Lecture Notes on Convex Analysis and Iterative Algorithms

İlker Bayram
ibayram@ieee.org

About These Notes

These are the lectures notes of a graduate course I offered in the Dept. of Electronics and Telecommunications Engineering at Istanbul Technical University. The goal of the course was to get students acquainted with methods of convex analysis, and proximal algorithms. In the first half of the course, convex analysis is introduced at a level suitable for graduate students in electrical engineering (i.e., some familiarity with the notion of a convex set, convex functions from other courses). My goal was to make students more comfortable in following arguments that appear in recent signal processing literature, and understand/analyze the proximal point algorithm. I then derived several other algorithms based on the proximal point algorithm, such the Douglas-Rachford algorithm, ADMM, and applications to saddle point problems. There are no references in this version. I hope to add some in the future.

İlker Bayram
December, 2018

Contents

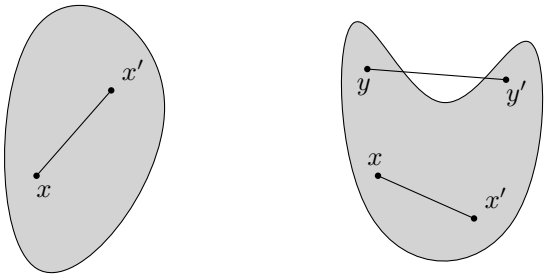
1 Convex Sets

This first chapter introduces convex sets and discusses some of their properties. Having a solid understanding of convex sets is very useful for convex analysis of functions because we can and will regard a convex function as a special representation of a convex set, namely its epighraph.

1.1 Basic Definitions

Definition 1. A set $C \in \mathbb{R}^n$ is said to be *convex* if $x \in C$, $x' \in C$ implies that $\alpha x + (1-\alpha)x \in C$ for any $\alpha \in [0,1]$. $\diamond$

Consider the sets below. Each pair of (x, x') we select in the set on the left defines a line segment which lies inside the set. However, this is not the case for the set on the right. Even though we are able to find line segments with endpoints inside the set (as in (x, x')), this is not true in general, as exemplified by (y, y').



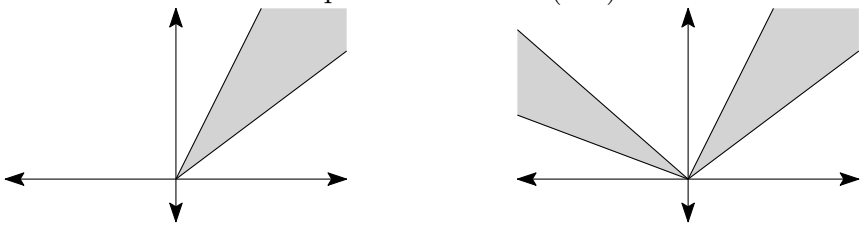
For the examples below, decide if the set is convex or not and prove whatever you think is true.

Example 1 (Hyperplane)**.** For given $s \in \mathbb{R}^n$, $r \in \mathbb{R}$, consider the set $H_{s,r} = \{x \in \mathbb{R}^n : \langle s, x\rangle = r\}$. Notice that this is a subspace for $r = 0$.

Example 2 (Affine Subspace)**.** This is a set $V \in \mathbb{R}^n$ such that if $x \in V$ and $x' \in V$, then $\alpha x + (1-\alpha)x' \in V$ for all $\alpha \in \mathbb{R}$.

Example 3 (Half Space)**.** For given $s \in \mathbb{R}^n$, $r \in \mathbb{R}$, consider the set $H^-_{s,r} = \{x \in \mathbb{R}^n : \langle s, x\rangle \leq r\}$.

Example 4 (Cone)**.** A cone $K \in \mathbb{R}^n$ is a set such that if $x \in K$, then $\alpha x \in K$ for all $\alpha > 0$. Note that a cone may be convex or non-convex. See below for an example of a convex (left) and a non-convex (right) cone.



Exercise 1. Let K be a cone. Show that K is convex if and only if $x, y \in K$ implies $x + y \in K$. ◇

1.2 Operations Preserving Convexity of Sets

Proposition 1 (Intersection of Convex Sets)**.** Let C_1, C_2, $\ldots C_k$ be convex sets. Show that $C = \cap_i C_i$ is convex.

Proof. Exercise! □

Question 1. What happens if the intersection is empty? Is the empty set convex? ◇

This simple result is useful for characterizing linear systems of equations or inequalities.

Example 5. For a matrix A, the solution set of $Ax = b$ is an intersection of hyperplanes. Therefore it is convex.

For the example above, we can in fact say more, thanks to the following variation of Prop. 1.

Exercise 2. Show that the intersection of a finite collection of affine subspaces is an affine subspace. ◇

Let us continue with systems of linear inequalities.

Example 6. For a matrix A, the solution set of $Ax \leq b$ is an intersection of half spaces. Therefore it is convex.

Proposition 2 (Cartesian Products of Convex Sets)**.** Suppose $C_1, \ldots, C_k$ are convex sets in $\mathbb{R}^n$. Then the Cartesian product $C_1 \times \cdots \times C_k$ is a convex set in $\mathbb{R}^{n \times \cdots \times n}$.

Proof. Exercise! □

Given an operator F and a set C, we can apply F to elements of C to obtain the image of C under F. We will denote that set as $F\,C$. If F is linear then it preserves convexity.

Proposition 3 (Linear Transformations of Convex Sets)**.** Let M be a matrix. If C is convex, then $M\,C$ is also convex.

Consider now an operator that just adds a vector $d$to its operand. This is a translation operator. Geometrically, it should be obvious that translation preserves convexity. It is a good exercise to translate this mental picture to an algebraic expression and show the following.

Proposition 4 (Translation)**.** If C is convex, then the set $C + d = \{x : x = v + d \text{ for some } v \in C\}$is also convex.

Given C_1, C_2, consider the set of points of the form $v = v_1 + v_2$, where $v_i \in C_i$. This set is denoted by $C_1 + C_2$and is called the Minkowski sum of $C_1$and C_2. We have the following result concerning Minkowski sums.

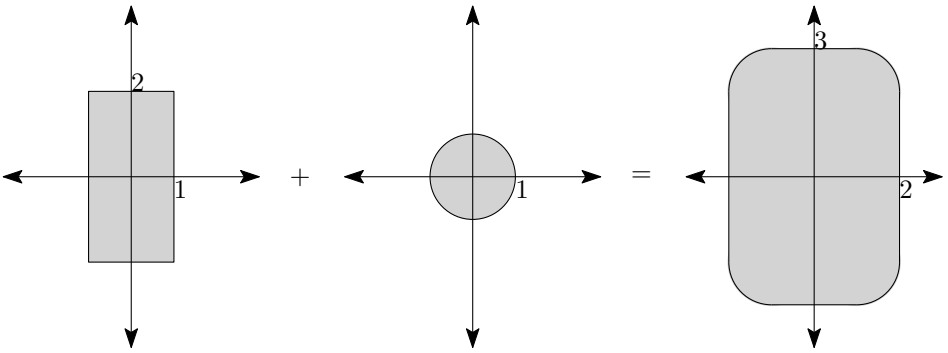
Proposition 5 (Minkowski Sums of Convex Sets)**.** If C_1 and C_2 are convex then $C_1 + C_2$is also convex.

Proof. Observe that

$$C_1 + C_2 = \begin{bmatrix} I & I \end{bmatrix} (C_1 \times C_2).$$

Thus it follows by Prop. 2 and Prop. 3 that $C_1 + C_2$ is convex. □

Example 7. The Minkowski sum of a rectangle and a disk in $\mathbb{R}^2$ in shown below.



◇

Definition 2 (Convex Combination)**.** Consider a finite collection of points x_1, $\ldots$, x_k. x is said to be a convex combination of x_i's if x satisfies

$$x = \alpha_1 x_1 + \cdots + \alpha_k x_k$$

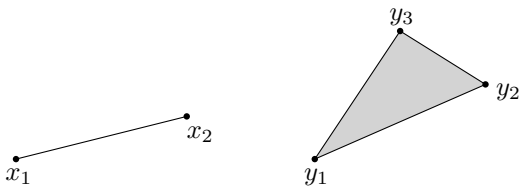
for some α_i such that

$$\alpha_i \geq 0 \text{ for all } i,$$
$$\sum_{i=1}^{k} \alpha_k = 1.$$

◇

Definition 3 (Convex Hull)**.** The set of all convex combinations of a set C is called the convex hull of C and is denoted as $\text{Co}(C)$. ◇

Below are two examples, showing the convex hull of the sets $C = \{x_1, x_2\}$, and $C' = \{y_1, y_2, y_3\}$.



Notice that in the definition of the convex hull, the set C does not have to be convex (in fact, C is not convex in the examples above). Also, regardless of the dimension of C, when constructing $\mathrm{Co}(C)$, we can consider convex combinations of any number of finite elements chosen from C. In fact, if we denote the set of all convex combinations of $k$elements from C as C_k, then we can show that $C_k \subset C_{k+m}$ for $m \geq 0$. An interesting result, which we will not use is in this course, is the following.

Exercise 3 (Caratheodory's Theorem)**.** Show that, if $C \in \mathbb{R}^n$, then $\mathrm{Co}(C) = C_{n+1}$. ◇

The following proposition justifies the term 'convex' in the definition of the convex hull.

Proposition 6. The convex hull of a set C is convex.

Proof. Exercise! □

The convex hull of C is the smallest convex set that contains C. More precisely, we have the following.

Proposition 7. If $D = \mathrm{Co}(C)$, and if E is a convex set with $C \subset E$, then $D \subset E$.

Proof. The idea is to show that for any integer k, Econtains all convex combinations involving k elements from C. For this, we will present an argument based on induction.

We start with $k = 2$. Suppose $x_1, x_2 \in C$. This implies $x_1, x_2 \in E$. Since E is convex, we have $\alpha x_1 + (1-\alpha)x_2 \in E$, for all $\alpha \in [0, 1]$. Since x_1, x_2 were arbitrary elements of C, it follows that Econtains all convex combinations of any two elements from C.

Suppose now that Econtains all convex combinations of any $k-1$ elements from C. That is, if $x_1, \ldots x_{k-1}$ are in C, then for $\sum_{i=1}^{k-1} \alpha_i = 1$, and $\alpha_i \geq 0$, we have $\sum_{i=1}^{k-1} \alpha_i\, x_i \in E$. Suppose we pick a k^{th} element, say x_k, from C. Also, let $\alpha_1, \ldots \alpha_k$be on the unit simplex, with $\alpha_k \neq 0$ (if $\alpha_k = 0$, we have nothing

to prove). Observe that

$$\begin{aligned}\sum_{i=1}^{k} \alpha_i \, x_i &= \alpha_k \, x_k + \sum_{i=1}^{k-1} \alpha_i \, x_i \\ &= \alpha_k \, x_k + (1 - \alpha_k) \sum_{i=1}^{k-1} \frac{\alpha_i}{1 - \alpha_k} \, x_i.\end{aligned}$$

Notice that

$$\begin{aligned}\sum_{i=1}^{k-1} \frac{\alpha_i}{1 - \alpha_k} &= 1, \\ \frac{\alpha_i}{1 - \alpha_k} &\geq 0, \text{ for all } i.\end{aligned}$$

Therefore,

$$y = \sum_{i=1}^{k-1} \frac{\alpha_i}{1 - \alpha_k} \, x_i$$

is an element of E since it is a convex combination of $k-1$ elements of C. But then,

$$\sum_{i=1}^{k} \alpha_i \, x_i = \alpha_k \, x_k + (1 - \alpha_k) \, y$$

is an element of E (why?). We are done. □

Another operation of interest is the affine hull. For that, let us introduce affine combinations.

Definition 4 (Affine Combination)**.** Consider a finite collection of points x_1, $\ldots$, x_k. x is said to be an affine combination of x_i's if x satisfies

$$x = \alpha_1 \, x_1 + \cdots + \alpha_k \, x_k$$

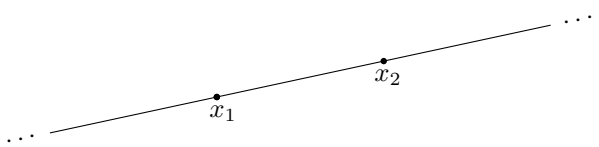
for some α_i such that

$$\sum_{i=1}^{k} \alpha_i = 1.$$

◇

Definition 5 (Affine Hull)**.** The set of all affine combinations of a set C is called the affine hull of C. ◇

The convex hull of two points x_1, x_2 is a line segment passing through the two points.



Exercise 4. Consider a set $C \subset \mathbb{R}^2$, composed of two points $C = \{x_1, x_2\}$. What is the difference between the affine and convex hull of C? ◇

Let us end this discussion with some definitions.

Definition 6 (Interior)**.** x is said to be in the interior of Cif there exists an open set Bsuch that $x \in B$ and $B \subset E$. The interior of C is denoted as $\text{int}(C)$. ◇

Definition 7 (Boundary)**.** Boundary of a set C is defined to be $C \cap \big(\text{int}(C)\big)^c$. ◇

1.3 Projections onto Closed Convex Sets

Definition 8 (Projection)**.** Let C be a set. For a given point y (inside or outside C), $x \in C$ is said to be a projection of y onto C if it is a minimizer of the following problem

$$\min_{z \in C} \|y - z\|_2.$$

◇

In general, projections may not exist, or may not be uniquely defined.

Example 8. Suppose D denotes the open disk in $\mathbb{R}^2$ and $y$be such that $\|y\|_2 > 1$. Then, the projection of y onto D does not exist. Notice that D is convex but not closed. ◇

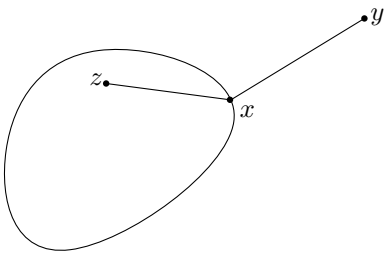
Example 9. Let C be the unit circle. Can you find a point $y \notin C$ that has infinitely many projections? Can you find a point which does not have a projection onto C? Notice in this case that C is not convex, but closed. ◇

The two examples above imply that projections are not always guaranteed to be well-defined. In fact, convexity alone is not sufficient. We will see later that convexity and closedness together ensure the existence of a unique minimizer. The following provides a useful characterization of the projection.

Proposition 8. Let C be a convex set. x is a projection of y onto x if and only if

$$\langle z - x, y - x \rangle \leq 0, \text{ for all } z \in C.$$

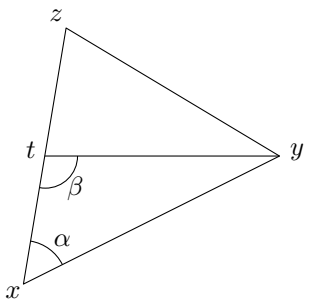
It is useful to understand what this proposition means geometrically. Consider the figure below. If x is the projection of y onto the convex set, then Prop.8 implies that the angle $\widehat{zxy}$ is obtuse.



Proof of Prop. 8. ($\Rightarrow$) Suppose $x = P_C(y)$ but there exists z such that

$$\langle z - x, y - x \rangle > 0.$$

The idea is as follows. Consider the figure below.



Pick t such that $\beta > \alpha$. But then

$$\|y - t\|_2 < \|y - x\|_2.$$

Thus, x cannot be the projection.

Let us now provide an algebraic proof. Consider

$$\begin{aligned} \|y - \big(\alpha z + (1-\alpha)\, x\big)\|_2^2 &= \|y - x + \alpha(x - z)\|_2^2 \\ &= \|y - x\|_2^2 + \alpha^2 \underbrace{\|x - z\|_2^2}_{d} + 2\alpha \underbrace{\langle x - z, y - x \rangle}_{-c} \end{aligned}$$

Notice now that, since $c > 0$ by assumption, the polynomial $\alpha^2 d - 2\alpha c$ is negative in the interval $(0, 2c/d)$. Thus we can find $\alpha \in (0,1)$ such that $\|y - \big(\alpha z + (1-\alpha)\, x\big)\|_2^2$ is strictly less than $\|y - x\|_2^2$. But this contradicts the assumption that $z = P_C(y)$.

($\Leftarrow$) Suppose $x \in C$ and $x \neq P_C(y)$. Let $z = P_C(y)$. Also, suppose that

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

Consider

$$\begin{aligned} \|y - z\|_2^2 &= \|y - x + x - z\|_2^2 \\ &= \|y - x\|_2^2 + \|x - z\|_2^2 + 2\underbrace{\langle x - z, y - x \rangle}_{c}. \end{aligned}$$

Now if $c > 0$, then $\|y - x\|_2 < \|y - z\|_2$. But this is a contradiction. □

Corollary 1. If C is closed, convex then $P_C(y)$ is a unique point.

Proof. Suppose $x_1, x_2 \in C$ and

$$\|x_1 - y\|_2 = \|x_2 - y\|_2 \leq \|z - y\|_2 \text{ for all } z \in C.$$

Then, we have, by Prop. 8 that

$$\langle y - x_1, x_2 - x_1 \rangle \leq 0,$$
$$\langle y - x_2, x_1 - x_2 \rangle \leq 0.$$

Adding these inequalities, we obtain

$$\langle x_1 - x_2, x_1 - x_2 \rangle \leq 0,$$

which implies $x_1 = x_2$. □

Projection operators enjoy useful properties. One of them is the following, which we will refer to as 'firm nonexpansivity' (to be properly defined later).

