Convex Optimization and Applications to Image Registration

Department of Mathematics, California Polytechnic State University

Brian Knight

Advisor: Dana Paquin

Spring 2020

Contents

Preface

The trouble with being interested in mostly everything is that it becomes particularly difficult to focus on any one thing. The following thesis is a broad overview of the theory of convex optimization, with a final chapter that applies some of this theory discussed. This topic was suggested to me by my advisor, Dr. Dana Paquin. as it is concerns itself with *providing solutions* to a huge variety of practical problems in a very elegant way. At the heart of convex optimization is the existence of global optimal solutions to these questions, and in this thesis we explore the mathematical structure of these sort of problems.

I began work on this feeling intimidated and incapable, and there are still many days where I feel the same. However, somewhere in the midst of the 3 quarters that were dedicated to this work, I realized that my education has given me the ability to tread water in the many roaring and tossing seas that make up the vast ocean of mathematics. Even more, I found myself doing something almost like swimming – moving and exploring as best I could, and staying afloat through each squall and tempest that came. I began with the ambitious goal of developing a basic understanding, implementing some existing applications, and then taking on some open problems. Though I did not reach this final step, I stand at the end of the road feeling equipped to ask interesting questions and to begin searching for answers to those questions – a skill that will be invaluable to me in the future. And really, there is no reason that this skill should apply only to mathematics. I am aware that my education has blessed me exponentially to think, to ask interesting questions, to listen to interesting answers, to struggle day in and day out, and to continue to learn and not grow weary. It is with this sentiment that I say I am incredibly grateful for the opportunity and privilege to study this material and to have a truly culminating experience.

To list every person that is owed thanks for my current state in life is a task no man in a hurry should undertake. Life is hard, and I have needed all of the help I have been given. The entire mathematics department here has welcomed me fully and without question. They have provided amazing opportunity for growth and development, and have challenged me daily to become a better mathematician and critical thinker, and humbled me thoroughly in the process. My advisor, Dana Paquin, has provided constant support and encouragement throughout this process and has been a significant mentor to me in all of my education. My fellow graduate students in the department have consistently been there to help me laugh and enjoy each moment spent in the shared offices. Whether it be grabbing the first, second, third, or n^{th} cup of coffee, or just gabbing together, they have become friends that I will have for a lifetime with whom I can share the wonders of mathematics. My time at Cal Poly could not be fully understood without the role my roommates played. They have never failed to make me laugh, and have helped me tremendously in becoming more of the human being that I desire to be. My friend Ian has taught me the importance and power of honesty, and shown me what it looks like to love people genuinely. My parents and both of my brothers have faithfully loved and supported me from tee ball up until today, and I know they will be there for me tomorrow as well.

If there is a will there is a way
by hook, by crook, by night or day.

I'll go and find it; I'll seek it out
with mind of grit and heart devout.

Some moments good, some of them tough;
when the mountain is tall the going is rough.

But I won't think of bumps and bruises anew,
only of the top – can you see it? – what a view.

And when I have it in my hand
I'll set back out to foreign land

with a touch more learned but much to know
to continue to adventure, to live, to grow.

Part I
Convex Optimization Overview

1 Introduction

1.1 Optimization Problems

We all have problems. Some are big, some are small, many are important and, in my humble opinion, all are interesting. Most of these problems, if one is willing to do a little bit of stretching, can be formulated as what we will call an *optimization problem*. In a very general sense, an optimization problem can be thought of in the following way:

$$\boxed{\begin{array}{ll} \text{minimize} & f(X) \\ \text{subject to} & \text{some restrictions on } X, \end{array}} \tag{1}$$

where X is some quantity, and f is a function of that quantity. These problems tend to be very easy to state and yet very difficult to solve. For example, most people are attempting to solve some sort of optimization problem related to happiness or joy or satisfaction in life. We want to maximize this quantity that is subject to the many different constraints faced day to day. We must eat, we must pay bills, we must do a lot of things we may not want to do. In the midst of what must be done, then, what is the best we can do? How can we squeeze the most juice out? Perhaps we could frame this problem as an optimization problem in the following way:

$$\begin{array}{ll} \text{maximize} & \text{Joy} \\ \text{subject to} & \text{Responsibilities.} \end{array} \tag{2}$$

