R}^{n\times m}$.

g is convex, l.s.c., and proper, with f being smooth and

$$\nabla f(x) = A^T(Ax - b).$$

Recall that ∇f is L-Lipschitz if and only if the spectral norm of the Hessian is bounded by L. Thus ∇f is L-Lipschitz for

$$L := \lambda_{\max}(A^T A).$$

To see the necessarily assumption that $S := \operatorname{argmin}_{x \in \mathbb{R}^m} F(x)$ holds, observe that $f(x)$ is continuous, convex, and coercive, with $\operatorname{dom} F = \mathbb{R}^m$.

Using the fact that if F is convex, l.s.c., proper, and coercive and $\varnothing \neq C$ is closed and convex with $\operatorname{dom} F \cap C \neq \varnothing$, then F has a minimizer over C.

Now, m can be very large and $\lambda_{\max}(A^T A)$ may be difficult to compute. It suffices to use some upper bound on eigenvalues such as the Frobenius norm

$$\begin{aligned} \|A\|_F^2 &= \sum_{j=1}^{m} \sum_{i=1}^{n} a_{ij}^2 \\ &= \operatorname{tr}\big(A^T A\big) \\ &= \sum_{i=1}^{m} \lambda_i(A^T A) \end{aligned}$$

3.7 Douglas-Rachford Algorithm

Consider the problem

$$\min_{x \in \mathbb{R}^m} F(x) := f(x) + g(x) \qquad (P)$$

where f, g are convex, l.s.c., and proper with

$$S := \operatorname{argmin}_{x \in \mathbb{R}^m} F(x) \neq \varnothing.$$

Suppose there exists some $s \in S$ such that

$$0 \in \partial f(s) + \partial g(s) \subseteq \partial (f+g)(s).$$

This happens for example when $\operatorname{ri} \operatorname{dom} f \cap \operatorname{ri} \operatorname{dom} g \neq \varnothing$.

Define

$$\begin{aligned} R_f &:= 2 \operatorname{Prox}_f - \operatorname{Id} \\ R_g &:= 2 \operatorname{Prox}_g - \operatorname{Id}. \end{aligned}$$

Definition 3.7.1 (Douglas-Rachford Operator)
Define

$$T := \operatorname{Id} - \operatorname{Prox}_f + \operatorname{Prox}_g R_f.$$

Lemma 3.7.1
The following hold:

(i) R_f, R_g are nonexpansive

(ii) $T = \frac{1}{2}(\operatorname{Id} + R_g R_f)$

(iii) T is firmly nonexpansive

Proof
Since $\operatorname{Prox}_f, \operatorname{Prox}_g$ are f.n.e., $2\operatorname{Prox}_f - \operatorname{Id}, 2\operatorname{Prox}_g - \operatorname{Id}$ are nonexpansive as shown in the assignments.

Expanding the definitions of R_g, R_f shows the equivalent expression

$$T = \frac{1}{2}(\operatorname{Id} + R_g R_g).$$

The above shows that T is $\frac{1}{2}$-averaged, which is equivalent to firm nonexpansiveness.

Proposition 3.7.2
$\operatorname{Fix} T = \operatorname{Fix} R_g R_f$.

Proof
Let $x \in \mathbb{R}^m$. Then

$$\begin{aligned} x \in \operatorname{Fix} T &\iff x = \frac{1}{2}(x + R_g R_f x) \\ &\iff x = R_g R_f x \\ &\iff x \in \operatorname{Fix} R_g R_f. \end{aligned}$$

Proposition 3.7.3
$\operatorname{Prox}_f(\operatorname{Fix} T) \subseteq S$.

Proof
Let $x \in \mathbb{R}^m$ and set $s = \operatorname{Prox}_f(x) = (\operatorname{Id} + \partial f)^{-1}(x)$. Then

$$\begin{aligned} x \in (\operatorname{Id} + \partial f)(s) = s + \partial f(s) &\iff 2s - (2s - x) \in s + \partial f(s) \\ &\iff 2s - R_f(x) - s \in \partial f(s) \\ &\iff s - R_f(x) \in \partial f(s). \end{aligned}$$

Moreover,

$$\begin{aligned} x \in \operatorname{Fix} T &\iff x = x - \operatorname{Prox}_f(x) + \operatorname{Prox}_g R_f(x) \\ &\iff s = \operatorname{Prox}_g R_f(x) = (\operatorname{Id} + \partial g)^{-1}(R_f(x)) \\ &\iff R_f(x) \in s + \partial g(s) \\ &\iff R_f(x) - s \in \partial g(s) \end{aligned}$$

It follows that

$$\begin{aligned} 0 &= s - R_f(x) + R_f(x) - s \\ &\in \partial f(s) + \partial g(s) \\ &\subseteq \partial(f+g)(s) \end{aligned}$$

and $s \in S$ as required for all $x \in \operatorname{Fix} T$.