Proposition 9. $\|P_C(x_1) - P_C(x_2)\|_2^2 \leq \langle P_C(x_1) - P_C(x_2), x_1 - x_2 \rangle$.

Proof. Let $p_i = P_C(x_i)$. Then, we have

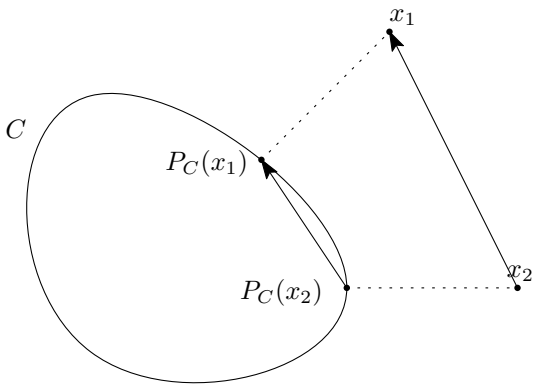
$$\langle x_1 - p_1, p_2 - p_1 \rangle \leq 0,$$
$$\langle x_2 - p_2, p_2 - p_1 \rangle \leq 0.$$

Summing these, we have

$$\langle (p_2 - p_1) + (x_1 - x_2), p_2 - p_1 \rangle \leq 0.$$

Rearranging, we obtain the desired inequality. □

Consider the following figure. The proposition states that the inner product of the two vectors is greater than the length of the shorter one squared.



As a first corollary of this proposition, we have :

Corollary 2. For a closed, convex C, we have $\langle P_C(x_1) - P_C(x_2), x_1 - x_2 \rangle \geq 0$.

Applying the Cauchy-Schwarz inequality, we obtain from Prop. 9 that projection operators are 'non-expansive' (also, to be discussed later).

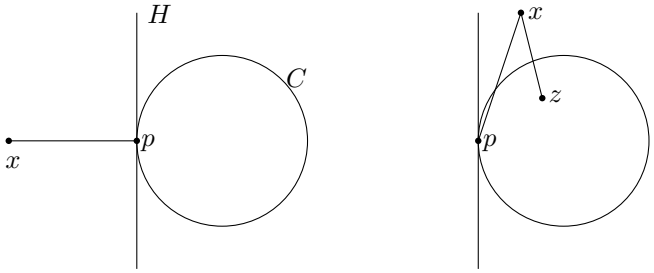
Corollary 3. For a closed, convex C, we have $\|P_C(x_1) - P_C(x_2)\|_2 \leq \|x_1 - x_2\|_2$.

1.4 Separation and Normal Cones

Proposition 10. Let C be a closed convex set and $x \notin C$. Then, there exists s such that

$$\langle s, x \rangle > \sup_y \langle s, y \rangle.$$

Proof. To outline the idea of the proof, consider the left figure below.



The hyperplane H, normal to $x - p$ touching C at p should separate x and C. If this is not the case, the situation resembles the right figure above, and we should have $\|x - z\| < \|x - p\|$.

Algebraically, $\langle x - P_C(x), z - P_C(x) \rangle \leq 0$, for all $z \in C$. Set $s = x - P_C(x)$. Then, we have

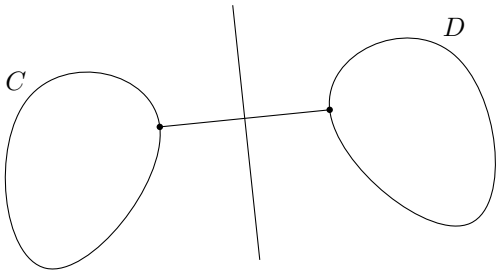
$$\begin{aligned} &\langle s, z - x + s \rangle \leq 0, \ \forall z \in C \\ \Longleftrightarrow &\langle s, x \rangle \geq \langle s, s \rangle + \langle s, z \rangle, \ \forall z \in C. \end{aligned}$$

□

As a generalization, we have the following result.

Proposition 11. Let C, D be disjoint closed convex sets. Suppose also that $C - D$ is closed. Then, there exists s such that $\langle s, x \rangle > \langle s, y \rangle$, for all $x \in C$, $y \in D$.

Proof. The idea is to consider the segment between the closest points of C and D, and construct a separating hyperplane that is orthogonal to this segment, as visualized below.



We want to find s such that

$$\langle s, y - x\rangle < 0, \forall x \in C,\ y \in D.$$

Note that $y - x \in D - C$. Also, since $C \cap D = \emptyset$, we have $0 \notin C - D$. Therefore, there exists s such that $\langle s, 0\rangle > \langle s, z\rangle$, for all $z \in D - C$. This s satisdies

$$\langle s, y - x\rangle < 0, \quad \forall\ y \in D, x \in C.$$

□

Definition 9. An affine hyperplane $H_{s,r}$ is said to support the set C if $\langle s, x\rangle \leq r$ for all $x \in C$. Notice that this is equivalent to $C \subset H^-_{s,r}$. ◇

For a given set C, let Σ_C denote the set of (s, r) such that $C \subset H^-_{s,r}$.

Proposition 12. If C is closed and convex, then

$$C = \cap_{(s,r)\in\Sigma_C} H^-_{s,r}.$$

Proof. The proof follows by showing that the expressions on the rhs and lhs are subsets of each other.

Obviously,

$$C \subset \cup_{(s,r)\in\Sigma_C} H^-_{s,r}.$$

Let us now show the other inclusion. Take $x \notin C$. Then, there exists p such that

$$\langle x, p\rangle < \langle z, p\rangle, \quad \forall\, z \in C.$$

Take $q = P_C(x)$. Then $H^-_{p,q} \supset C$, and $x \notin H^-_{p,q}$. Since $(p, q) \in \Sigma_C$, we find that $x \notin \cap_{(s,r)\in\Sigma_C} H^-_{s,r}$. Thus,

$$C \supset \cup_{(s,r)\in\Sigma_C} H^-_{s,r}.$$

□

We also state, without proof, the following result, that will be useful in the discussion of duality.

Proposition 13. Suppose C is a convex set. There exists a supporting hyperplane for any $x \in \mathrm{bd}(C)$.

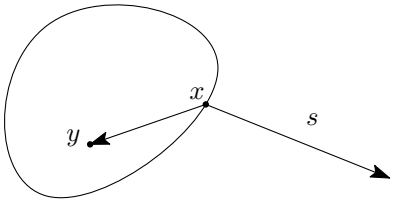
1.5 Tangent and Normal Cones

Definition 10. Let C be a closed convex set. The direction $s \in \mathbb{R}^n$ is said to be normal to C at x when

$$\langle s, y - x \rangle \leq 0, \quad \forall\, y \in C.$$

$\diamond$

According to the definition, for any $y \in C$, the angle between the two vectors shown below is obtuse.



Notice that if s is normal to C at x, then αs is also normal, for $\alpha \geq 0$. The set of normal directions is therefore a cone.

Definition 11. The cone mentioned above is called the normal cone of C at x, and is denoted as $N_C(x)$. $\diamond$

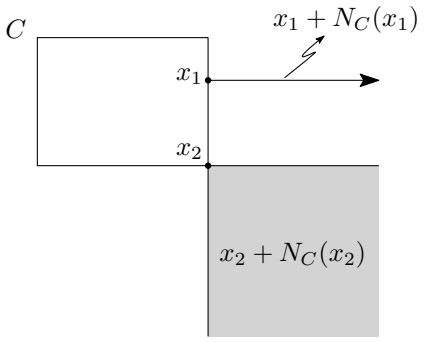
Proposition 14. $N_C(x)$ is a convex cone.

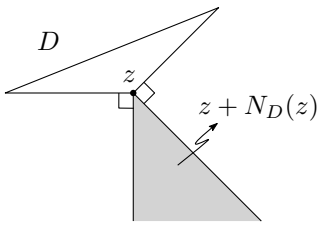
Proof. If s_1, s_2 are in $N_C(x)$, then

$$\langle \alpha s_1 + (1-\alpha) s_2, y - x \rangle = \alpha \langle s_1, y - x \rangle + (1-\alpha) \langle s_2, y - x \rangle \leq 0,$$

implying that $\alpha s_1 + (1-\alpha) s_2 \in N_C(x)$. $\square$

Below are two examples.





From the definition, we directly obtain the following results on normal cones.

Proposition 15. Let C be closed, convex. If $s \in N_C(x)$, then

$$P_C(x + \alpha s) = x, \quad \forall\ \alpha \geq 0.$$

Normal cones will be of interest when we discuss constrained minimization, and subdifferentials of functions.

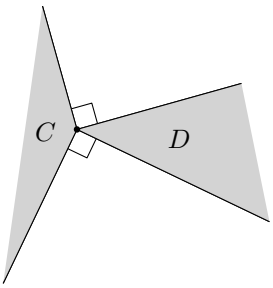
Let us now define a related cone through 'polarity'.

Definition 12. Given a closed cone C, the polar cone of C is the set of p such that

$$\langle p, s\rangle \leq 0, \quad \forall\, s \in C.$$

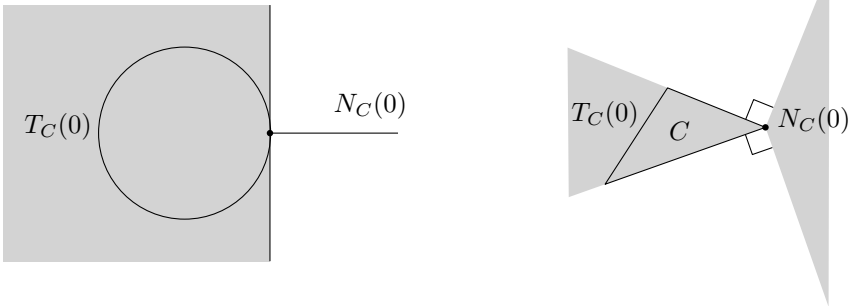
$\diamond$

In the figure below, the dot indicates the origin, and D is the polar cone of C. Note also that C is the polar cone of D.



Definition 13. The polar cone of $N_C(x)$ is called the tangent cone of C at x, and is denoted by $T_C(x)$. $\diamond$

The figures below show the tangent cones of the sets at the origin (the origin is indicated by a dot).



Proposition 16. For a convex C, we have $x + T_C(x) \supset C$.

Proof. If $p \in C$ and $s \in N_C(x)$, then

$$\begin{aligned} & \langle p - x, s\rangle \leq 0 \\ \Rightarrow \quad & p - x \in T_C(x). \end{aligned}$$

$\square$

Proposition 17. Suppose C is closed and convex. Then

$$\cap_{x \in C} x + T_C(x) = C.$$

Proof. Let us denote the set on the lhs as D. Note that, by the previous proposition, $D \supset C$.

To see the converse inclusion, take $z \notin C$. Let $x = P_C(z)$. Then $z - x \in N_C(x)$ and $z - x \notin T_C(x)$. Thus $z \notin x + T_C(x)$, and thus $z \notin D$. □

Proposition 18. $T_C(x)$ is the closure of the cone generated by $C - x$.

Exercise 5. Show that C is convex if and only if $\frac{1}{2}(x+y) \in C$, for all $x, y \in C$. ◇

Proposition 19. Let $x \in \mathrm{bd}(C)$, where C is convex. There exists s such that

$$\langle s, x \rangle \leq \langle s, y \rangle, \quad \forall y \in C.$$

Proof. Consider a sequence $\{x_k\}_k$ with $x_k \notin C$ and $\lim_k x_k = x$. Also, let $y \in C$. We can find a sequence $\{s_k\}_k$ with $\|s_k\|_2 = 1$ such that $\langle s_k, x_k \rangle \leq \langle s_k, y \rangle$ for all k. Now, extract a convergent subsequence s_{k_i} with limit s (such a subsequence exists by the Bolzano-Weierstrass theorem, since s_k are bounded, and are in $\mathbb{R}^n$.). Then, we must have

$$\langle s_{k_i}, x_{k_i} \rangle \leq \langle s_{k_i}, y \rangle, \quad \forall i.$$

Taking limits, we obtain $\langle s, x \rangle \leq \langle s, y \rangle$. □

Alternative Proof (Sketch). Assume $N_C(x) \neq \emptyset$. Take $z \in N_C(x)$. Then, $\langle z, y - x \rangle \leq 0$. This is equivalent to $\langle -z, x \rangle \leq \langle -z, y \rangle$. □

2 Convex Functions

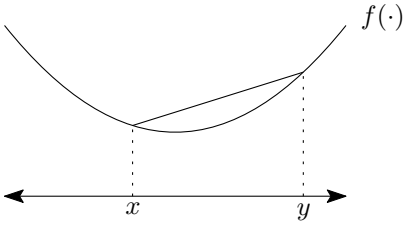
The standard definition is as follows.

Definition 14. $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if for all x, $y \in \mathbb{R}^n$, and $\alpha \in [0, 1]$, we have

$$f\big(\alpha x + (1-\alpha)y\big) \leq \alpha f(x) + (1-\alpha)\, f(y).$$

If the above inequality can be made strict, the function is said to be strictly convex.

The inequality is demonstrated in the figure below.

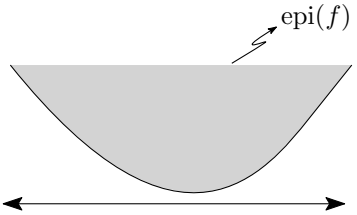


◇

In order to link convex functions and sets, let us also introduce the following.

Definition 15. Given $f : \mathbb{R}^n \to \mathbb{R}$, the epigraph of f is the subset of $\mathbb{R}^{n+1}$ defined as

$$\text{epi}(f) = \Big\{(x, r) \in \mathbb{R}^n \times \mathbb{R} : r \geq f(x)\Big\}.$$



◇

Proposition 20. f is convex if and only if its epigraph is a convex set in $\mathbb{R}^{n+1}$.

Proof. ($\Rightarrow$) Suppose f is convex. Pick (x_1, r_1), (x_2, r_2) from epi(f). Then,

$$r_1 \geq f(x_1)$$
$$r_2 \geq f(x_2).$$

Using the convexity of f, we have,

$$f\big(\frac{1}{2}x_1 + \frac{1}{2}x_2\big) \leq \frac{1}{2}\Big(f(x_1) + f(x_2)\Big) \leq \frac{1}{2}(r_1 + r_2).$$

Therefore

$$\frac{1}{2}\Big((x_1, r_1) + (x_2, r_2)\Big) \in \text{epi}(f),$$

and so epi(f) is convex.

($\Leftarrow$) Suppose epi(f) is convex. Notice that since $(x_1, f(x_1))$, $(x_2, f(x_2)) \in$ epi(f), we have

$$\big(\alpha x_1 + (1-\alpha)x_2, \alpha f(x_1) + (1-\alpha)f(x_2)\big) \in \text{epi}(f).$$

But this means that

$$f\big(\alpha x_1 + (1-\alpha)x_2\big) \leq \alpha f(x_1) + (1-\alpha)\, f(x_2).$$

Thus, f is convex. □

Definition 16. The domain of $f : \mathbb{R}^n \to \mathbb{R}$ is the set

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

◇

Example 10. Let C be a set in $\mathbb{R}^n$. Consider the function

$$i_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ \infty, & \text{if } x \notin C. \end{cases}$$

$i_C(x)$ is called the indicator function of the set C. Its domain is the set C.

Exercise 6. Show that $i_C(x)$ is convex if and only if C is convex. ◇

Exercise 7. Consider the function

$$u_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ 1, & \text{if } x \notin C. \end{cases}$$

Determine if $u_C(x)$ is convex. If so, under what conditions? ◇

Proposition 21 (Jensen's inequality)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $x_1, \ldots, x_k$, be points in its domain. Also, let $\alpha_1, \ldots, \alpha_k \in [0,1]$ with $\sum_i \alpha_i = 1$. Then,

$$f\Big(\sum_i \alpha_i x_i\Big) \leq \sum_i \alpha_i f(x_i). \tag{2.1}$$

Proof. Notice that $\big(x_i, f(x_i)\big) \in$ epi(f), for $i = 1, \ldots, k$. Therefore,

$$\left(\sum_i \alpha_i x_i, \sum_i \alpha_i f(x_i)\right) \in \text{epi}(f).$$

But this is equivalent to (2.1). □

Definition 17. f is said to be concave if $-f$ is convex. $\diamond$

Below are some examples of convex functions.

Example 11. Affine functions : $f(x) = \langle s, x\rangle + b$, for some s and b. In relation with this, determine if $f(x, y) = xy$ is convex.

Example 12 (Norms)**.** Suppose $\|\cdot\|$ is a norm. Then by the triangle inequality, and the homogeneity of the norm, we have

$$\|\alpha x + (1-\alpha)y\| \leq \alpha\|x\| + (1-\alpha)\|y\|.$$

In particular, for $1 \leq p \leq \infty$, the ℓ_p norm is defined as

$$\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}.$$

Show that $\|x\|_p$ is actually a norm. What happens if $p < 1$?