Less existentially and yet still abstractly, X could be the number of hours of sleep a person gets on average and $f(X)$ could be a numerical measurement of the irritability of that person. Some reasonable restrictions on X, then, would be that $X < 24$, if we were to assume the person spent some time awake, and $X > 0$ to assume a person sleeps for some non-zero time, and so to account for our inability to sleep a negative number of hours. In this case minimizing $f(X)$ corresponds to minimizing a person's irritability as a function of average number of hours of sleep, subject to a realistic average number of hours sleeping time. Immediately a lot of questions arise:

- How is f defined?
- How do we measure irritability numerically?
- Can this problem even be solved?
- If so, how do we solve it?

The process of formulating these problems in any given context is a wonderfully creative art form that will not be a focus of this document. So we are not abundantly concerned with the first two questions. Given an optimization problem, however, we want to be able to answer the last two questions, and in this way we may help the creative people solve their existential problems.

To get started let's look at one more example which may be more familiar and is certainly more tangible.

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad f(x) = ax^2 + bx + c \tag{3}$$

Here we are asked to minimize a quadratic function of $x \in \mathbb{R}$. Because there are no constraints we refer to this problem as an *unconstrained* optimization problem. In this case we know how f is defined, and don't have to deal with measuring irritability. We now ask: *Can this problem be solved?* If $a > 0$ then we know that the function will have a minimum value, but if $a \leq 0$ and $b \neq 0$, the possible outputs will be unbounded below. Since there are no constraints on x, we will be able to run off to $-\infty$ unquestioned, and so we will not have a solution to our problem. So we must have $a > 0$ to move forward. Now, *how do we solve it?* This minimum value will occur at the vertex of the parabola, which, using some basic calculus, is given by finding the point at which

$$f'(x) = 0 \implies x = -\frac{b}{2a},\ f(x) = c - \frac{b^2}{4a}.$$

Thus, the point $x = -\dfrac{b}{2a}$ minimizes the function $f(x)$ and the minimum value is $f(-\frac{b}{2a}) = c - \frac{b^2}{4a}$. So we have solved our first optimization problem. Our solution is given in the form of the optimal point, $-\frac{b}{2a}$, and the corresponding optimal value, $c - \dfrac{b^2}{4a}$. So we have the optimal pair

$$\left(-\frac{b}{2a},\ \ c - \frac{b^2}{4a}\right).$$

Still, even if our function is well behaved, how can we be sure an optimal pair exists? And, furthermore, how can we find it? Of course if we answer the second question, the first comes for free, but they are both of interest in their own right.

1.2 Mathematical Optimization

To begin answering these questions, we will first further develop the infrastructure of an optimization problem by putting some mathematical restrictions to our very general problem (1) given in Section (1). Specifically, a *mathematical optimization problem* is a problem of the form

$$\begin{aligned}&\text{minimize } f_0(x)\\ &\text{subject to } f_i(x) \leq b_i \text{ for } i = 1, 2, \ldots, m, \qquad \text{where } b_i \in \mathbb{R} \text{ for } i = 1, 2, \ldots, m.\end{aligned} \tag{4}$$

If the mathematical context is understood, we call this simply an *optimization problem.* Here $x = (x_0, x_1, \ldots, x_n) \in \mathbb{R}^n$ is the *optimization variable* of the problem, and the function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the *objective function.* The functions $f_i : \mathbb{R}^n \to \mathbb{R}$ are the (inequality) *constraint functions* and the b_i are the *constraints.* We say that a point x^* is *optimal*, or a *solution* to the problem if x^* has the smallest objective value among all of the different vectors that satisfy the constraints. Thus if x^* is optimal, then for any $z \in \mathbb{R}^n$ such that $f_i(z) \leq b_i$ for all $i = 1, \ldots, m$, it follows that $f(x^*) \leq f(z)$.