Recall that (firmly) nonexpansive operators are continuous and iterating a f.n.e. operator tends to a fixed point.

Theorem 3.7.4
Let $x_0 \in \mathbb{R}^m$. Update via

$$x_{n+1} := x_n - \operatorname{Prox}_g x_n + \operatorname{Prox}_g(2\operatorname{Prox}_f x_n - x_n).$$

Then $\operatorname{Prox}_f(x_n)$ tends to a minimizer of $f + g$.

Proof
Remark that $x_{n+1} = Tx_n = T^{n+1}x_0$. Since T is f.n.e., we know that $x_n \to \bar{x} \in \operatorname{Fix} T$.

But since Prox_f is continuous,

$$\operatorname{Prox}_f x_n \to \operatorname{Prox}_f \bar{x} \in \operatorname{Prox}_f(\operatorname{Fix} T) \subseteq S.$$

3.8 Stochastic Projected Subgradient Method

Consider the problem

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

f is convex, l.s.c., and proper, $\varnothing \neq C \subseteq \operatorname{int} \operatorname{dom} f$ is closed and convex, and $S :=$

$\operatorname{argmin}_{x \in C} f(x) \neq \varnothing$.

Set

$$\mu := \min f(C).$$

Given $x_0 \in C$, update via

$$x_{n+1} := P_C(x_n - t_n g_n).$$

We assume that $t_n > 0$ and

$$\sum_{n=0}^{\infty} t_n \to \infty$$
$$\frac{\sum_{k=0}^{n} t_k^2}{\sum_{k=0}^{n} t_k} \to 0 \qquad k \to \infty$$

for example $t_n = \frac{\alpha}{n+1}$ for some $\alpha > 0$.

We choose g_n to be a random vector satisfying the following assumptions

(A1) For each $n \in \mathbb{N}$, $E[g_n \mid x_n] \in \partial f(x_n)$ (unbiased subgradient)

(A2) For each $n \in \mathbb{N}$, there is some $L > 0$, $E[\|g_n\|^2 \mid x_n] \leq L^2$

Let us write

$$\mu_k := \min\{f(x_i) : 0 \leq i \leq k\}.$$

Theorem 3.8.1
Assuming the previous assumptions hold, then $E[\mu_k] \to \mu$ as $k \to \infty$.

Proof
Pick $s \in S$ and let $n \in \mathbb{N}$. Then

$$\begin{aligned} 0 &\leq \|x_{n+1} - s\|^2 \\ &= \|P_C(x_n - t_n g_n) - P_C s\|^2 \\ &\leq \|(x_n - t_n g_n) - s\|^2 \\ &= \|(x_n - s) - t_n g_n\|^2 \\ &= \|x_n - s\|^2 - 2t_n \langle g_n, x_n - s\rangle + t_n^2 \|g_n\|^2 \end{aligned}$$

Taking the conditional expectation given x_n yields

$$\begin{aligned}E[\|x_{n+1} - s\|^2 \mid x_n] &\leq \|x_n - s\|^2 + 2t_n \langle E[g_n \mid x_n], s - x_n \rangle + t_n^2 E[\|g_n\|^2 \mid x_n] \\ &\leq \|x_n - s\|^2 + 2t_n (f(x) - f(x_n)) + t_n^2 L^2 && (A1), (A2) \\ &= \|x_n - s\|^2 + 2t_n (\mu - f(x_n)) + t_n^2 L^2.\end{aligned}$$

Now, taking the expection with respect to x_n yields

$$E[\|x_{n+1} - s\|^2] \leq E[\|x_n - s\|^2] + 2t_n(\mu - E[f(x_n)]) + t_n^2 L^2.$$

Let $k \in \mathbb{N}$. Summing the inequality from $n = 0$ to k yields

$$\begin{aligned}0 &\leq E[\|x_{n+1} - s\|^2] \\ &\leq \|x_0 - s\|^2 - 2 \sum_{n=0}^{k} t_n (E[f(x_n)] - \mu) + L^2 \sum_{n=0}^{k} t_n^2.\end{aligned}$$

Rearranging yields

$$\begin{aligned}0 &\leq E[\mu_k] - \mu \\ &\leq \frac{\|x_0 - s\|^2 + L^2 \sum_{n=0}^{k} t_n^2}{2 \sum_{n=0}^{k} t_n} \\ &\to 0 && k \to \infty\end{aligned}$$

3.8.1 Minimizing a Sum of Functions

Consider the problem

$$\begin{aligned}\min f(x) := \sum_{i \in [r]} f_i(x) && (P) \\ x \in C\end{aligned}$$

Suppose $f_1, \ldots, f_r : \mathbb{R}^m \to (-\infty, \infty]$ are convex, l.s.c., and proper.