Example 13 (Quadratic Forms)**.** $f(x) = x^T Q x$ is convex if $Q + Q^T$ is positive semidefinite. Show this!

Exercise 8. Show that $f(x)$ above is not convex if $Q + Q^T$ is not positive semi-definite. $\diamond$

2.1 Operations That Preserve the Convexity of Functions

1. Translation by an arbitrary amount, multiplication with a non-negative constant.
2. Dilation : If $f(x)$ is convex, so is $f(\alpha x)$, for $\alpha \in \mathbb{R}$. (Follows by considering the epigraph.)
3. Pre-Composition with a matrix : If $f(x)$ is convex, so is $f(Ax)$. Show this!
4. Post-Composition with an increasing convex function : Suppose $g : \mathbb{R} \to \mathbb{R}$ is an increasing convex function and $f : \mathbb{R}^n \to \mathbb{R}$ is convex. Then, $g\big(f(\cdot)\big)$ is convex.

   Proof. Since g is increasing, and convex we have

   $$\begin{aligned} g\Big(f\big(\alpha x_1 + (1-\alpha)x_2\big)\Big) &\leq g\Big(\alpha f(x_1) + (1-\alpha) f(x_2)\Big) \\ &\leq \alpha\, g\Big(f(x_1)\Big) + (1-\alpha)\, g\Big(f(x_2)\Big). \end{aligned}$$

   $\square$

5. Pointwise supremum of convex functions : Suppose $f_1(\cdot), \dots f_k(\cdot)$ are convex functions, and define

$$g(x) = \max_i f_i(x).$$

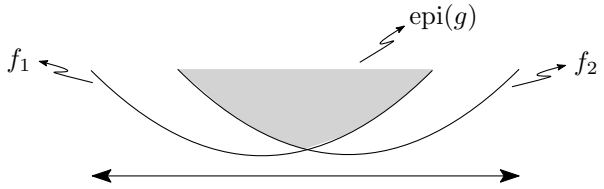
Then g is convex.

Proof. Notice that

$$\text{epi}(g) = \bigcap_i \text{epi}(f_i).$$

Since intersections of convex sets are convex, $\text{epi}(g)$ is convex, and so g is convex.

Below is a visual demonstration of the proof.



□

2.2 First Order Differentiation

Proposition 22. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. Then, f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in \mathbb{R}^n.$$

Proof. ($\Rightarrow$) Suppose f is convex. Then,

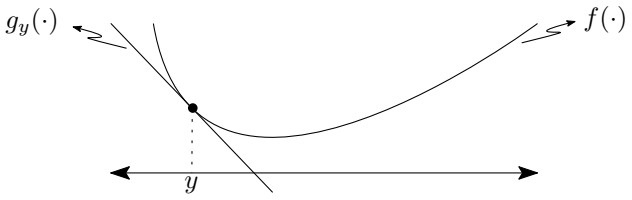
$$f\big(x + \alpha(y - x)\big) \leq (1 - \alpha) f(x) + \alpha f(y), \text{ for } 0 \leq \alpha \leq 1.$$

Rearranging, we obtain

$$\frac{f\big(x + \alpha(y - x)\big) - f(x)}{\alpha} \leq f(y) - f(x) \text{ for } 0 \leq \alpha \leq 1.$$

Letting $\alpha \to 0$, the left hand side converges to $\langle \nabla f(x), y - x \rangle$.

($\Rightarrow$) Consider the function $g_y(x) = f(y) + \langle \nabla f(y), x - y \rangle$. Then, $g_y(x) \leq f(x)$ and $g_y(y) = f(y)$ (see below). Also, $g_y(x)$ is convex (in fact, affine).



Now set

$$h(x) = \sup_y g_y(x).$$

But, by the two properties of $g_y(\cdot)$ above, it follows that $h(x) = f(x)$. But since $h(x)$ is the supremum of a collection of convex functions, it is convex. Thus, it follows that $f(x)$ is convex. □

We remark that a similar construction as in the second part of the proof will lead to the conjugate of $f(x)$, which will be discussed later.

Note that if $f : \mathbb{R} \to \mathbb{R}$ is differentiable and convex, then $f'(\cdot)$ is a monotonously increasing function. To see this, suppose that $f'(x_0) = 0$. Then, for $y \geq x_0$, we can show that $f'(y) \geq f'(x_0) = 0$. Indeed, $f(y) \geq f(x_0)$ because of the proposition. If $f'(y) < 0$, then

$$f(x_0) \geq \underbrace{f(y)}_{\geq f(x_0)} + \underbrace{\langle f'(y), x - y \rangle}_{>0} > f(x_0),$$

which is a contradiction. Therefore, we must have $f'(y) \geq 0$.

To generalize this argument, apply it to

$$h_{x_0}(x) = f(x) - \big(f'(x_0)\, x\big)$$

Observe that $h'(x_0) = 0$, and $h'_{x_0}(x) = f'(x) - f'(x_0)$.

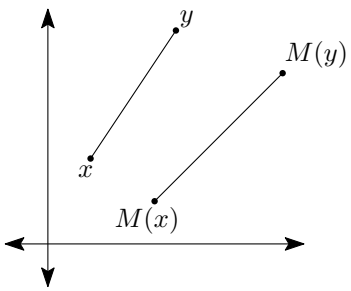
Question 2. How does the foregoing discussion generalize to convex functions defined on $\mathbb{R}^n$? ◇

Definition 18. An operator $M : \mathbb{R}^n \to \mathbb{R}^n$ is said to be monotone if

$$\langle M(x) - M(y), x - y \rangle \geq 0, \quad \text{for all } x, y \in \mathbb{R}^n.$$

◇

Below is an instance of this relation. The two vectors $x - y$ and $M(x) - M(y)$ approximately point in the same direction.



Proposition 23. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then, f is convex if and only if ∇f is monotone.

Proof. ($\Rightarrow$) Suppose f is convex. This implies the following two inequalities.

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

Rearranging these inequalities, we obtain

$$0 \geq \langle \nabla f(x) - \nabla f(y), y - x \rangle.$$

($Leftarrow$) Suppose ∇f is monotone. For $y, x \in \mathbb{R}^n$, let $z = y - x$. Then

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + \alpha z), z \rangle \, d\alpha.$$

Rearranging,

$$f(y) - f(x) - \int_0^1 \langle \nabla f(x), z \rangle d\alpha = \int_0^1 \langle \nabla f(x + \alpha z) - \nabla f(x), z \rangle \, d\alpha.$$

But the right hand side is non-negative by the monotonicity of ∇f. It follows that

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

It then follows by Prop. 22 that f is convex. □

2.3 Second Order Differentiation

We have seen that $f : \mathbb{R} \to \mathbb{R}$ is convex if and only if its first derivative if monotone increasing. But the latter property is equivalent to the second derivative being non-negative. Therefore, we also have an additional equivalent condition that involves the second derivative. This generalizes as follows.

Proposition 24. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function. Then, f is convex if and only if $\nabla^2 f$ is positive semi-definite.

Proof. ($\Rightarrow$) If f is convex, then $F = \nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone mapping, by Prop. 23. Let $d \in \mathbb{R}^n$. Then,

$$\langle F(x + \alpha d) - F(x), \alpha d \rangle \geq 0 \quad \text{for all } \alpha > 0.$$

This implies

$$\Big\langle \frac{F(x + \alpha d) - F(x)}{\alpha}, d \Big\rangle \geq 0 \quad \text{for all } \alpha > 0.$$

Taking limits (which exist because f is twice differentiable), we obtain $\langle G(x)\, d, d\rangle \geq 0$, where

$$G(x) = \nabla^2 f = \begin{bmatrix} \partial_1^2 f & \partial_2\partial_1 f & \dots & \partial_n\partial_1 f \\ \vdots & & \ddots & \\ \partial_1\partial_n f & \partial_2\partial_n f & \dots & \partial_n^2 f \end{bmatrix}$$

($\Leftarrow$) Conversely, assume that $G(x) = \nabla F$ is positive semi-definite. We will show that $F = \nabla f$ is monotone. Notice that

$$F(x+d) - F(x) = \int_0^1 G(x+\alpha d)\, d\, d\alpha.$$

This implies

$$\langle F(x+d) - F(x), d\rangle = \int_0^1 \underbrace{\langle G(x+\alpha d)\, d, d\rangle}_{\geq 0} \geq 0.$$

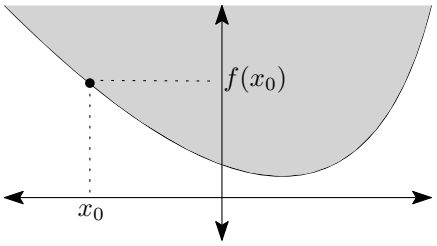
Therefore F is monotone. □

3 Conjugate Functions

This chapter introduces the notion of a conjugate function, along with some basic properties.

Suppose $\text{epi}(f)$ is a closed set. We will now consider a dual representation of this set.

Consider a point in $\text{epi}(f)$: $(x_0, f(x_o)) \in \mathbb{R}^{n+1}$, where $f : \mathbb{R}^n \to \mathbb{R}$.



Notice that this point is on the boundary of $\text{epi}(f)$. Therefore we can find a point (z, c) such that

$$\langle z, x_0\rangle + cf(x_0) \geq \langle z, y\rangle + cr \quad \text{for all } (y, r) \in \text{epi}(f). \tag{3.2}$$

Notice that here $c \leq 0$ since if $y \in \text{dom}(f)$, r can be arbitrarily large.

Now, if $c \neq 0$, we can find s, and r such that $f_{s,r}(x) = \langle s, x\rangle + r$ minorizes $f(x)$ at x_0. That is,

$$\begin{aligned} f_{s,r}(x_0) &= f(x_0) \\ f_{s,r}(x) &\leq f(x) \end{aligned}$$

To be concrete, we obtain s and r as follows.

$$\begin{aligned} & \langle z/c, x_0\rangle + f(x_0) \leq \langle z/c, x\rangle + f(x) \\ \iff \quad & \underbrace{\langle -z/c}_{s}, x\rangle + \underbrace{\langle z/c, x_0\rangle + f(x_0)}_{r} \leq f(x) \end{aligned}$$

The remaining question is : can we always find a (z, c) pair with $c \neq 0$ such that (3.2) holds?

Fortunately, the answer is yes, and it is easier to see if we assume $\text{dom}(f) = \mathbb{R}^n$. Note that, in this case, if $(z, c) \neq (0, 0)$ and (3.2) holds, then $c = 0$ implies that

$$\langle z, x_0\rangle \geq \langle z, y\rangle \quad \text{for all } y \in \mathbb{R}^n.$$

But this inequality is not valid for $y = 2z\|x_0\|_2/\|z\|_2$. Thus, we must have $c \neq 0$. The general case is considered below.

Lemma 1. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is closed, convex. Then, there exist (z, c) with $z \neq 0$, $c \neq 0$ such that (3.2) holds.

Proof. To see that we can find a pair (z, c) with $c \neq 0$, let $x_k \in \text{relint}(\text{dom}(f))$ with $x_k \to x$. Then, we can find (s_k, c_k) such that

$$\langle s_k, x_k \rangle + c_k f(x_k) \leq \langle s_k, y \rangle + c_k f(y).$$

Here, if $c_k = 0$, then

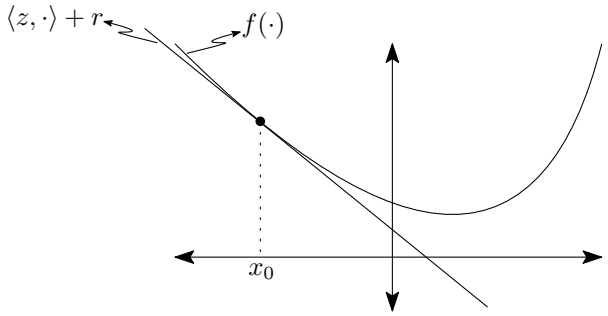
$$\langle s_k, x_k - y \rangle \leq 0, \quad \forall y \in \text{dom}(f).$$

But since $x_k \in \text{relint}(\text{dom}(f))$, $x_k + \alpha(x_k - y) \in \text{dom}(f)$ for sufficiently small $\alpha > 0$. This implies $\langle s_k, y - x_k \rangle \leq 0$, which is a contradiction (here, we assume that s_k is included in the subspace parallel to aff(dom(f))). Therefore, $c_k \neq 0$. □

The foregoing discussion implies that, for a closed convex $f : \mathbb{R}^n \to \mathbb{R}$, given any $x_0 \in \mathbb{R}^n$, we can find $z \in \mathbb{R}^n$ and $r \in \mathbb{R}$ such that the following two conditions hold:

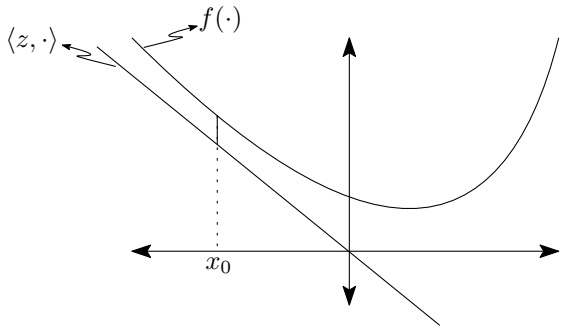
$$\langle z, x_0 \rangle + r = f(x_0)$$
$$\langle z, x \rangle + r \leq f(x).$$

This is depicted below.



Notice that there is a maximum value of r, associated with a z, so that the above two conditions are valid. How can we find this value?

Consider the following figure.



In order to find the maximum r, we can look for the minimum vertical distance between $f(x)$ and $\langle z, x \rangle$. That is, we set

$$r = \inf_x f(x) - \langle z, x \rangle$$

Observe that this definition implies the two conditions in (3.3).

In order to work with sup rather than inf, and to emphasize the dependence on z, and we define

$$f^*(z) = -r = \sup_x \langle z, x\rangle - f(x).$$

The function $f^*(z)$ is called the Fenchel conjugate of f, and thanks to the supremum operation, it is convex with respect to z – note that in the definition of f^*, x acts like an index. In addition to convexity, $f^*\!*$ is also closed, because its epigraph is the intersection of closed sets (half spaces).The following inequality follows immediately from the definition.

Proposition 25 (Fenchel's inequality)**.** For a closed convex $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$f(x) + f^*(z) \leq \langle z, x\rangle, \quad \text{for all } x, z.$$

Consider now the conjugate of the conjugate:

$$f^{**}(x) = \sup_z \langle z, x\rangle - f^*(z).$$

Since the function $g(z) = \langle z, x\rangle - f^*(z)$ minorizes $f(x)$, we have $f^{**}(x) \leq f(x)$. However, by the previous discussion, we also know that for any x^*, there is a pair (z^*, r^*) such that $\langle z^*, x\rangle + r^* = f(x^*)$. Therefore, we deduce that $f^{**}(x) = f(x)$. Thus, we have shown the following.

Proposition 26. If $f : \mathbb{R}^n \to \mathbb{R}$ is a closed convex function, then

$$f(x) = \sup_z \langle z, x\rangle - f^*(z).$$

Example 14. Let $f(x) = \frac{1}{2} x^T Q\, x$, where Q is positive definite. Then,

$$f^*(z) = \sup_x \langle z, x\rangle - \frac{1}{2} x^T Q\, x.$$

Maximum is achieved at x^* such that $z = Q\, x^*$. Plugging in $x^* = Q^{-1}\, z$, we obtain

$$\begin{aligned} f^*(z) &= z^T Q^{-1} z - \frac{1}{2} z^T Q^{-T} Q\, Q^{-1} z \\ &= \frac{1}{2} z^T Q^{-1}\, z. \end{aligned}$$

◇

Example 15. Let C be a convex set. The indicator function is defined as

$$i_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ \infty, & \text{if } x \notin C. \end{cases}$$

The conjugate

$$\sigma_C(z) = \sup_x \langle z, x\rangle - i_C(x) = \sup_{x \in C} \langle z, x\rangle.$$

◇

Some Properties of the Conjugate Function

(i) If $g(x) = f(x - s)$, then

$$\begin{aligned} g^*(z) &= \sup_x \langle z, x\rangle - f(x - s) \\ &= \sup_y \langle z, y + s\rangle - f(y) \\ &= f^*(z) + \langle z, s\rangle. \end{aligned}$$

(ii) If $g(x) = t\, f(x)$, with $t > 0$, then

$$\begin{aligned} g^*(z) &= \sup_x .\langle z, x\rangle - tf(x) \\ &= t \sup_x \langle z/t, x\rangle - f(x) \\ &= tf^*\left(\frac{z}{t}\right). \end{aligned}$$

(iii) If $g(x) = f(tx)$, then

$$\begin{aligned} g^*(z) &= \sup_{y=tx} \langle z, y/t\rangle - f(y) \\ &= f^*\left(\frac{z}{t}\right). \end{aligned}$$

(iv) If A is invertible, and $g(x) = f(Ax)$, then

$$\begin{aligned} g^*(z) &= \sup_x \langle z, x\rangle - f(Ax) \\ &= \sup_y \langle A^{-T} z, y\rangle - f(y) \\ &= f^*(A^{-T}\, z). \end{aligned}$$

(v) If $g(x) = f(x) + \langle s_0, x\rangle$, then

$$\begin{aligned} g^*(z) &= \sup_x \langle z - s_0, x\rangle - f(x) \\ &= f^*(z - s_0). \end{aligned}$$

(vi) If $g(x_1, x_2) = f_1(x_1) + f_2(x_2)$, then

$$\begin{aligned} g^*(z_1, z_2) &= \sup_{x_1, x_2} \langle z_1, x_1\rangle + \langle z_2, x_2\rangle - f(x) \\ &= f^*(z_1) + f^*(z_2). \end{aligned}$$

(vii) If $g(x) = f_1(x) + f_2(x)$, then

$$\begin{aligned} g^*(z) &= \sup_x \langle z, x\rangle - f_1(x) - \left[\sup_y \langle y, x\rangle - f_2^*(y)\right] \\ &= \sup_x \inf_y \langle z, x\rangle - f_1(x) - \langle y, x\rangle + f_2^*(y). \end{aligned}$$

We will later dicsuss that whenever there is a 'sadlle point', we can exchange the order of inf and sup. Doing so, we obtain

$$\begin{aligned} g^*(z) &= \inf_y \sup_x \langle z - y, x\rangle - f_1(x) + f_2^*(y) \\ &= \inf_y f_1^*(z - y) + f_2^*(y). \end{aligned}$$

The operation that appears in the last line is called infimal convolution.