In this general problem we are seeking to minimize the objective f_0, but this can be stated equivalently as

$$\begin{aligned}&\text{maximize } -f_0(x)\\ &\text{subject to } -f_i(x) \geq b_i \text{ for } i = 1, 2, \ldots, m.\end{aligned} \tag{5}$$

For this reason we will predominately work with minimization problems, understanding that for each statement we make, there is an equivalent statement related to the corresponding maximization problem.

We may also work with *equality constraints,* $f_i(x) = b_i$, which can be formed by combining two inequality constraints $f_i(x) \leq b_i$ and $f_i(x) \geq b_i$.

In this new language, we see that the previous quadratic example (3) is a mathematical optimization problem with $f_0(x) = ax^2 + bx + c$, which has an optimal solution $x^* = -\frac{b}{2a}$.

For a simple example with constraint functions, we could consider a similar problem,

$$\begin{aligned}&\underset{x \in \mathbb{R}}{\text{minimize}} \quad f(x) = x^2\\ &\text{subject to} \quad |x| \geq 4\end{aligned} \tag{6}$$

In this case, to match the form of (4), we have the objective function $f_0(x) = x^2$, and $f_1(x) = -|x|$, and $b_1 = -4$. Here we have two solutions $x_1^* = 4$, $x_2^* = -4$ which both have optimal value $f(4) = f(-4) = 16$.

Rather than specific examples, however, we are curious about the quantifiers that must be put in place in order to ensure we can solve the optimization problem, and will be primarily interested in problems where solutions are guaranteed.

1.3 Convex Optimization

The study of convex optimization, broadly, concerns itself with the solvability of mathematical optimization problems, and methods of how to compute solutions when they exist. A convex optimization problem is one of the form

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{subject to } f_i(x) \le b_i \text{ for } i = 1, 2, \ldots, m, \end{aligned} \tag{7}$$

where the functions $f_0, f_1, \ldots, f_m$ satisfy

$$f_i(\alpha x + \beta y) \le \alpha f_i(x) + \beta f_i(y) \tag{8}$$

for all $x, y \in \mathbb{R}^n$, and $\alpha, \beta \in \mathbb{R}$ with

$$\alpha + \beta = 1, \quad \alpha, \beta \ge 0.$$

Convex optimization problems form a proper subset of optimization problems, and it turns out that a problem of the form above will have a solution granted we have points in our domain that satisfy all the constraints (the problem must be feasible), and that we can't run off to $-\infty$ (as we saw for the parabola $ax^2 + bx + c$ where $a < 0$). The inequalities satisfied by the objective function and inequality functions will be the exact definition for the f_i's to be convex functions.

This rest of this document will address why it is the case that we can solve convex optimization problems by discussing and proving many major results in the study of convex optimization, closely following the classic work *Convex Optimization* by Boyd and Vandenberghe [**boyd**]. It will also discuss how this theory can be used to aid in the pursuit of solving non-convex problems, and consider a particular problem, the image registration problem, to exemplify this.