Set $I := [r]$ and assume that for each $i \in I$,

$$\varnothing \neq C \subseteq \operatorname{int} \operatorname{dom} f_i.$$

for some closed convex C.

We also assuem that for each $i \in I$, there is some $L_i \geq 0$ for which
$$\|\partial f_i(C)\| \leq L_i.$$

Proposition 3.8.2
$\sup\|\partial f_i(C)\| \leq L_i$ if and only if $f_i\big|_C$ is L_i-Lipchistz.

For example, this holds if C is bounded.

Let us assume that (P) has a solution. We verify (A1), (A2) to justify solving the problem with SPSM.

By the triangle inequality,
$$\sup\|\partial f(C)\| \leq \sum_{i\in I} L_i.$$

Let $x_0 \in C$. Given $x_n \in C$, we pick an index $i_n \in I$ uniformly randomly and set
$$g_n := r f'_{i_n}(x_n) \in \partial f_{i_n}(x_n).$$
Observe that
$$\begin{aligned}
E[g_n \mid x_n] &= \sum_{i=1}^{r} \frac{1}{r} r f'_i(x_n) \\
&= \sum_{i=1}^{r} f'_i(x_n) \\
&\in \partial f_1(x_n) + \cdots + \partial f_r(x_n) \\
&= \partial(f_1 + \cdots + f_r)(x_n) && \text{Sum Rule} \\
&= \partial f(x_n)
\end{aligned}$$
hence (A1) holds.

Next,
$$\begin{aligned}
E[\|g_n\|^2 \mid x_n] &= \sum_{i=1}^{r} \frac{1}{r} \|r f'_i(x_n)\|^2 \\
&= \sum_{i=1}^{r} r\|f'_i(x_n)\|^2 \\
&\leq r \sum_{i=1}^{r} L_i^2.
\end{aligned}$$
Thus (A2) holds with $L := \sqrt{r \sum_{i=1}^{r} L_i^2}$.

Having verified the assumptions, we may apply SPSM.

3.9 Duality

3.9.1 Fenchel Duality

Consider the problem

$$\min_{x \in \mathbb{R}^m} f(x) + g(x) \tag{P}$$

$f, g : \mathbb{R}^m \to (-\infty, \infty]$ are convex, l.s.c., and proper.

We can rewrite the problem as

$$\min_{x,z \in \mathbb{R}^m} \{f(x) + g(z) : x = z\}.$$

Construct the Lagrangian

$$L(x, z; y) := f(x) + g(z) + \langle y, z - x \rangle.$$

The dual objective function is obtained by minimizing the Lagrangian with respect to x, z.

$$\begin{aligned}
d(u) &:= \inf_{x,z} L(x, z; u) \\
&= \inf_{x,z} \{f(x) - \langle u, x \rangle + g(z) + \langle u, z \rangle\} \\
&= -\sup_x (\langle u, x \rangle - f(x)) - \sup_z (\langle -u, z \rangle - g(z)) \\
&= -f^*(u) - g^*(-u).
\end{aligned}$$

Let

$$\begin{aligned}
p &:= \inf_{x \in \mathbb{R}^m} f(x) + g(x) \\
d &:= \inf_{u \in \mathbb{R}^m} f^*(u) + g^*(-u)
\end{aligned}$$

and recall that $p \geq -d$ from assignments.

3.9.2 Fenchel-Rockafellar Duality

Consider the problem

$$\min_{x \in \mathbb{R}^m} f(x) + g(Ax) \tag{P}$$

where $f : \mathbb{R}^m, \to (-\infty, \infty]$ is convex, l.s.c., and proper, $g : \mathbb{R}^n, \to (-\infty, \infty]$ is convex, l.s.c., and proper, and $A \in \mathbb{R}^{n\times m}$.