(viii) More generally, if $g(x) = f_1(x) + f_2(x) + \ldots + f_n(x)$, then

$$g^*(z) = \inf_{\substack{z_1,\ldots,z_n \\ \text{s.t. } z=z_1+\ldots+z_n}} f_1^*(z_1) + \ldots + f_n^*(z_n).$$

Example 16. The last property will be useful when we consider multiple constraints. In particular, let $C = A \cap B$, where A, B are convex sets. Then we have that

$$\sigma_C(x) = \sup_{z\in C}\langle z, x\rangle = \inf_y \sigma_A(x - y) + \sigma_B(y).$$

To see this, note that $\sigma^*(z) = i_C(z)$. But $i_{A\int B}(z) = i_A(z) + i_B(z)$. Computing the conjugate, we obtain

$$i_C^*(x) = \sigma_C(x) = \inf_y \sigma_A(x - y) + \sigma_B(y).$$

$\diamond$

4 Duality

This chapter introduces the notion of duality. We start with a general discussion of duality. We then pass to Lagrangian duality, and finally consider the Karush-Kuhn-Tucker conditions of optimality.

4.1 A General Discussion of Duality

Consider a minimization problem like

$$\min_{x \in C} f(x), \text{ where } C \text{ is a closed convex set.}$$

Suppose there exists a function $K(x, \lambda)$, which is

(i) convex for fixed λ, as a function of x,

(ii) concave for fixed x, as a function of λ,

$$\max_{\lambda \in D} K(x, \lambda) = \begin{cases} f(x), \text{if } x \in C, \\ \infty, \text{if } x \notin C, \end{cases}$$

for some closed convex set D.

Example 17. Consider the problem

$$\min_x \frac{1}{2} \|y - x\|_2^2 + \|x\|_2, \tag{4.4}$$

for $x, y \in C = \mathbb{R}^n$. Here, by Cauchy-Schwarz inequality, we can write

$$\|x\|_2 = \max_{\lambda \in B_2} \langle x, \lambda \rangle,$$

where B_2 is the unit ℓ_2 ball of $\mathbb{R}^n$. In this case, the problem (4.4) is equivalent to

$$\min_x \max_{\lambda \in B_2} \underbrace{\frac{1}{2} \|y - x\|_2^2 + \langle y, x \rangle}_{K(x,\lambda)}.$$

◇

In short, "$\min_{x \in C} f(x)$" is equivalent to

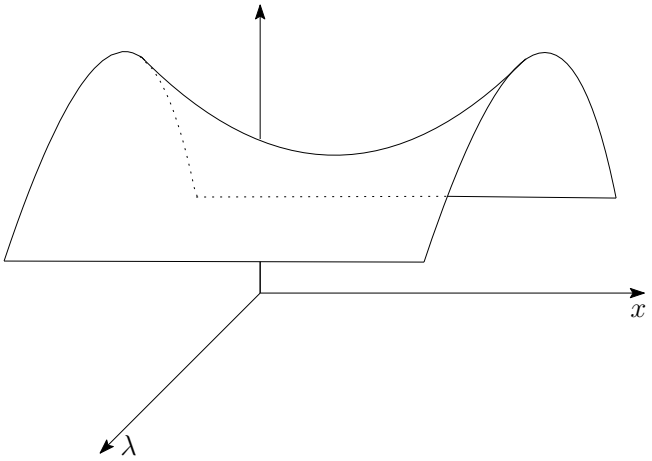
$$\min_{x \in C} \max_{\lambda \in D} K(x, \lambda).$$

We call (x^*, λ^*) a solution of this problem if

$$K(x^*, \lambda) \leq K(x^*, \lambda^*) \leq K(x, \lambda^*) \quad \text{for all } x \in C, \lambda \in D.$$

If such a (x^*, λ^*) exists, it is called a saddle point of (x, λ).

See below for a depiction.



This approach is useful especiallly if for fixed λ, $K(x, \lambda)$ is easy to minimize with respect to x. In that case, if $\lambda = \lambda^*$, then minimizing $K(x, \lambda^*)$ over $x \in C$ is equivalent to solving the problem.

Example 18. Suppose λ is fixed. Then to maximize

$$\frac{1}{2}\|y - x\|_2^2 + \langle \lambda, x \rangle,$$

set the gradient to zero. This gives

$$x - y + \lambda = 0 \quad \Leftrightarrow \quad x = y - \lambda.$$

$\diamond$

The question now is, how do we obtain λ^*? For that, define the function

$$g(\lambda) = \min_{x \in C} K(x, \lambda).$$

Notice that

$$g(x, \lambda) \leq K(x, \lambda) \quad \text{for all } x \in C.$$

Consider now the problem

$$\max_{\lambda \in D} g(\lambda). \tag{4.5}$$

Exercise 9. Show that $g(\lambda)$ is concave for $\lambda \in D$. $\diamond$

Now let $\bar{\lambda} = \max_{\lambda \in D} g(\lambda)$. Also, let $\bar{x} = \arg\min_{x \in C} K(x, \bar{\lambda})$. Then,

$$K(\bar{x}, \bar{\lambda}) \leq K(x, \bar{\lambda}) \text{ for } x \in C,$$

and

$$g(\bar{\lambda}) = K(\bar{x}, \bar{\lambda}) \geq g(\lambda) \geq K(\bar{x}, \lambda) \text{ for } \lambda \in D.$$

Combining these, we obtain

$$K(\bar{x}, \lambda) \leq K(\bar{x}, \bar{\lambda}) \leq K(x, \bar{\lambda}) \text{ for } x \in C, \lambda \in D.$$

Therefore, $(\bar{x}, \bar{\lambda})$ is a saddle point, and $\bar{x}$ solves the problem

$$\min_{x \in C} f(x).$$

We have shown that *if a saddle point exists*, we can obtain it either by solving a $\min - \max$ or a $\max - \min$ problem.

Here, (4.4) is called the primal problem and (4.5) is called the dual problem. Note that there might be different dual problems depending on how we choose $K(x, \lambda)$. Notice also that if a saddle point exists, we have

$$d^* = \max_{\lambda \in D} g(\lambda) = \min_{x \in C} f(x) = p^*.$$

Example 19. Consider the problem

$$\min_x \frac{1}{2} \|y - x\|_2^2 + \|x\|_2.$$

An equivalent problem is

$$\min_x \max_{\lambda \in B_2} \frac{1}{2} \|y - x\|_2^2 + \langle x, \lambda \rangle,$$

where B_2 is the unit ball of the ℓ_2 norm. Let us define

$$g(\lambda) = \min_x \frac{1}{2} \|y - x\|_2^2 + \langle x, \lambda \rangle.$$

Carrying out the minimization, we find

$$\begin{aligned} g(\lambda) &= \frac{1}{2} \|\lambda\|_2^2 + \langle y - \lambda, \lambda \rangle \\ &= -\frac{1}{2} \|y - \lambda\|_2^2 + c, \end{aligned}$$

for some constant c. Therefore, the dual problem is

$$\max_{\lambda \in B_2} -\|y - \lambda\|_2^2.$$

Or, equivalently

$$\min_{\lambda \in B_2} \|y - \lambda\|_2^2.$$

The minimizer is the projection of y onto B_2, and is given by

$$\lambda^* = \begin{cases} y, & \text{if } \|y\|_2 \leq 1, \\ y/\|y\|_2, & \text{if } 1 < \|y\|_2. \end{cases}$$

The solution of the primal problem is

$$x^* = y - \lambda^* = y - P_{B_2}(y) = \begin{cases} 0, & \text{if } \|y\|_2 \leq 1, \\ y\,\dfrac{\|y\|_2 - 1}{\|y\|_2}, & \text{if } 1 < \|y\|_2. \end{cases}$$

$\diamond$

Notice that this discussion depends heavily on the existence of a saddle point. However, even if such a point does not exist, we can define a dual problem, but in this case, the maximum of the dual d^* and the minimum of the primal problem p^* are not necessarily equal. Instead, we will have $d^* \leq p^*$. To see this, note that

$$g(\lambda) \leq K(x, \lambda) \leq f(x) \quad \text{for all } x, \lambda.$$

Therefore,

$$d^* = g(\lambda^*) \leq K(x, \lambda^*) \leq f(x^*) = p^*.$$

Therefore, $p^* - d^*$ is always nonnegative. This difference is called the duality gap.

The following proposition is a summary of the foregoing discussion. We provide a proof for the sake of completeness.

Proposition 27. Let $K(x, \lambda)$ be convexoconcave, and define

$$f(x) = \sup_{\lambda} K(x, \lambda)$$
$$g(\lambda) = \inf_{x} K(x, \lambda).$$

Then,

$$K(x^*, \lambda) \leq K(x^*, \lambda^*) \leq K(x, \lambda^*) \tag{4.6}$$

if and only if

$$x^* \in \arg\min_{x} f(x), \tag{4.7}$$
$$\lambda^* \in \arg\max_{\lambda} g(\lambda),$$
$$\inf_{x} \sup_{\lambda} K(x, \lambda) = \sup_{\lambda} \inf_{x} K(x, \lambda).$$

Proof. ($\Rightarrow$) Suppose (4.6) holds. Then, since $K(x^*, \lambda^*) = f(x^*) = g(\lambda^*)$, and since $f(x) \geq g(\lambda)$ for arbitrary x, λ, it follows that (4.7) holds too.

($\Leftarrow$) Suppose (4.7) holds. Then, by the last equality in (4.7), we have $f(x^*) = g(\lambda^*)$. Now by the definition of $f(\cdot)$, we have

$$f(x^*) \geq K(x^*, \lambda), \text{ for any } \lambda.$$

Similarly, by the definition of $g(\cdot)$, we obtain

$$g(\lambda^*) \leq K(x, \lambda^*) \text{ for any } x.$$

Using the definition of f, and g once again, we can write,

$$f(x^*) \geq K(x^*, \lambda^*) \geq g(\lambda^*).$$

Since $f(x^*) = g(\lambda^*)$, we therefore obtain the desired inequality (4.6) by combining these inequalities. □

4.2 Lagrangian Duality

We now discuss a specific dual, associated with a constrained minimization problem.

$$\min_x f(x) \text{ subject to } \begin{cases} g_1(x) & \leq 0, \\ g_2(x) & \leq 0, \\ & \vdots \\ g_m(x) & \leq 0, \end{cases} \tag{4.8}$$

where all of the functions are closed, convex, and defined on $\mathbb{R}^n$.

In this setting, we define the Lagrangian function as

$$L(x, \lambda) = \begin{cases} f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x) + \ldots \lambda_m g_m(x), \text{if } \lambda_i \geq 0 \\ -\infty, \text{ if at least one } \lambda_i \leq 0. \end{cases}$$

Notice that

$$\max_{\lambda \geq 0} L(x, \lambda) = \begin{cases} f(x), & \text{if } g_i(x) \leq 0, \text{ for all } i, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, (4.8) is equivalent to

$$\min_x \max_{\lambda \geq 0} L(x, \lambda).$$

Also, notice that for fixed x, $L(x, \lambda)$ is concave (in fact affine), with respect to λ. It follows from the previous discussion that if (x^*, λ^*) is a saddle point of $L(x, \lambda)$ if x^* solves (4.8), or

$$\begin{aligned}\lambda^* &= \arg\min_x \; L(x, \lambda^*) \\ &= \arg\min_x \; f(x) + \lambda_1^* \, g_1(x) + \lambda_2^* \, g_2(x) + \ldots + \lambda_m^* \, g_m(x).\end{aligned}$$

Notice that the problem is transformed into an unconstrained problem, with the help of λ^*. To obtain λ^*, we consider the Lagrangian dual, defined as

$$g(\lambda) = \min_x L(x, \lambda).$$

The dual problem is,

$$\max_{\lambda \geq 0} g(\lambda).$$

If a saddle point exists, λ^* is the solution of this dual problem. In that case, we obtain a minimizer (which need not be unique) as

$$x \in \arg\min_x L(x, \lambda^*).$$

Example 20. Consider the problem

$$\min x \text{ s.t. } x^2 + 2x \leq 0.$$

The dual function is

$$g(\lambda) = \min_x \underbrace{x + \lambda(x^2 + 2x)}_{L(x,\lambda) \text{ for } \lambda \geq 0}.$$

At the minimum we must have

$$1 + \lambda \, (2x + 2) = 0 \iff -\frac{1}{2\lambda} - 1 = x.$$

Plugging in, we obtain,

$$\begin{aligned}g(\lambda) &= -\frac{1}{2\lambda} - 1 + \lambda \left(\frac{1}{4\lambda^2} + 1 + \frac{1}{\lambda} - \frac{1}{\lambda} - 2 \right) \\ &= -\lambda - \frac{1}{4\lambda} - 1.\end{aligned}$$

λ^* satisfies

$$-1 + \frac{1}{(2\lambda^*)^2}.$$

Solving for λ^*, and taking the positive root (since $\lambda^* \geq 0$), we find $\lambda^* = \frac{1}{2}$. Therefore,

$$x^* = \arg\min_x \; x + \frac{1}{2}(x^2 + 2x),$$

which can be easily solved, to give $x^* = -2$. Note that we can see this easily if we sketch the constraint function. $\diamond$

So far, the discussion relied on the assumption that the Lagrangian has a saddle point, so that the duality gap is zero. A natural question is to ask when this can be guaranteed. The conditions that ensure the existence of a saddle point are called constraint qualification. A simple one to state is Slater's condition.

Proposition 28 (Slater's condition)**.** Suppose $f(\cdot)$ and $g_i(\cdot)$ for $i = 1, 2, \ldots, n$ are convex functions. Consider the problem

$$\min_x f(x), \text{ s.t. } g_i(x) \leq 0, \quad \forall i. \tag{4.9}$$

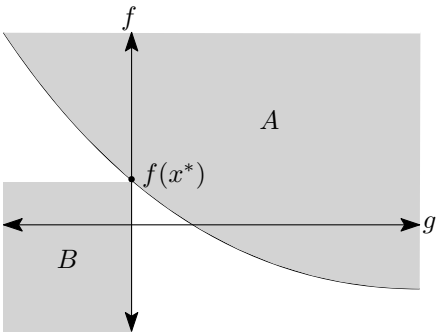
Suppose there exists $\bar{x}$ such that $g_i(\bar{x}) \leq 0$ for all i, and $g_j(x) < 0$ for some j. Then, x^* solves (4.9) if and only if there exists $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of the function

$$L(x, \lambda) = \begin{cases} f(x) + \sum_{i=1}^{n} \lambda_i \, g_i(x), & \text{if } \lambda_i \geq 0 \text{ for all } i, \\ -\infty, & \text{otherwise.} \end{cases}$$

Proof. For simplicity, we assume that $n = 1$, so there is only one constraint function $g(\cdot)$.