We outline the rest of the paper in the following way. In Chapter 2 we define and explore many properties and characterizations of convex and affine sets. With basic definitions in place we determine the convexity of different spaces and their subspaces, many of which will be of use later on. Chapter 3 defines a convex function and proves several core results about the behavior of convex functions, most notably first and second order conditions for convexity. It also explores operations which preserve the convexity of a function, such as positive scaling and composition with affine functions. Chapter 4 reintroduces the formal definition convex optimization problem and goes through various examples. In Chapter 5 we develop an understanding of the dual problem to a general optimization problem and discuss various remarkable results relating the dual problem to the original (primal) optimization problem. These results provide key insight into the process of actually solving a convex optimization problem. In Chapter 6 we discuss a humble subset of these solving methods, specifically focusing on an interior point method known as the barrier method. Chapter 7 concludes our formal background of convex optimization. In Chapter 8, we formulate the image registration problem as a convex optimization problem based on a paper by Taylor and Bhusnurmath [**Taylor08**] and discuss the results and limitations of the algorithm, as well as the possibilities for future work in the area. Chapter 9 provides some concluding remarks of the thesis.

2 Affine and Convex Sets

Before we can rightly discuss convex optimization, we must first develop an understanding of convexity and why it is a central idea in finding global optimal solutions.

2.1 A non-convex set

Consider the polynomial function given by

$$f(x) = \frac{1}{12}(x+3)(x+2)(x+1)(x-1)(x-2)(x-3),$$

the graph of which can be seen below.

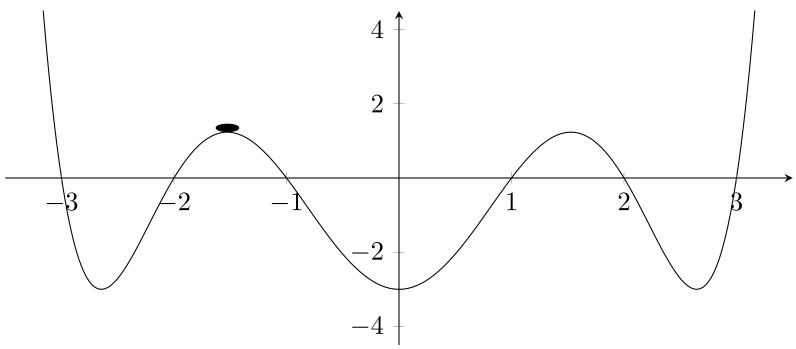


Figure 1: Graph of the function $\frac{1}{12}(x+3)(x+2)(x+1)(x-1)(x-2)(x-3)$, with a 2-dimensional bug on the graph between $x = -1$ and $x = -2$.

Now imagine there is a 2-dimensional bug walking around in the 'hills' and 'valleys' formed by the graph of this polynomial, as seen by the node in Figure (1). The only job that this bug has is to find the lowest possible altitude it can reach while staying on the hills. But the bug cannot see; it can only notice, when it moves, whether its altitude is increasing or decreasing. From its current position, if the bug goes in either direction its altitude will decrease. Noting the change, the bug will continue along until reaching the bottom of the valley where it will decide it has found the lowest point. It has no reason to climb back up a hill, so it has found its final resting ground. The fundamental issue with this strategy is the existence of local minimums, or the fact that there are multiple valleys that must be checked. If the bug were traversing a parabola with a minimum point, it would walk to the correct position every time.

Perhaps unsurprisingly, the existence of a global minimum, and no local minima, can be characterized in the shape of the function, or the shape of the space of points above the function's graph. The illustration in Figure (2) emphasize the fact that two in the shaded region are connected via a line that does not lie in the shaded region, illustrating that the shaded region is not a convex region. This region is called the *epigraph* of the function, and we will discuss these types of sets as they relate to solving minimization problems in the following sections.

2.2 Affine Sets

We begin with a discussion of affine sets.

Definition 1 (*Affine*)**.** A set $A \subseteq \mathbb{R}^n$ is *affine* if the line through any two distinct points in A lies in A.