The Fenchel-Rockafellar dual is given by

$$\begin{aligned} &\min f^*(-A^T y) + g^*(y) \qquad (D) \\ &y \in \mathbb{R}^n \end{aligned}$$

As before, let

$$\begin{aligned} p &:= \inf_{x\in\mathbb{R}^m} f(x) + g(Ax) \\ d &:= \inf_{y\in\mathbb{R}^n} f^*(-A^T y) + g^*(y) \end{aligned}$$

and recall that $p \geq -d$ from assignments.

Lemma 3.9.1
Let $h : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. For each $x \in \mathbb{R}^m$,

$$h^v(x) := h(-x).$$

The following hold:

(i) h^v is convex, l.s.c., and proper

(ii) $\partial h^v = -\partial h \circ (-\operatorname{Id})$

Proof
The convexity, l.s.c., and properness is verified by definition.

Let $u \in \mathbb{R}^m$ and $x \in \operatorname{dom} \partial h \circ (-\operatorname{Id})$. Then

$$\begin{aligned} u \in -\partial h \circ (-\operatorname{Id})(x) = -\partial f(-x) &\iff -u \in \partial h(-x) \\ &\iff h(y) \geq h(-x) + \langle -u, y - (-x)\rangle \quad \forall y \in \mathbb{R}^m \\ &\iff h(-y) \geq h(-x) + \langle -u, -y + x\rangle \quad \forall y \in \mathbb{R}^m \\ &\iff h^v(y) \geq h^v(x) + \langle u, y - x\rangle \quad \forall y \in \mathbb{R}^m \\ &\iff u \in \partial h^v(x). \end{aligned}$$

3.9.3 Self-Duality of Douglas-Rachford

Recal that the DR operator to solve (P) is given by

$$T_p := \frac{1}{2}(\operatorname{Id} + R_g R_f)$$

where $R_f := 2\,\mathrm{Prox}_f - \mathrm{Id}$ and similarly for R_g.

Similarly, the DR operator to solve (D) is defined as

$$T_d := \frac{1}{2}(\mathrm{Id} + R_{(g^*)^v} R_{f^*}).$$

Lemma 3.9.2
Let $h : \mathbb{R}^m \to (-\infty, \infty]$ be convex, l.s.c., and proper. The following hold:

(i) $\mathrm{Prox}_{h^v} = -\mathrm{Prox}_h \circ (\mathrm{Id})$

(ii) $R_{h^*} = -R_h$

(iii) $R_{(h^*)^v} = R_h \circ (-\mathrm{Id})$

Proof
(i): This is shown using the relation $\mathrm{Prox}_f = (\mathrm{Id} + \partial f)^{-1}$ as well as the lemma $\partial h^v = -\partial h \circ (-\mathrm{Id})$.

(ii): This can be proven by expanding the definition of R_{h^*} as well as the relation $\mathrm{Prox}_{h^*} = (\mathrm{Id} - \mathrm{Prox}_h)$ proven in A4.

(iii): First, we can shown by definition that

$$\mathrm{Prox}_{(h^*)^v} = -\mathrm{Prox}_{h^*} \circ (-\mathrm{Id}).$$

The proof is completed using this fact as well as the relation $\mathrm{Prox}_{h^*} = (\mathrm{Id} - \mathrm{Prox}_h)$

Theorem 3.9.3
$T_p = T_d$.

Proof
From our previous lemma,

$$\begin{aligned} T_d &:= \frac{1}{2}(\mathrm{Id} + R_{(g^*)^v} R_{f^*}) \\ &= \frac{1}{2}(\mathrm{Id} + [R_g \circ (-\mathrm{Id})] \circ (-R_f)) \\ &= \frac{1}{2}(\mathrm{Id} + R_g R_f) \\ &= T_p. \end{aligned}$$

Theorem 3.9.4
Let $x_0 \in \mathbb{R}^m$. Update via

$$x_{n+1} := x_n - \operatorname{Prox}_f(x_n) + \operatorname{Prox}_g(2\operatorname{Prox}_f x_n - x_n) = T_p x_n.$$

Then $\operatorname{Prox}_f(x_n)$ converges to a minimizer of $f + g$ and $x_n - \operatorname{Prox}_f(x_n)$ converges to a minimizer of $f^* + (g^*)^v$.

Proof
We already know that $\operatorname{Prox}_f(x_n)$ converges to a minimizer of $f + g$. Since $T_p = T_d$, $\operatorname{Prox}_{f^*}(x_n)$ converges to a minimizer of $f^* + (g^*)^v$. Using the fact that $\operatorname{Prox}_{f^*} = \operatorname{Id} - \operatorname{Prox}_f$, we conclude the proof.