Assume that x^* solves (4.9). Consider the sets

$$A = \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \geq \begin{pmatrix} f(x) \\ g(x) \end{pmatrix} \text{ for some } x \right\},$$
$$B = \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} < \begin{pmatrix} f(x^*) \\ 0 \end{pmatrix} \right\}.$$



It can be shown that both A and B are convex (show this!). Further, we have $A \cap B = \emptyset$ (show this!). Therefore, there exists μ_1, μ_2 such that

$$\mu_1 \, z_1 + \mu_2 \, z_2 \leq \mu_1 \, t_1 + \mu_2 \, t_2 \text{ for all } z \in B, t \in A.$$

Note here that $\mu_2 \geq 0$ since z_2 can be taken as small as desired. Similarly, $\mu_1 \geq 0$, since z_1 can be taken as small as desired. Also notice that $\mu_1 \neq 0$, since otherwise we would have $0 \leq g(x)$ for all x, but we already know that $g(\bar{x}) < 0$. Therefore, for $\lambda^* = \mu_2/\mu_1$, we can write

$$z_1 + \lambda^* z_2 \leq 0 \cdot t_1 + \lambda^* \, t_2, \text{ for all } z \in B, \, t \in A.$$

Consequently,

$$f(x^*) \leq f(x) + \lambda^* g(x) \text{ for all } x \text{ and } \lambda^* \geq 0.$$

In particular, we have that $f(x^*) \leq f(x^*) + \lambda^* g(x^*)$. Since $g(x^*) \leq 0$, and $\lambda^* \geq 0$, it follows that $\lambda^* g(x^*) = 0$. So, we have

$$f(x) + \lambda^* g(x) \geq f(x^*) = f(x^*) + \lambda^* g(x^*) \geq f(x^*) + \lambda g(x^*) \text{ for all } \lambda \geq 0.$$

Thus, (x^*, λ^*) is a saddle point. □

4.3 Karush-Kuhn-Tucker (KKT) Conditions

Suppose now that Slater's conditions are satisfied so that x^* solves the problem so that (x^*, λ^*) is a saddle point of the Lagrangian $L(x, \lambda)$. Notice that in this case, x^* is a minimizer for the problem,

$$\min_x f(x) + \lambda_1^* g_1(x) + \lambda_2^* g_2(x) + \ldots + \lambda_m^* g_m(x)$$

If all of the functions above are differentiable, we can write

$$\nabla f(x^*) + \lambda_1^* \nabla g_1(x^*) + \lambda_2^* \nabla g_2(x^*) + \ldots + \lambda_m^* \nabla g_m(x^*) = 0.$$

But since $\lambda_i^* \geq 0$, we have that if $g_i(x^*) < 0$, then we must have $\lambda_i^*. = 0$. Therefore, $\lambda_i^* g_i(x^*) = 0$ for all i. Collected together, these conditions are known as KKT conditions.

$$\begin{aligned}
&\lambda_i^* \geq 0, \\
&g_i(x^*) \leq 0, \\
&\lambda_i^* g_i(x^*) = 0, \quad \text{('Complementary slackness')} \\
&\nabla f(x^*) + \sum_i \lambda_i^* \nabla g_i(x^*) = 0. \qquad (4.10a)
\end{aligned}$$

By the above discussion these conditions are necessary for optimality. It turns out that they are also sufficient conditions. To see this, first observe that (4.10a) implies

$$g(\lambda^*) = L(x^*, \lambda^*).$$

But since $\lambda_i^* g_i(x^*) = 0$, we have

$$L(x^*, \lambda^*) = f(x^*) + \sum_i \lambda_i^* g_i(x^*) = f(x^*).$$

Therefore, $g(\lambda^*) = L(x^*, \lambda^*) = f(x^*)$, i.e., (x^*, λ^*) is a saddle point of $L(x, \lambda)$. Thus x^* solves the primal problem.

5 Subdifferentials

A convex function does not have to be differentiable. However, even when it is not differentiable, there is considerable structure in how it varies locally. Subdifferentials generalize derivatives (or gradients) and capture this structure for convex functions. In addition, they enjoy a certain calculus, which proves very useful in deriving minimization algorithms. We introduce, and discuss some basic properties of subdifferentials in this chapter. We start with some motivations underlying definitions and basic properties. We then explore connections with KKT conditions. Finally, we dicsuss the notion of monotonicity, a fundamental property of subdifferentials.

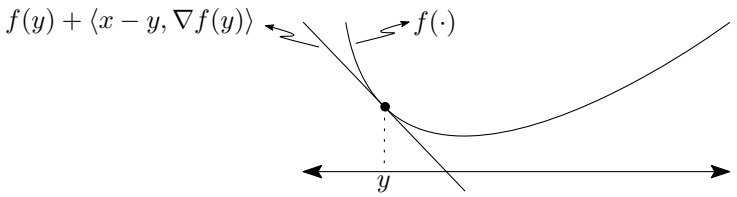
5.1 Motivation, Definition, Properties of Subdifferentials

Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and convex, then

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

for all x, y. In fact, $s = \nabla f(y)$ is the unique vector that satisfies the inequality below

$$f(x) \geq f(y) + \langle x - y, s \rangle \quad \text{for all } x, y. \tag{5.11}$$



This useful observation has the shortcoming that it requires f to be differentiable. In general, a convex function may not be differentiable – consider for instance $f(x) = |x|$. We define the subdifferential by making use of the inequality (5.11).

Definition 19. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. The subdifferential of f at y is the set of s that satisfy

$$f(x) \geq f(y) + \langle x - y, s \rangle \quad \text{for all } x.$$

This set is denoted by $\partial f(y)$. ◇

Since for every $y \in \text{dom}(f)$, we can find r, s, such that

(i) $r + \langle s, y \rangle = f(y)$,

(ii) $r + \langle s, y\rangle \leq f(x)$,

it follows that $\partial f(y)$ is non-empty. $\partial f(\cdot)$ has other interesting properties as well.

Proposition 29. For every $y \in \mathrm{dom}(f)$, $\partial f(y)$ is a convex set.

Proof. Suppose $(s_1,\ s_2) \in \partial f(y)$. Then, we have

$$f(y) + \langle x - y, s_1\rangle \leq f(x)$$
$$f(y) + \langle x - y, s_2\rangle \leq f(x).$$

Taking a convex combination of these inequalities, we find

$$f(y) + \langle x - y, \alpha s_1 + (1-\alpha)s_2\rangle \leq f(x).$$

□

Example 21. Let $f(x) = |x|$. Then

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

◇

Notice the example above, wherever the function is differentiable, the subdifferential coincides with the gradient. This holds in general.

Proposition 30. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

Proof. Suppose $s \in \partial f(x)$. In this case, for any $d \in \mathbb{R}^n$, $t \in \mathbb{R}$, we have

$$f(x + t\,d) \geq f(x) + \langle s, t\,d\rangle \iff \langle s, d\rangle \leq \frac{f(x+td) - f(x)}{t} \quad \text{for all } t, d.$$

Now let $t \to 0$. The inequality above implies

$$\langle s, d\rangle \leq \langle \nabla f(x), d\rangle, \quad \text{for all } d. \tag{5.12}$$

Now notice that

$$\langle s, -d\rangle \leq \frac{f(x-td) - f(x)}{t} \quad \text{for all } t, d.$$

Therefore,

$$\langle s, -d\rangle \leq \langle \nabla f(x), -d\rangle, \quad \text{for all } d. \tag{5.13}$$

Taken together, (5.12), (5.13) imply $\langle s, d\rangle = \langle \nabla f, d\rangle$ for all d. Thus, $s = \nabla f(x)$. □

The subdifferential can be used to characterize the minimizers of convex functions.

Proposition 31. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Then,

$$x^* \in \arg\min_x f(x)$$

if and only if

$$0 \in \partial f(x^*).$$

Proof. This follows from the equivalences

$$\begin{aligned} & x^* \in \arg\min_x f(x) \\ \iff \quad & f(x^*) \leq f(x) \quad \text{for all } x, \\ \iff \quad & f(x) \leq f(x^*) + \langle 0, x - x^* \rangle \quad \text{for all } x, \\ \iff \quad & 0 \in \partial f(x^*). \end{aligned}$$

□

Proposition 32. Let f, $g : \mathbb{R}^n \to \mathbb{R}$ be convex functions. Also let $h = f + g$. If $x \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$, then

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

Proof. Let $s_1 \in A$, $s_2 \in B$. Then,

$$\big(f(x) + \langle y - x, s_1 \rangle\big) + \big(g(x) + \langle y - x, s_2 \rangle\big) \leq f(y) + g(y).$$

Therefore $s_1 + s_2 \in C$. Since s_1 and s_2 are arbitrary members of A, B, it follows that $A + B \subset C$.

For the converse (i.e., $C \subset A. + B$), we need to show that any $z \in \partial h(x)$ we can find $u \in \partial f(x)$ such that $z - u \in \partial g(x)$. This is equivalent to saying that, we can find $u \in \partial f(x)$, $v \in \partial g(x)$ such that $z = u + v$. We take a detour to show this result. □

Let us now study the link between conjugate functions and subdifferentials. This will complete the remaining part of Prop. 5.1.

Proposition 33. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a closed, convex function and f^* denote its conjugate. Then $s \in \partial f(x)$ if and only if

$$f(x) + f^*(x) = \langle s, x \rangle. \tag{5.14}$$

Proof. Since $f(x) = \sup_z \langle z, x\rangle - f^*(z)$, we have

$$f(x) + f^*(x) \geq \langle z, x\rangle, \quad \text{for all } z, x. \tag{5.15}$$

Consider now the following chain of equivalences

$$\begin{aligned} & s \in \partial f(x) \\ \iff \quad & f(x) - \langle s, x\rangle \leq f(y) - \langle s, y\rangle \quad \text{for all } y \\ \iff \quad & f(x) - \langle s, x\rangle \leq \inf_y f(y) - \langle s, y\rangle = -\sup_y \langle s, y\rangle - f(y) = -f^*(s) \end{aligned} \tag{5.16}$$

Combining the two inequalities (5.15), (5.16), we obtain (5.14). □

Corollary 4. $x \in \partial f^*(s)$ if and only if $f(x) + f^*(s) = \langle x, s\rangle$.

Corollary 5. $0 \in \partial f(x)$ if and only if $x \in f^*(0)$.

Proof. Consider the following chain of equivalences.

$$0 \in \partial f(x) \quad \iff \quad f(x) + f^*(x) = \langle 0, x\rangle \quad \iff \quad x \in \partial f^*(x).$$

□

Example 22. Recall the definition of a support function : for a closed convex set C, let $\sigma_C(x) = \sup_{z\in C} \langle z, x\rangle$. Recall also that $\sigma_C^*(s) = i_C(s)$. Therefore,

$$\begin{aligned} s \in \partial\sigma_C(x) \quad & \iff \quad \sigma_C(x) + i_C(s) = \langle s, x\rangle \\ & \iff \quad \sigma_C(x) = \langle s, x\rangle, \quad \forall s \in C. \end{aligned}$$

In words, $\partial\sigma_C(x)$ is the set of $s \in C$ for which $\sigma_C(x) = \langle s, x\rangle$.

[Insert Fig. on p37] ◇

Example 23. Consider now a closed convex set C, and let us obtain a description of the subdifferential of the characteristic function of C at x, i.e., $\partial i_C(x)$.

Observe that

$$s \in \partial i_C(x) \quad \iff \quad i_C(y) \geq i_C(x) + \langle s, y - x\rangle, \quad \text{for all } y.$$

If $x \notin C$, the inequality is not satisfied for any s. In this case, $\partial i_C(x) 0 \emptyset$.

If $x \in C$, then there are two cases to consider

(i) If $y \notin C$, then $i_C(y) = \infty$, and the inequality is always satisfied.

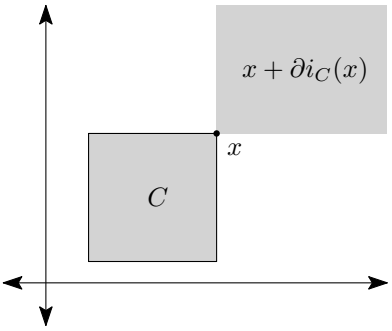
(ii) If $y \in C$, then

$$0 \geq \langle s, y - x\rangle, \quad \forall y \in C \quad \iff \quad s \in N_C(x).$$

In summary,

$$\partial i_C(x) = \begin{cases} N_C(x), & \text{if } x \in C, \\ \emptyset, & \text{if } x \notin C. \end{cases}$$

See below for when C is a rectangular set.



◇

We now go back to Prop. 5.1, and complete the proof. Recall that what remains is to show that $\partial h(x) \subset \partial f(x) + \partial g(x)$, where $h(x) = f(x) + g(x)$ for convex functions $f(x)$, $g(x)$.

Proof of Prop. cont'd: Let $u \in \partial h(x)$. This implies,

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

Plugging in the definition of h, we can write

$$f(x) + g(x) + \inf_z \underbrace{f^*(z) + g^*(u - z)}_{m} = \langle x, u \rangle.$$

Suppose the infimum of m is achieved for when $z = s$.

$$\Big[f(x) + f^*(x) - \langle x, s \rangle \Big] + \Big[g(x) + g^*(u - s) - \langle u - s - x \rangle \Big] = 0$$

Both terms in square bracksts are non-negative by Fenchel's inequality. Therefore, for equality to hold, we need both terms to be zero. Thus, we can write

$$\begin{aligned} f(x) + f^*(s) &= \langle x, s \rangle \\ g(x) + g^*(u - s) &= \langle u - s, x \rangle. \end{aligned}$$

Thus,

$$u = \underbrace{s}_{\in \partial f(x)} + \underbrace{u - s}_{\in \partial g(x)} \in \partial f(x) + \partial g(x).$$

Since u was arbitrary, this implies $\partial h(x) \subset \partial f(x) + \partial g(x)$. □

From this dicsussion, we obtain the following corollary concerning constrained minimization.

Corollary 6. Consider the problem

$$\min_{x \in C} f(x),$$

where f is a convex function, C is a convex set. x^* is a solution of this problem if and only if

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

5.2 Connection with the KKT Conditions

Let us now study what the condition stated in Corollary 6 means. Let us consider the problem

$$\min_x \ f(x) \text{ subject to } \begin{cases} g_1(x) \leq 0, \\ g_2(x) \leq 0, \end{cases}$$

where g_i are convex and differentiable. Let $C_i = \{x : g_i(x) \leq 0\}$. Note that both C_i's are convex sets. Suppose $g_i(x) < 0$. Then $N_{C_i}(x) = \{0\}$. However, if $g_i(x) = 0$, then

$$N_{C_i}(x) = \{s : \langle s, y - x \rangle \leq 0 = g_i(x) \text{ for all } y \text{ with } g_i(y) \leq g_i(x) = 0\}.$$

That is, $N_{C_i}(x) = \alpha \nabla g_i(x)$, where $\alpha \geq 0$. Also, if $g_i(x) > 0$, then $N_{C_i}(x) = \emptyset$. Therefore, the condition

$$0 \in \nabla f(x) + N_{C_1}(x) + N_{C_2}(x)$$

is equivalent to

$$0 = \nabla f(x) + \alpha_1 \nabla g_1(x) + \alpha_2 g(x) \text{ where } \begin{cases} \alpha_i \geq 0, \\ \alpha_i \, g_i(x) = 0, \\ g_i(x) \leq 0. \end{cases}$$

These are precisely the KKT conditions.

5.3 Monotonicity of the Subdifferential

Recall that for a convex differentiable f, we had

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0, \text{ for all } x, y.$$

A similar property holds for the subdifferential.

Proposition 34. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex. If $s \in \partial f(x)$ and $z \in \partial f(y)$, then

$$\langle s - z, x - y \rangle \geq 0. \tag{5.17}$$

Proof. Observe that

$$\begin{aligned} s \in \partial f(x) \quad &\implies f(y) \geq f(x) + \langle s, y - x \rangle \\ z \in \partial f(y) \quad &\implies f(x) \geq f(y) + \langle z, x - y \rangle. \end{aligned}$$

Summing the inequalities, we obtain (5.17). □

Definition 20. An operator $T : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is said to be monotone if $\langle s - z, x - y \rangle \geq 0$, for all x, y, and $s \in T(x)$, $z \in T(y)$. ◇

It is useful to think of set-valued operators in terms of their graphs.

Definition 21. The graph of $T : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is the set of u, v such that $v \in T(u)$. ◇

Notice that for a convex function, the graph of ∂f is the set of (x, u) such that $x \in \mathrm{dom}(f)$ and $u \in \partial f(x)$.

A curious property satisfied by $\partial f(x)$ is 'maximal' monotonicity.

Definition 22. $T : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is said to be maximal monotone if there is no monotone operator F, such that the graph of T is a strict subset of the graph of F. ◇

Example 24. For $f(x) = |x|$, the graph of $\partial f(x)$ is shown below.

[Insert Fig on p.41]. ◇

We state the following fact without proof. See Rockafellar's book for a proof.

Proposition 35. If $T = \partial f$ for a closed convex function f, then T is maximal monotone.

Let us now revisit a minimization problem like $\min_x f(x)$, for a convex f. In terms of the subdifferential of f, this is equivalent to looking for x such that $0 \in T(x)$, where $T = \partial f$. An equivalent problem is to find x such that $x \in x + \lambda T(x)$, for $\lambda > 0$.

Definition 23. Given $S : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, S^{-1} is the operator whose graph is the set of (u, v) where $u \in S(v)$. ◇

Note that, by the foregoing discussion, $\min_x f(x)$ is equivalent to finding x such that $x = (I + \lambda T)^{-1} x$. In the following, we study the properties of $J_{\lambda T} = (I + \lambda T)^{-1}$.

Let us first write down our previous observation as a proposition.

Proposition 36. If $f(z) \leq f(x)$ for all x, then $J_{\lambda_T}(z) = z$.

Definition 24. An operator $U : \mathbb{R}^n \to \mathbb{R}^n$ is said to be firmly-nonexpansive if

$$\|U(x) - U(y)\|_2^2 + \|(I - U)(x) - (I - U)(y)\|_2^2 \leq \|x - y\|_2^2.$$

$\diamond$

Proposition 37. If T is monotone, then $(I + T)^{-1}$ is firmly non-expansive.