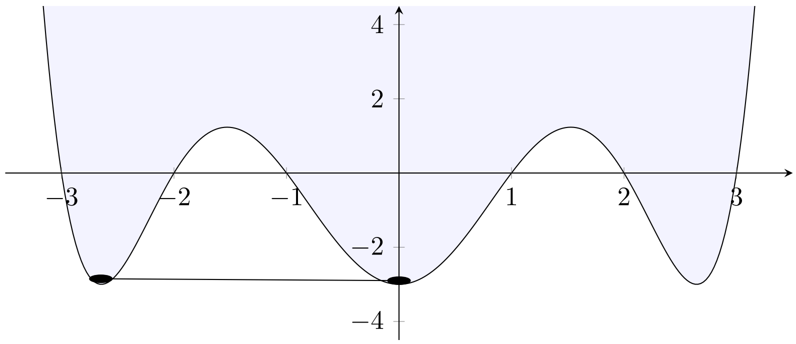


Figure 2: Epigraph of $f(x)$, with to possible locations where the bug would be 'stuck' and have no reason to search for another minima.

In other words, A is affine if and only if for any $x_1, x_2 \in A$, and $\theta \in \mathbb{R}$, we have

$$\theta x_1 + (1-\theta)x_2 \in A.$$

So A contains linear combinations of any two points in A, provided the coefficients sum to 1.

We may also define an *affine combination*, and state our definition of an affine set in an equivalent manner.

Definition 2. (*affine combination*) Let A be a set, and let $x_1, \ldots, x_k \in A$ for $k \in \mathbb{N}$. An *affine combination* of the points $x_1, \ldots, x_k$ is a linear combination

$$\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

such that $\theta_1 + \theta_2 + \cdots + \theta_k = 1$.

A set A is affine if and only if every affine combination of points in A is also in A.

Given any affine set A, and a point $x_0 \in A$, we define its corresponding subspace V by

$$V_{x_0} = A - x_0 = \{x - x_0 \mid x \in A\}.$$

To see that V_{x_0} is indeed a subspace, we need to check that it is closed under sums and scalar multiplication. So let $\alpha, \beta \in \mathbb{R}$, and $v_1, v_2 \in V_{x_0}$. Then

$$\alpha v_1 + \beta v_2 = \alpha(v_1 + x_0) + \beta(v_2 + x_0) + (1 - \alpha - \beta)x_0 \in A.$$

Since $v_1 + x_0, v_2 + x_0, x_0 \in A$, and $\alpha + \beta + (1 - \alpha - \beta) = 1$, it follows that $\alpha v_1 + \beta v_2 \in A$, and thus $\alpha v_1 + \beta v_2 \in V_{x_0}$.

Lastly, we remark that A can be written in terms of an offset of its corresponding subspace,

$$A = V_{x_0} + x_0 = \{v + x_0 \mid v \in V\}.$$

Remark. If A is an affine set, and $x_0 \in A$, the corresponding subspace $V_{x_0} = A - x_0$ is invariant under the choice of x_0. In other words, for any two elements $x_0, y_0 \in A$, we have that $V_{x_0} = V_{y_0}$, and so we may define the corresponding subspace of A independent of a specific point in A.

Proof. Let $v \in V_{x_0}$, so that $v = a - x_0$ for some $a \in A$. We must show $v = a' - y_0$ for some $a' \in A$. Observe that

$$v = a - x_0 = a - x_0 + y_0 - y_0,$$

But $a - x_0 + y_0$ is an affine combination in A, and thus is in A. Letting $a' = a - x_0 + y_0$, we see have

$$v = a' - y_0 \in V_{y_0}.$$

Thus $V_{x_0} \subseteq V_{y_0}$, and by a symmetric argument we conclude $V_{x_0} \supseteq V_{y_0}$, and thus

$$V_{x_0} = V_{y_0}.$$

□

Per the preceding result, we may drop the subscript of from V_{x_0}, and refer the subspace is invariant under the choice of a representative element from A. This allows us to construct a well-defined dimension of any affine set.

Definition 3 (*Dimension* (affine set))**.** Given an affine set A, the *dimension* of A is the dimension of its corresponding subspace V.