Proof. Note that,

$$\|(I - J)x - (I - J)y\|_2^2 = \|x - y\|_2^2 + \|Jx - Jy\|_2^2 - 2\langle Jx - Jy, x - y\rangle.$$

Therefore, if we can show that

$$\langle Jx - Jy, x - y\rangle \geq \|Jx - Jy\|_2^2, \tag{5.18}$$

we are done.

Now suppose

$$\begin{aligned} x &= u + v, \text{ with } v \in \partial T(u), \\ y &= z + t, \text{ with } t \in \partial T(z). \end{aligned}$$

Then, $J(x) = v$, and $J(y) = t$. This implies

$$\langle v - t, u + v - z - t\rangle = \langle Jx - Jy, x - y\rangle = \|v - t\|_2^2 + \langle v - t, u - z\rangle.$$

The last term is non-negative by the monotonicity of T. Also, since $\|Jx - Jy\|_2^2 = \|v - t\|_2^2$, the (5.18) follows. $\square$

Alternative Proof :

$$\begin{aligned} &T \text{ is monotone} \\ &\iff \langle x' - x, y' - y\rangle \geq 0 \quad \forall (x, y),\ (x', y') \in T, \\ &\iff \langle x' - x + y' - y, x' - x\rangle \geq \|x' - x\|_2^2 \quad \forall (x, y),\ (x', y') \in T. \end{aligned}$$

$\square$

Proposition 38. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, convex and

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \tag{5.19}$$

for any x, y pair. Then,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|_2^2.$$

Proof. Consider the function $g(s) = f\big(y+s\,(x-y)\big)$. Observe that $g(0) = f(y)$, $g(1) = f(x)$, and

$$\begin{aligned}
g(1) - g(0) &= \int_0^1 g'(s)\,ds \\
&= \int_0^1 \langle \nabla f\big(y + s\,(x-y), x-y\big)\rangle\,ds \\
&\leq \int_0^1 \big|\langle \nabla f\big(y + s\,(x-y) - \nabla f(y), x-y\big)\rangle\,\big| + \langle \nabla f(y), x-y\rangle ds \\
&\leq \int_0^1 \|\nabla f\big(y + s\,(x-y) - \nabla f(y)\|_2\,\|x-y\|_2 + \langle \nabla f(y), x-y\rangle ds \\
&\leq \int_0^1 L\,s\,\|x-y\|_2^2 + \langle \nabla f(y), x-y\rangle\,ds \\
&= \frac{L}{2}\|x-y\|_2^2 + \langle \nabla f(y), x-y\rangle.
\end{aligned}$$

where we applied the Cauchy-Schwarz inequality and (5.19) to obtain the last two inequalities. □

Now suppose we want to minimize f and it satisfies the hypothesis of Prop. 38, namely (5.19). We will derive an algorithm which will start from some initial point x_0 and produce a sequence that converges to the minimizer.

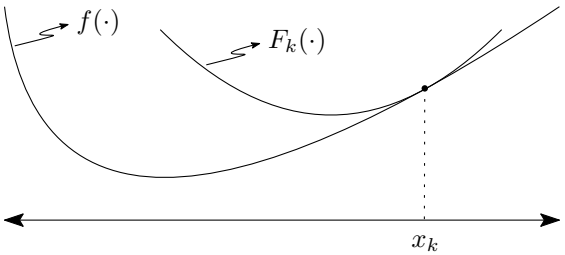
Suppose that at the k^{th} step, we have x_k, and we set x_{k+1} as

$$x_{k+1} = \arg\min_x \Big\{ F_k(x) = f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{\alpha}{2}\|x - x_k\|_2^2 \Big\}.$$

Observe that

(i) $F_k(x) = f(x_k)$,

(ii) $F_k(x) \geq f(x)$ for all x.



Therefore,

$$f(x_{k+1}) \leq F(x_{k+1}) \leq F_k(x_k) \leq f(x_k).$$

In words, at each iteration, we reduce the value of f. Let us derive what x_{k+1} is. Observe that

$$0 = \nabla f(x_k) + \alpha(x_{k+1} - x_k).$$

This is equivalent to

$$x_{k+1} = x_k - \frac{1}{\alpha} \nabla f(x_k).$$

This is an instance of the steepest-descent algorithm with a fixed step-size.

6 Applications to Algorithms

We now consider the application of the foregoing discussion convex analysis to algorithms. We start with the proximal point algorithm. Although the proximal point algorithm in its first form is almost never used in minimization algorithms, its convergence proof contains key ideas that are relevant for more complicated algorithms. In fact, the algorithms discussed in the sequel can be written as proximal point algorithms. We then discuss firmly non-expansive operators, which may be regarded as building blocks for developing convergent algorithms. Following this, we discuss the augmented Lagrangian, the Douglas-Rachford algorithm, and the alternating direction method of multipliers algorithm (ADMM). The last section considers a generalization of the proximal point algorithm, along with an application to a saddle point problem. I hope to add more algorithms to this collection...

6.1 The Proximal Point Algorithm

Consider the minimization of a convex function $f(x)$. The proximal point algorithm constructs a sequence x^k that converges to a minimizer. The sequence is defined as

$$x^{k+1} = \arg\min_x \underbrace{\frac{1}{2\alpha}\|x - x^k\|_2^2 + f(x)}_{F_k(x)} \tag{6.20}$$

The function $F_k(x)$ is a perturbation of $f(x)$ around x_k. The quadratic term ensures that x^{k+1} is close to x^k.

In fact, this algorithm may actually be regarded as an MM algorithm since

- $F_k(x_k) = f(x_k)$,
- $F_k(x) \geq f(x)$ for all x.

It follows immediately that

$$f(x^{k+1}) \leq F_k(x^{k+1}) \leq F_k(x^k) = f(x^k)$$

In terms of subdifferentials, we have

$$0 \in (x^{k+1} - x^k) + \alpha\, \partial f(x^{k+1}).$$

Equivalently, we can write

$$x^k \in (I + \alpha \partial f)\, x^{k+1}.$$

Thus, at the k^{th} step of PPA, we are essentially computing the inverse of $(I + \alpha\, \partial f)$ at x^k. Notice that, since $F_k(x)$ in (6.20) is strictly convex, x^{k+1} is uniquely defined. Therefore, $(I + \alpha\, \partial f)^{-1}$ is a single-valued operator.

Definition 25. For a convex f, the operator $\mathcal{J}_f = (I + \partial f)^{-1}$ is called the *proximity operator* of f. ◇

The proximity operator of a convex function acts like a projection operator in the following sense.

Proposition 39. For a proximity operator $\mathcal{J}_{\alpha f}$, we have

$$\langle \mathcal{J}_{\alpha f}(x) - \mathcal{J}_{\alpha f}(y), x - y \rangle \geq \|\mathcal{J}_{\alpha f}(x) - \mathcal{J}_{\alpha f}(y)\|_2^2. \tag{6.21}$$

Proof. Let $x' = \mathcal{J}_{\alpha f}(x)$ and $y' = \mathcal{J}_{\alpha f}(y)$. Then, we have

$$\begin{aligned} x \in (I + \alpha\, \partial f)\, x' \quad &\iff \quad (x - x') \in \alpha \partial f(x'), \\ y \in (I + \alpha\, \partial f)\, y' \quad &\iff \quad (y - y') \in \alpha \partial f(y'). \end{aligned}$$

Thanks to the monotonicity of $\alpha \partial f$, we can therefore write

$$\begin{aligned} & \langle x' - y', (x - x') - (y - y') \rangle \geq 0 \\ \iff \quad & \langle x' - y', x - y \rangle \geq \|x' - y'\|_2^2, \end{aligned}$$

which is what we wanted to show. □

The property in (6.21) is a key observation so we give it a name.

Definition 26. An operator F is said to be *firmly-nonexpansive* if

$$\langle Fx - Fy, x - y \rangle \geq \|Fx - Fy\|_2^2.$$

◇

Thus, proximity operators are firmly-nonexpansive.

Let us now make another observation regarding the proximity operator. Suppose x^* minimizes f. For $\alpha > 0$, this is equivalent to

$$\begin{aligned} & 0 \in \partial f(x^*) \\ \iff \quad & 0 \in \alpha\, \partial f(x^*) \\ \iff \quad & x^* \in x^* + \alpha\, \partial f(x^*) \\ \iff \quad & x^* = \mathcal{J}_{\alpha f}(x^*). \end{aligned}$$

The last equality is a very special one.

Definition 27. A point x is said to be a *fixed point* of an operator F if $x = F(x)$. ◇

We have thus shown the following.

Proposition 40. A point x^* minimizes the convex function f if and only if $x^* = \mathcal{J}_{\alpha f}(x^*)$ for all $\alpha > 0$.

The following theorem is known as the Krasnoselskii-Mann theorem.

Theorem 1 (Krasnoselskii-Mann)**.** Suppose F is a firmly-nonexpansive mapping on $\mathbb{R}^N$ and its set of fixed points is non-empty. Let $x^0 \in \mathbb{R}^N$. If the sequence x^k is defined as $x^{k+1} = F\,x^k$, then x^k converges to a fixed point of F.

Before we prove the theorem, let us state an auxiliary result of interest, that provides an alternative definition of firm-nonexpansivity.

Lemma 2. For an operator F, the following conditions are equivalent.

- $\langle Fx - Fy, x - y\rangle \geq \|Fx - Fy\|_2^2$
- $\|Fx - Fy\|_2^2 + \big\|(I - F)x - (I - F)y\|_2^2 \leq \|x - y\|_2^2$

Proof. Exercise!
Hint : Expand the expression

$$\|Fx - Fy\|_2^2 + \big\|(x - y) - (Fx - Fy)\|_2^2.$$

□

We are now ready for the proof of the Krasnoselskii-Mann Theorem.

Proof of the Krasnoselskii-Mann Theorem. Pick some z such that $z = Fz$. We have,

$$\begin{aligned}\|Fx^k - z\|_2^2 &= \|Fx^k - Fz\|_2^2 \\ &\leq \|x^k - z\|_2^2 - \|(I - F)\,x^k - \underbrace{(I - F)\,z}_{=0}\,\|_2^2\end{aligned}$$

Therefore,

$$\|x^{k+1} - z\|_2^2 \leq \|x^k - z\|_2^2 - \|x^k - x^{k+1}\|_2^2. \tag{6.22}$$

Summing this inequality from $k = 0$ to n, and rearranging, we obtain

$$\|x^{n+1} - z\|_2^2 \leq \|x^0 - z\|_2^2 - \sum_{k=0}^{n} \|x^k - x^{k+1}\|_2^2.$$

From this inequality, we obtain

$$\sum_{k=0}^{n} \|x^k - x^{k+1}\|_2^2 \leq \|x^0 - z\|_2^2.$$

Therefore, $\|x^k - x^{k+1}\|_2 \to 0$, which implies $(I - F)x^k \to 0$.

Also, since $\|x^n - z\|_2 \leq \|x^0 - z\|_2$ for any n, the sequence x^k is bounded. Therefore, we can pick a convergent subsequence $\{x^{k_n}\}_n$ with limit, say x^*. By the continuity of $I - F$, we have,

$$0 = \lim_{n\to\infty} (I - F)x^{k_n} = (I - F)x^*.$$

Therefore $x^* = Fx^*$. If we now plug in $z = x^*$ in (6.22), we have

$$\|x^{k_n+m} - x^*\|_2 \leq \|x^{k_n} - x^*\|_2 \text{ for all } m > 0.$$

Since the $\|x^{k_n} - x^*\|_2$ can be made arbitrarily small by choosing n large enough, it follows that $x^k \to x^*$. □

An immediate corollary of this general result (which we will refer to later) is the convergence of PPA.

Corollary 7. The sequence x^k, constructed by the proximal point algorithm in (6.20) converges to a minimizer of f for any $\alpha > 0$.

Thanks to the Krasnoselskii-Mann theorem, firmly-nonexpansive operators play a central role in the convergence study of a number of algorithms. We now provide a brief study of these operators. We state the results in their most general form, for later reference.

6.2 Firmly-Nonexpansive Operators

We already saw that proximity operators are firmly-nonexpansive. However, not all firmly-nonexpansive operators are proximity operators. Firmly-nonexpansive operators can be generated from monotone, or maximal monotone operators. Specifically, suppose T is a monotone operator with domain $\mathbb{R}^N$ and consider $\mathcal{S}_{\alpha T} = (I + \alpha T)^{-1}$. Let us name this operator, for it has interesting properties that will be useful later.

Definition 28. For a monotone operator T, the operator $\mathcal{S}_T = (I + T)^{-1}$ is called the *resolvent* of T. ◇

We pose two questions regarding resolvent operators.

- Is $\mathcal{S}_{\alpha T}$ defined everywhere?
- Is $\mathcal{S}_{\alpha T}$ single-valued?

We noted earlier that for a convex function f, if $T = \partial f$, then the answer to both questions is affirmative. However, for a general monotone operator, $\mathcal{J}_{\alpha T}$ may not be defined everywhere. Let us consider an example to demonstrate this.

Example 25. Suppose we defined $T : \mathbb{R} \to \mathbb{R}$ as the unit step function, i.e.,

$$T(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1, & \text{if } x > 0. \end{cases}$$

Consider the operator $I + T$. Observe that

$$(I + T)(x) = \begin{cases} x, & \text{if } x \leq 0, \\ x + 1, & \text{if } x > 0. \end{cases}$$

Notice that the interval $(0, 1]$ is not included in the range of $I + T$. Therefore, $\mathcal{S}_T = (I + T)^{-1}$ is not defined for any point in $(0, 1]$. ◇

Nevertheless, in the range of $I + \alpha\, T$, the inverse $(I + \alpha\, T)^{-1}$ is single-valued.

Proposition 41. Suppose T is a monotone mapping and $\alpha > 0$. Then, for any x in the range of the operator $I + \alpha T$, we can find a unique z such that $(I + \alpha\, T)\, z = x$. That is, $(I + \alpha T)^{-1}$ is a single-valued operator on its domain.

Proof. First, observe that for $\alpha > 0$, the operator T is monotone if and only if αT is monotone. Therefore, without loss of generality, we take $\alpha = 1$.

We first show that for $w \in T\, u$ and $z \in T\, y$, if $u + w = y + z$, then $u = y$. To see this, observe that

$$u + w - y - z = 0$$

implies

$$\|(u - y) + (w - z)\|_2^2 = 0,$$

which is equivalent to

$$\|u - y\|_2^2 + \|w - z\|_2^2 + 2\langle u - y, w - z\rangle = 0. \tag{6.23}$$

But the inner product term in (6.23) is non-negative thanks to the monotonicity of T. Therefore, in order for equality to hold in (6.23), we must have $u = y$ and $w = z$.

Now, if x is in the range of $I + \alpha T$, then there is a point v and $u \in \alpha T v$ such that $u + v = x$. But the observation above implies that if this is the case, then u is unique and in fact, $u = (I + \alpha T)^{-1} x$. □

We noted that for a monotone operator T, the range of $I + \alpha T$ may be a strict subset of $\mathbb{R}^n$. This restricts the domain of $\mathcal{S}_{\alpha T}$. For maximal monotone operators, the range of $I + \alpha T$ is in fact the whole $\mathbb{R}^N$. This non-trivial result (which we will not prove) is known as Minty's theorem.

Theorem 2 (Minty's Theorem)**.** Suppose T is a maximal monotone operator defined on $\mathbb{R}^N$ and $\alpha > 0$. Then, the range of $I + \alpha T$ is $\mathbb{R}^N$.

To demonstrate Minty's theorem, let us consider the maximal monotone operator whose graph is a superset of the graph of the operator in Example 25.

Example 26. Suppose the set valued operator $\tilde{T}$ is defined as the set valued operator

$$\tilde{T}x = \begin{cases} \{0\}, & \text{if } x < 0, \\ [0,1], & \text{if } x = 0, \\ \{1\}, & \text{if } x > 0. \end{cases}$$

Consider now the operator $I + T$. Observe that

$$(I + \tilde{T})(x) = \begin{cases} \{x\}, & \text{if } x < 0, \\ [0,1], & \text{if } x = 0, \\ \{x+1\}, & \text{if } x > 0. \end{cases}$$

Notice that for any $y \in \mathbb{R}^n$, we can find x such that $y \in (I + \tilde{T})x$. Therefore, the range of $I + \tilde{T}$ is the whole $\mathbb{R}^n$.

Minty's theorem ensures that the domain of $\mathcal{S}_{\alpha T}$ is the whole space if T is maximal monotone. Therefore, we can state the following corollary.

Corollary 8. If T is a maximal monotone operator on $\mathbb{R}^N$, then $\mathcal{S}_{\alpha T}$ is single valued and defined for all $x \in \mathbb{R}^N$.

We have the following generalization of Prop. 39 in this case.

Proposition 42. Suppose T is maximal monotone and $\alpha > 0$. Then, $\mathcal{S}_{\alpha T}$ is firmly-nonexpansive.

Proof. Exercise!
(Hint : Consider the argument used in the proof of Prop. 39.) □

We remark that a subdifferential is maximal monotone. Therefore, a proximity operator is in fact the resolvent of a maximal monotone operator and Prop. 39 is a special case of Prop. 42.

We introduce a final object called the 'reflected resolvent' that will be of interest in the convergence analysis of algorithms that will follow. Let us first state a result to motivate the definition.

Proposition 43. An operator S is firmly non-expansive if and only if $2S - I$ is non-expansive.

Proof. Suppose S is firmly nonexpansive and let $N = 2S - I$. Observe that

$$\|Nx - Ny\|_2^2 = \|x - y\|_2^2 + \Big\{4\|Sx - Sy\|_2^2 - 4\langle Sx - Sy, x - y\rangle\Big\}.$$

But the term inside the curly brackets above is non-negative thanks to the firm-nonexpansivity of S. Thus we obtain

$$\|Nx - Ny\|_2 \leq \|x - y\|_2.$$

Conversely, suppose $N = 2S - I$ is non-expansive. Then, $S = \frac{1}{2}I + \frac{1}{2}N$. Observe also that $I - S = \frac{1}{2}I - \frac{1}{2}N$. We compute

$$\begin{aligned}\|Sx - Sy\|_2^2 + \|(I - S)x - (I - S)y\|_2^2 &= \frac{1}{2}\|x - y\|_2^2 + \frac{1}{2}\|Nx - Ny\|_2^2 \\ &\leq \|x - y\|_2^2.\end{aligned}$$

Thus, S is firmly nonexpansive by Lemma 2. □

Definition 29. Suppose T is maximal monotone and $\alpha > 0$. The operator $\mathcal{N}_T = 2\,\mathcal{S}_T - I$ is called the *reflected resolvent* of T. ◇

Thus, the reflected resolvent of a maximal monotone operator is non-expansive.

Let us state Prop. 43 from another viewpoint for later reference.

Corollary 9. N is nonexpansive if and only if $\left(\frac{1}{2}I + \frac{1}{2}N\right)$ is firmly-nonexpansive.

Another useful observation is the following.

Corollary 10. If f is a convex function and $\alpha > 0$, then $(2\mathcal{J}_{\alpha f} - I)$ is non-expansive.

The following is a useful property to know concerning reflected resolvents (which, in fact, justifies the term 'reflected').

Proposition 44. Suppose T is maximal monotone. Then, $\mathcal{N}_T(x + y) = x - y$ if and only if $y \in T\,x$.

Proof. The claim follows from the following chain of equivalences.

$$\begin{aligned}& \mathcal{N}_T(x + y) = x - y \\ \iff\ & 2\,\mathcal{S}_T(x + y) - (x + y) = x - y \\ \iff\ & \mathcal{S}_T(x + y) = x \\ \iff\ & x + y \in (I + T)\,x \\ \iff\ & y \in T\,x.\end{aligned}$$

□

6.3 The Dual PPA and the Augmented Lagrangian

Consider now a problem of the form

$$\min_x f(x) \text{ subject to } Ex = d, \tag{6.24}$$

where E is a matrix.

In order to solve this problem, we will apply PPA on the dual problem. For this let us derive a dual problem through the use of Lagrangians. Let

$$L(x, \lambda) = f(x) + \langle \lambda, Ex - d \rangle.$$

Then, (6.24) can be expressed as,

$$\min_x \max_\lambda f(x) + \langle \lambda, Ex - d \rangle.$$

In order to obtain the dual problem, we change the order of min and max. This gives us the dual problem,

$$\max_\lambda g(\lambda),$$

where

$$g(\lambda) = \min_x f(x) + \langle \lambda, Ex - d \rangle.$$

Recall that if $\lambda^* \in \arg\max_\lambda g(\lambda)$, then $x^* \in \arg\min_x L(x, \lambda^*)$ is a solution of the original problem (6.24). To find λ^*, we apply PPA on the dual problem and define a sequence as

$$\lambda^{k+1} = \arg\max_\lambda g(\lambda) - \frac{\alpha}{2}\|\lambda - \lambda^k\|_2^2.$$

To find λ^{k+1}, we need to solve

$$\max_\lambda \min_x f(x) + \langle \lambda, Ex - d \rangle - \frac{\alpha}{2}\|\lambda - \lambda^k\|_2^2. \tag{6.25}$$

Let (x^{k+1}, λ^{k+1}) denote the solution of this saddle point problem. To find this point, suppose we first tackle the maximization. Observe that, for fixed x, the optimality condition for the maximization part implies

$$\begin{aligned} 0 &= (Ex - d) - \alpha(\lambda^* - \lambda^k) \\ \iff \lambda^* &= \lambda^k + \frac{1}{\alpha}(Ex - d) \end{aligned}$$

To find x^{k+1}, plug this expression in the saddle point problem (6.25).

$$\begin{aligned} x^{k+1} &= \arg\min_x f(x) + \langle \lambda^k + \frac{1}{\alpha}(Ex - d), Ex - d \rangle - \frac{\alpha}{2}\|\lambda^k + \frac{1}{\alpha}(Ex - d) - \lambda^k\|_2^2 \\ &= \arg\min_x f(x) + \langle \lambda^k, Ex - d \rangle + \frac{\alpha}{2}\|Ex - d\|_2^2. \end{aligned}$$

To summarize, the dual PPA algorithm is

$$x^{k+1} = \arg\min_x \; L_A(x, \lambda^k)$$
$$\lambda^{k+1} = \lambda^k + \frac{1}{\alpha}(Ex^k - d),$$

where

$$L_A(x, \lambda) = f(x) + \langle \lambda, Ex - d \rangle + \frac{\alpha}{2}\|Ex - d\|_2^2.$$

The function L_A is called the augmented Lagrangian. Notice that L_A is similar to the Lagrangian L but contains the additional term $\langle \lambda, Ex - d \rangle$, justifying the expression 'augmented'.

Let us now consider the convergence of this algorithm.

6.4 The Douglas-Rachford Algorithm

Consider now a problem of the form

$$\min_x \; f(x) + g(x), \tag{6.26}$$

where both f and g are convex. Let $F = \partial f$, and $G = \partial G$. If x is a solution of this problem, we should have

$$0 \in Fx + Gx. \tag{6.27}$$

Let us now derive equivalent expressions of this inclusion, to obtain a fixed point iteration. Suppose we fix $\alpha > 0$. Then, (6.27) and the following statements are equivalent.

$$\begin{aligned}
& \text{There exist } u \in Fx,\ z \in Gx \text{ such that } 0 = u + z \\
\iff\ & \text{There exist } u \in Fx,\ z \in Gx \text{ such that } x + \alpha z = x - \alpha u \\
\iff\ & \text{There exist } u \in Fx \text{ such that } x = (I + \alpha G)^{-1}(x - \alpha u) && (6.28) \\
\iff\ & x \in (I + \alpha G)^{-1}(I - \alpha F)\, x && (6.29)
\end{aligned}$$

At this point, observe that if f is differentiable then F is single-valued. In that case, the inclusion in (6.29) is actually an equality (may I confuse matters : this statement is correct if we consider the range of F as $\mathbb{R}^n$ and not $2^{\mathbb{R}^n}$). This suggests the following fixed point iterations, which is known as the forward-backward splitting algorithm.

$$x^{k+1} = (I + \alpha G)^{-1}(I - \alpha F)\, x$$

We will discuss this algorithm later.

We would like to derive an algorithm that employs $\mathcal{J}_{\alpha f} = (I+\alpha F)^{-1}$. For this, let us now backtrack to (6.28) and write down another equivalent statement.

$$\iff \quad \text{There exist } u \in Fx \text{ such that } x+\alpha u = (I+\alpha G)^{-1}\,(x-\alpha u)+\alpha u \quad (6.30)$$

If we now define a new variable $t = x + \alpha u$, we have the following useful equalities :

$$\begin{aligned} x &= (I + \alpha\, F)^{-1}\, t = \mathcal{J}_{\alpha f}(t) \\ \alpha u &= t - x = (I - \mathcal{J}_{\alpha f})\, t \end{aligned}$$

Plugging these in (6.30), we obtain the following proposition.

Proposition 45. Suppose f and g are convex and the minimization problem (6.26) has at least one solution. A point x is a solution of (6.26) if and only if

$$x = \mathcal{J}_{\alpha f}(t), \text{ for some } t \text{ that satisfies } t = \Big(\mathcal{J}_{\alpha g}\big(2\mathcal{J}_{\alpha f} - I\big) + (I - \mathcal{J}_{\alpha f})\Big)(t). \quad (6.31)$$

Thus, if we can obtain t that satisfies the fixed point equation in (6.31), we can obtain the solution to our minimization problem as $\mathcal{J}_{\alpha f}(t)$. A natural choice is to consider the fixed point iterations

$$t^{k+1} = \Big(\mathcal{J}_{\alpha g}\big(2\mathcal{J}_{\alpha f} - I\big) + (\mathcal{J}_{\alpha f} - I)\Big)(t^k).$$

These constitute the Douglas-Rachford iterations. By a little algebra, we can put them in a form that is easier to interpret. For this observe that,

$$\begin{aligned} \mathcal{J}_{\alpha g}\big(2\mathcal{J}_{\alpha f} - I\big) + (I - \mathcal{J}_{\alpha f}) &= \mathcal{J}_{\alpha g}\big(2\mathcal{J}_{\alpha f} - I\big) - \frac{1}{2}(2\mathcal{J}_{\alpha f} - I) + \frac{1}{2}I \\ &= \frac{1}{2}I + \frac{1}{2}(2\mathcal{J}_{\alpha g} - I)\,(2\mathcal{J}_{\alpha f} - I). \end{aligned} \quad (6.32)$$

The convergence of the algorithm is easier to see in this form. Recall that, Corollary 10 implies the non-expansivity of $(2\mathcal{J}_{\alpha f} - I)$ and $(2\mathcal{J}_{\alpha g} - I)$. But composition of non-expansive operators is also non-expansive. Therefore, the composite operator $(2\mathcal{J}_{\alpha g} - I)\,(2\mathcal{J}_{\alpha f} - I)$ is non-expansive. Finally, by Corollary 9, we can conclude that the operator in (6.32) is firmly-nonexpansive. Combining this observation with Prop. 47, we obtain the following convergence result as a consequence of the Krasnoselskii-Mann theorem (i.e., Thm. 1).

Proposition 46. Suppose f and g are convex and the minimization problem (6.26) has at least one solution. The sequence constructed as

$$t^{k+1} = \left(\frac{1}{2}I + \frac{1}{2}(2\mathcal{J}_{\alpha g} - I)\,(2\mathcal{J}_{\alpha f} - I)\right)(t^k), \quad (6.33)$$

is convergent. If t^* denotes the limit of this sequence, then $x^* = \mathcal{J}_{\alpha f}(t^*)$ is a solution of (6.26).

Generalized Douglas-Rachford Algorithm

In this section, we obtain a small variation on the Douglas-Rachford iterations in (6.31). Let us start with the following general observation. Suppose T is a single valued operator and $z = \big((1-\beta^*)I + \beta^* T\big)z$ for some $\beta^* \neq 0$. Then, actually $z = \big((1-\beta)I + \beta T\big)z$, for any β. Therefore, we have the following generalization of Prop. 47.

Proposition 47. Suppose f and g are convex and the minimization problem (6.26) has at least one solution. A point x is a solution of (6.26) if and only if $x = \mathcal{J}_{\alpha f}(t)$ for some t that satisfies

$$t = \Big((1-\beta)I + \beta(2\mathcal{J}_{\alpha g} - I)\,(2\mathcal{J}_{\alpha f} - I)\Big)\,t, \text{ for all } \beta.$$

Although this proposition is true for any $\beta \in \mathbb{R}$, we will specifically be interested in $\beta \in (0,1)$, because that is the interval for which we can construct a convergent sequence that can be used to obtain a minimizer of (6.26). The generalized DR iterations are as follows.

Proposition 48. Suppose f and g are convex and the minimization problem (6.26) has at least one solution. Also, let $\beta \in (0,1)$. Then, the sequence constructed as

$$t^{k+1} = \Big((1-\beta)I + \beta(2\mathcal{J}_{\alpha g} - I)\,(2\mathcal{J}_{\alpha f} - I)\Big)\,(t^k), \tag{6.34}$$

is convergent. If t^* denotes the limit of this sequence, then $x^* = \mathcal{J}_{\alpha f}(t^*)$ is a solution of (6.26).

In order to prove this result, we need a generalization of the Krasnoselskii-Mann theorem. Let us start with a definition.

Definition 30. An operator T is said to be β-averaged with $\beta \in (0,1)$ if T can be written as

$$T = (1-\beta)I + \beta N,$$

for a non-expansive operator N. ◇

Specifically, a firmly-nonexpansive operator is $\frac{1}{2}$-averaged. Notice also that if T is β-averaged and $\beta < \beta'$, then T is also β'-averaged.

Let us now consider how we can generalize the Krasnoselskii-Mann theorem. Recall that, in the proof of that theorem, a key inequality used in the beginning of the proof was of the form

$$\|Tx - Ty\|_2^2 + \|(I-T)x - (I-T)y\|_2^2 \leq \|x-y\|_2^2, \tag{6.35}$$

where T is firmly nonexpansive. However, this depends heavily on the firm non-expansivity of T. If T is only β averaged, with $\beta > 1/2$, then (6.35) does not hold any more. Let us now see if we can come up with an alternative. So, let $T = (1-\beta)I + \beta N$, for a nonexpansive N. Then, we have,

$$\|Tx - Ty\|_2^2 = (1-\beta)^2\|x-y\|_2^2 + \beta^2\|Nx - Ny\|_2^2 + 2\beta(1-\beta)\langle Nx - Ny, x - y\rangle$$

$$\|(I-T)x - (I-T)y\|_2^2 = \beta^2\|x-y\|_2^2 + \beta^2\|Nx - Ny\|_2^2 - 2\beta^2\langle Nx - Ny, x - y\rangle.$$

In order to cancel the inner product terms, let us consider a weighted sum.

$$\begin{aligned}\|Tx - Ty\|_2^2 + \frac{1-\beta}{\beta}\|(I-T)x - (I-T)y\|_2^2 \\ &= \big[(1-\beta)^2 + \beta(1-\beta)\big]\|x-y\|_2^2 + \big[\beta^2 + \beta(1-\beta)\big]\|Nx - Ny\|_2^2 \\ &\leq \|x-y\|_2^2.\end{aligned}$$

Thus, we have shown the following.

Proposition 49. Suppose T is β-averaged. Then,

$$\|Tx - Ty\|_2^2 + \frac{1-\beta}{\beta}\|(I-T)x - (I-T)y\|_2^2 \leq \|x-y\|_2^2. \tag{6.36}$$

We can now state a generalization of Theorem 1, which we also refer to as the Krasnoselskii-Mann theorem.

Theorem 3 (Krasnoselskii-Mann (General Statement))**.** Suppose F is a β-averaged mapping on $\mathbb{R}^N$ and its set of fixed points is non-empty. Given an initial x^0, suppose we define a sequence as $x^{k+1} = F\,x^0$. Then, x^k converges to a fixed point of F.

Proof. Exactly the same arguments as in the proof of Thm 1 work, except that we now start with the inequality (6.36). □

Proposition 48 now follows as a corollary of Theorem 3 because the operator in (6.34) is β-averaged with $\beta \in (0,1)$.

6.5 Alternating Direction Method of Multipliers

Consider the problem

$$\min_x f(x) + g(M\,x). \tag{6.37}$$

In order to solve this problem, we will apply the Douglas-Rachford algorithm on the dual problem. The resulting algorithm is known as the alternating

direction method of multipliers algorithm (ADMM). We will assume that $M$is full column rank.

We first split variables and write (6.37) as a saddle point problem

$$\min_{x,z} \max_{\lambda} f(x) + g(z) + \langle \lambda, M\, x - z \rangle.$$

The dual problem is then

$$\max_{\lambda} \underbrace{\left[\min_{x}\ f(x) + \langle \lambda, M\, x \rangle\right]}_{-d_1(\lambda)} + \underbrace{\left[\min_{z}\ g(z) + \langle \lambda, z \rangle\right]}_{-d_2(\lambda)}$$

Equivalently, the dual problem can be expressed as,

$$\min_{\lambda} d_1(\lambda) + d_2(\lambda), \tag{6.38}$$

for

$$d_1(\lambda) = \max_{x} \langle -M^T \lambda, x \rangle - f(x) = f^*(-M^T\, \lambda) \tag{6.39a}$$

$$d_2(\lambda) = \max_{z} \langle \lambda, z \rangle - g(z) = g^*(\lambda). \tag{6.39b}$$

For this problem, the Douglas-Rachford algorithm, starting from some y^0, can be written as follows.

$$\begin{aligned}
\bar{y}^k &= \mathcal{N}_{\alpha d_2}(y^k), \\
\hat{y}^k &= \mathcal{N}_{\alpha d_1}(\bar{y}^k) \\
y^{k+1} &= \frac{1}{2}\, y^k + \frac{1}{2}\, \hat{y}^k \\
\lambda^{k+1} &= \mathcal{J}_{\alpha d_2}\, (y^{k+1})
\end{aligned}$$

Recall that the sequence of λ^k's converge to a solution of (6.38). Let us now find expressions for the terms above in terms of f and g.

Suppose $y^k = \lambda^k + \alpha\, z^k$, with $z^k \in \partial d_2(\lambda^k)$. This implies (also recall Prop. 44) that $\lambda^k = \mathcal{J}_{\alpha d_2}(y^k)$, so that

$$\begin{aligned}
\bar{y}^k &= 2\, \mathcal{J}_{\alpha d_2}(y^k) - y^k \\
&= 2\lambda\ - (\lambda^k + \alpha\, z^k) \\
&= \lambda^k - \alpha z^k.
\end{aligned}$$

Now observe that

$$\begin{aligned}
\mathcal{J}_{\alpha\, d_1}(\bar{y}^k) &= \arg\min_{y}\ \frac{1}{2\,\alpha} \|y - \bar{y}^k\|_2^2 + d_1(y) \\
&= \arg\min_{y}\ \left[\max_{x}\ \frac{1}{2\alpha} \|y - \bar{y}^k\|_2^2 - f(x) - \langle M^T\, y, x \rangle\right]
\end{aligned}$$

Changing the order of min and max, we find that $\mathcal{J}_{\alpha d_1}(\bar{y}^k)$ must satisfy

$$\mathcal{J}_{\alpha\, d_1}(\bar{y}^k) = \bar{y}^k + \alpha\, M\, x^{k+1},$$

where

$$\begin{aligned} x^{k+1} &:= \arg\max_x \frac{1}{2\alpha}\|\alpha M x\|_2^2 - f(x) - \langle \bar{y}^k + \alpha M x, Mx\rangle \\ &= \arg\max_x \frac{1}{2\alpha}\|\alpha M x\|_2^2 - f(x) - \langle \lambda^k - \alpha z^k + \alpha M x, Mx\rangle \\ &= \arg\min_x f(x) + \langle \lambda^k, Mx\rangle + \frac{\alpha}{2}\|Mx - z^k\|_2^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{y}^k &= 2\mathcal{J}_{\alpha d_1}(y^k) - \bar{y}^k \\ &= \bar{y}^k + 2\alpha M x^k \\ &= \lambda^k - \alpha z^k + 2\alpha M x^{k+1}. \end{aligned}$$

We also have

$$\begin{aligned} y^{k+1} &= \frac{1}{2}y^k + \frac{1}{2}\hat{y}^k \\ &= \lambda^k + \alpha\, M\, x^{k+1} \end{aligned}$$

Let us finally show that y^{k+1} can be expressed as $y^{k+1} = \lambda^{k+1} + \alpha\, z^{k+1}$ for some $z^{k+1} \in \partial d_2(\lambda^{k+1})$, so that the assumption stated in the beginning of the k^{th} iteration is also valid at the $(k+1)^{\text{st}}$ iteration, when we define z^{k+1} properly. We have,

$$\begin{aligned} \lambda^{k+1} &= \arg\min_y \frac{1}{2\alpha}\|y - y^{k+1}\|_2^2 + d_2(y) \\ &= \arg\min_y \left[\max_z \frac{1}{2\alpha}\|y - y^{k+1}\|_2^2 + \langle y, z\rangle - g(z)\right] \end{aligned}$$

Changing the order of min-max, we find

$$\begin{aligned} \lambda^{k+1} &= y^{k+1} - \alpha z^{k+1} \\ &= \lambda^k + \alpha(Mx^{k+1} - z^{k+1}), \end{aligned}$$

where

$$\begin{aligned} z^{k+1} &= \arg\max_z \frac{1}{2\alpha}\|\alpha z\|_2^2 + \langle y^{k+1} - \alpha z, z\rangle - g(z) \\ &= \arg\max_z \frac{1}{2\alpha}\|\alpha z\|_2^2 + \langle \lambda^k + \alpha M x^{k+1} - \alpha z, z\rangle - g(z) \\ &= \arg\min_z\ g(z) - \langle \lambda^k, z\rangle + \frac{\alpha}{2}\|M\, x^{k+1} - z\|_2^2. \end{aligned}$$

The optimality conditions for the last equation are

$$\begin{aligned} & 0 \in \partial g(z^{k+1}) - \lambda^k + \alpha(z^{k+1} - Mx^{k+1}) \\ \iff \quad & \lambda^k - \alpha(z^{k+1} - Mx^{k+1}) \in \partial g(z^{k+1}) \\ \iff \quad & \lambda^{k+1} \in \partial g(z^{k+1}) \\ \iff \quad & z^{k+1} \in \partial g^*(\lambda^{k+1}) = \partial d_2(\lambda^{k+1}). \end{aligned}$$

ADMM in Terms of the Primal Variables

We can rewrite the iterations solely in terms of x^k, z^k, λ^k. This produces the following algorithm, referred to as the ADMM.

$$\begin{aligned} x^{k+1} &= \arg\min_x \; f(x) + \langle \lambda^k, Mx \rangle + \frac{\alpha}{2}\|Mx - z^k\|_2^2 \\ z^{k+1} &= \arg\min_z \; g(z) - \langle \lambda^k, z \rangle + \frac{\alpha}{2}\|Mx^{k+1} - z\|_2^2 \\ \lambda^{k+1} &= \lambda^k + \alpha(Mx^{k+1} - z^{k+1}). \end{aligned}$$

Convergence of ADMM

Recall that we derived ADMM as an instance of Douglas-Rachford algorithm applied on the dual problem in (6.38). The iterations we started with, and their relation to x^k, z^k, are given below.

$$\bar{y}^k = \mathcal{N}_{\alpha d_2}(y^k) \qquad \left[\bar{y}^k = \lambda^k - \alpha z^k\right] \tag{6.42a}$$

$$\hat{y}^k = \mathcal{N}_{\alpha d_1}(\bar{y}^k) \qquad \left[\hat{y}^k = \lambda^k - \alpha z^k + 2\alpha M x^{k+1}\right] \tag{6.42b}$$

$$y^{k+1} = \frac{1}{2}\, y^k + \frac{1}{2}\, \hat{y}^k \qquad \left[y^{k+1} = \lambda^k + \alpha M x^{k+1}\right] \tag{6.42c}$$

$$\lambda^{k+1} = \mathcal{J}_{\alpha d_2}\,(y^{k+1}) \qquad \left[\lambda^{k+1} = \lambda^k + \alpha(Mx^{k+1} - z^{k+1})\right] \tag{6.42d}$$

The convergence results on the Douglas-Rachford iterations therefore ensure that in (6.40), y^k is convergent. Since the operators $\mathcal{N}_{\alpha d_2}$, $\mathcal{N}_{\alpha d_1}$, $\mathcal{J}_{\alpha d_2}$ are continuous, this in turn implies that $\bar{y}^k$, $\hat{y}^k$ and λ^k are also convergent sequences. Thanks to the relations in (6.42), and the full column rank property of M we then obtain that x^k and z^k are also convergent. Let λ^*, x^*, z^* denote the corresponding limits.

First, notice that (6.42d) implies

$$\lambda^* = \lambda^* + \alpha(Mx^* - z^*).$$

Thus, we have $Mx^* = z^*$.

Now, using $Mx^* = z^*$, from (6.42c), and (6.42a) we obtain,

$$\mathcal{N}_{\alpha d_2}(\lambda^* + \alpha z^*) = \lambda^* - \alpha z^*.$$

But by Prop. 44, this implies that $z^* \in \partial d_2(\lambda^*)$. In view of (6.39b), equivalently $\lambda^* \in \partial g(z^*)$.

By a similar argument, we obtain from (6.42a) that $\bar{y}^* = \lambda^* - \alpha M x^*$, from (6.42b) that $\hat{y} = \lambda^* + \alpha M x^*$, and

$$\mathcal{N}_{\alpha d_1}(\lambda^* - \alpha M x^*) = \lambda^* + \alpha M x^*.$$

This time, Prop. 44, implies $-Mx^* \in \partial d_1(\lambda^*)$. In view of (6.39a), equivalently $-M^T\lambda^* \in \partial f(x^*)$. Using this and the previous observation $\lambda^* \in \partial g(z^*)$, we thus can write

$$\begin{aligned} & 0 \in \partial f(x^*) + M^T \lambda^* \\ \iff \quad & 0 \in \partial f(x^*) + M^T \partial g(z^*) \\ \iff \quad & 0 \in \partial f(x^*) + M^T \big(\partial g\big)(Mx^*). \end{aligned}$$

In words, x^* is a solution of the primal problem (6.37).

6.6 A Generalized Proximal Point Algorithm

Recall that, given a maximal monotone T, the proximal point algorithm consists of

$$x^{k+1} = (I + \alpha T)^{-1} x^k.$$

The sequence x^k converges to some x^* such that $0 \in T(x^*)$. Recall that the convergence of PPA depended on the firm-nonexpansivity of $S = (I + \alpha T)^{-1}$, which is equivalent to

$$\langle S\, x_1 - S\, x_2, x_1 - x_2 \rangle \geq \|S\, x_1 - S\, x_2\|_2^2.$$

To derive a generalized PPA, suppose M is a positive definite matrix, and consider the following train of equivalences

$$\begin{aligned} & 0 \in T(x), \\ \iff \quad & Mx \in Mx + \alpha T(x), \\ \iff \quad & x = (M + \alpha T)^{-1}\, M\, x. \end{aligned}$$

Note that, the last line assumes that $(M + \alpha T)$ has an inverse. This is indeed the case, since M can be written as $M = cI + U$ for some positive definite matrix U and $c > 0$. Thanks to positive definiteness, U is maximal monotone. Consequently, $U + \alpha T$ is also maximal monotone and we can then resort to Minty's theorem and Prop. 42 to conclude that $(M + \alpha T) = c\, I + (U + \alpha T)$ has a well-defined inverse. We also remark at this point that M does not have to be symmetric.

We will study the operator $(M + \alpha T)^{-1} M$ in a modified norm.

Lemma 3. Suppose M is a positive definite matrix. Then, the mapping $\langle x, y\rangle_M = \langle x, My\rangle$ defines an inner product.

Proof. To be added... □

In the following, we denote the induced norm $\sqrt{\langle \cdot, \cdot \rangle_M}$ as $\| \cdot \|_M$.

Proposition 50. Suppose M is positive definite and T is maximal monotone with respect to the inner product $\langle \cdot, \cdot \rangle_I$. Then, the operator $S = (M+\alpha T)^{-1}M$ is firmly-nonexpansive with respect to the inner product $\langle \cdot, \cdot \rangle_M$. That is,

$$\langle Sx - Sy, x - y\rangle_M \geq \|Sx - Sy\|_M^2.$$

Proof. Without loss of generality, take $\alpha = 1$. Since the range of $M + T$ is the whole space, and M is invertible, for any $y_i$we can find x_i and $u_i \in T(x)$ such that $My_i = Mx_i + u_i$. Notice that $x_i = Sy_i$. But then,

$$\begin{aligned}\langle Sy_1 - Sy_2, y_1 - y_2\rangle_M &= \langle x_1 - x_2, x_1 - x_2\rangle_M + \langle x_1 - x_2, M^{-1}(u_1 - u_2)\rangle_M \\ &= \|x_1 - x_2\|_M^2 + \langle x_1 - x_2, u_1 - u_2\rangle_I \\ &\geq \|x_1 - x_2\|_M^2.\end{aligned}$$

□

To be added : generalization of the Krasnoselskii-Mann theorem...

In summary, the generalized PPA consists of the following iterations

$$x^{k+1} = (M + \alpha T)^{-1} M\, x^k.$$

Application to a Saddle Point Problem

Consider a problem of the form

$$\min_x \max_{z \in B} f(x) + \langle Ax, z\rangle,$$

where B is a closed, convex set, and f is a convex function. Rewriting, we obtain an equivalent problem as,

$$\min_x \max_z \big\{ L(x, z) = f(x) + \langle Ax, z\rangle - i_B(z) \big\}.$$

Observe that $L(x, z)$ is a convexoconcave function. For such functions, the following operator replaces the subdifferential

$$\begin{aligned} T(x, z) &= \begin{bmatrix} \partial_x L(x, z) \\ \partial_z\big(-L(x, z)\big) \end{bmatrix} \\ &= \begin{bmatrix} \partial f(x) + A^T z \\ \partial i_B(z) - Ax \end{bmatrix} \end{aligned} \tag{6.43}$$

Proposition 51. The operator T defined in (6.43) is maximal monotone.

Proof. Let us first show monotonicity. Observe that

$$\begin{aligned}
&\left\langle T(x_1, z_2) - T(x_2, z_2), \begin{bmatrix} x_1 \\ z_1 \end{bmatrix} - \begin{bmatrix} x_2 \\ z_2 \end{bmatrix} \right\rangle \\
&= \left\langle \begin{bmatrix} \partial f(x_1) \\ \partial i_B(z_1) \end{bmatrix} - \begin{bmatrix} \partial f(x_2) \\ \partial i_B(z_2) \end{bmatrix}, \begin{bmatrix} x_1 - x_2 \\ z_1 - z_2 \end{bmatrix} \right\rangle + \underbrace{\begin{bmatrix} x_1 - x_2 \\ z_1 - z_2 \end{bmatrix}^T \begin{bmatrix} 0 & A^T \\ -A & 0 \end{bmatrix} \begin{bmatrix} x_1 - x_2 \\ z_1 - z_2 \end{bmatrix}}_{=0} \\
&\geq 0.
\end{aligned}$$

Proof of maximality to be added... □

If we apply PPA for obtaining a zero of $T(x, z)$ defined in (6.43), the resulting iterations are complicated. Consider now a generalized PPA (GPPA) with the choice

$$M = \begin{bmatrix} I & \alpha A^T \\ \alpha A & I \end{bmatrix}.$$

Observe that M is positive definite if $\alpha^2\, \sigma(A^T\, A) < 1$. In order to apply GPPA, we need an expression for $(M + \alpha T)^{-1}$. Notice that

$$\begin{aligned}
(M + \alpha T) \begin{bmatrix} x \\ z \end{bmatrix} &= \left(\begin{bmatrix} I & \alpha A^T \\ \alpha A & I \end{bmatrix} + \alpha \begin{bmatrix} \partial f & A^T \\ -A & \partial i_B \end{bmatrix} \right) \begin{bmatrix} x \\ z \end{bmatrix} \\
&= \begin{bmatrix} I + \alpha \partial f & 2\alpha A^T \\ 0 & I + \alpha \partial i_B \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix}
\end{aligned}$$

Thus,

$$(M + \alpha T) \begin{bmatrix} \hat{x} \\ \hat{z} \end{bmatrix} \in \begin{bmatrix} x \\ z \end{bmatrix}$$

means

$$(I + \alpha \partial f)\hat{x} + 2\alpha A^T\, \hat{z} = x, \tag{6.44a}$$

$$(I + \alpha\, \partial i_B)\hat{z} = z. \tag{6.44b}$$

Solving (6.44b) and plugging this back in (6.44a), we obtain the expressions for $\hat{x}$and $\hat{z}$as,

$$\begin{aligned}
\hat{z} &= P_B(z), \\
\hat{x} &= \mathcal{J}_{\alpha f}(x - 2\alpha A^T\, \hat{z}),
\end{aligned}$$

where $P_B = \mathcal{J}_{\alpha i_B}$ denotes the projection operator onto B and $\mathcal{J}_{\alpha f}$ is proximity operator for f.

Observe also that,

$$M \begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} x + \alpha\, A^T z \\ \alpha A x + z \end{bmatrix}.$$

Therefore, the GPPA for this problem is,

$$\begin{aligned} z^{k+1} &= P_B(z^k + \alpha A x^k) \\ x^{k+1} &= \mathcal{J}_{\alpha f}\big(x^k + \alpha A^T\, z^k - 2\alpha A^T\, z^{k+1}\big) \\ &= \mathcal{J}_{\alpha f}\big(x^k - \alpha A^T\, (2z^{k+1} - z^k)\big) \end{aligned}$$

By the analysis on GPPA, we can state that this algorithm converges if $\alpha^2 > \sigma(A^T\, A)$.

Application to an Analysis Prior Problem

Consider now the problem

$$\min_x \frac{1}{2}\|y - H\, x\|_2^2 + \lambda \|A\, x\|_1$$

By making use of the dual expression of the ℓ_1 norm, we can express this problem as a saddle point problem. For this, recall that,

$$\begin{aligned} \|x\|_1 &= \max_{z \in B_\infty} \langle x, z \rangle, \\ &= \max_z \langle x, z \rangle - i_{B_\infty}(z), \end{aligned}$$

where B_∞ denotes the unit ball of the ℓ_∞ norm. The equivalent saddle problem is,

$$\min_x \max_z \underbrace{\frac{1}{2}\|y - H\, x\|_2^2}_{f(x)} + \lambda\, \langle A\, x, z \rangle - \lambda\, i_{B_\infty}(z),$$

This is exactly the same form considered above. The choice

$$M = \begin{bmatrix} I & -\alpha A^T \\ -\alpha A & I \end{bmatrix}$$

leads to the iterations

$$\begin{aligned} x^{k+1} &= \mathcal{J}_{\alpha f}(x^k - \alpha A^T z^k) \\ z^{k+1} &= P_{B_\infty}\big(z^k + \alpha A(2x^{k+1} - x^k)\big). \end{aligned}$$

These are the iterations proposed by Chambolle and Pock.