Example 1. If $x_0, y_0 \in \mathbb{R}^n$, we may define the line between these two points as

$$A = \{\theta x + (1 - \theta)y \mid \theta \in \mathbb{R}\}.$$

The corresponding subspace of A is the set

$$V = a - x_0,$$

which is a line through he origin in $\mathbb{R}^n$, and so we the dimension of A is 1.

Example 2. A singleton point is also an affine set, and has the corresponding subspace $V = \{0\}$, which has dimension 0.

We can further define the affine dimension of any set $C \subseteq \mathbb{R}^n$ by finding the affine dimension of the smallest affine set containing C.

Definition 4 (*Affine Hull*)**.** The set of all affine combinations of points in some set $C \subseteq \mathbb{R}^n$ is called the *affine hull* of C, and is denoted $\mathbf{aff}\, C$:

$$\mathbf{aff}\, C = \{\theta_1 x_1 + \cdots + \theta_k x_k \mid x_1, \ldots, x_k \in C, \theta_1 + \cdots + \theta_k = 1\}.$$

The affine hull of a set C is the smallest affine set that contains C. i.e., if C is an affine set, then $\mathbf{aff} C = C$, and if S is any affine set such that $C \subseteq S$, then $\mathbf{aff}\, C \subseteq S$.

Example 3. To get a feel for the affine hull, if we let C be the unit circle in $\mathbb{R}^2$ we see that

$$\mathbf{aff}\, \left\{x \in \mathbb{R}^2 \mid x_1^2 + x_2^2 = 1\right\} = \mathbb{R}^2$$

So the affine dimension of the unit circle is 2, which may seem odd considering the unit circle is typically thought of as having dimension 1. Indeed, for any point $x \in \mathbb{R}^2$, a ball of arbitrary radius r around it has affine dimension 2.

In contrast, if the affine dimension of a set C is not the dimension of the whole space, or $\mathbf{aff}\, C \subset \mathbb{R}^n$, then we define the relative interior of the set C as the interior relative to $\mathbf{aff}\, C$.

Definition 5. The *relative interior* of a set $C \subseteq \mathbb{R}^n$, denoted $\mathbf{relint}\, C$, is the set

$$\mathbf{relint}\, C = \{x \in C \mid B(x, r) \cap \mathbf{aff}\, C \subseteq C \text{ for some } r > 0\}$$

where $B(x, r)$ is the ball of radius r for some chosen norm on $\mathbb{R}^n$ (which is typically the Euclidean norm).

Example 4. If $C \subset \mathbb{R}^3$ is the closed unit disk in the xy-plane then $\mathbf{aff}\, C$ is the xy-plane and $\mathbf{relint}\, C = \{(x, y, 0) \mid x^2 + y^2 < 1\}$, or the open unit disk in the xy-plane.

Lastly, we can now define the *relative boundary* of a set C as

$$\mathbf{relbd}\, C = \mathbf{cl}\, C \setminus \mathbf{relint}\, C,$$

where $\mathbf{cl}\, C$ is the union of the set C and all of the limit points of C.

In our previous example C is closed and so we have

$$\begin{aligned} \mathbf{relbd}\, C &= C \setminus \mathbf{relint}\, C \\ &= \left\{(x, y, 0) \mid x^2 + y^2 = 1\right\}. \end{aligned}$$

2.3 Convex Sets

Definition 6 (*Convex Set*)**.** A set C is *convex* if for any two points $x_1, x_2 \in C$, the line segment between them lies in C. i.e.

$$\forall\, \theta \in [0, 1], \quad \theta x_1 + (1 - \theta) x_2 \in C$$

We see immediately that every affine set is convex. Similarly to affine sets, we call a point of the form

$$\theta_1 x_1 + \cdots + \theta_k x_k$$

a convex combination if

$$x_1, x_2, \ldots, x_k \in C, \theta_i \geq 0, \text{ and } \sum_{i=1}^{k} \theta_i = 1.$$

Theorem 1. A set is convex if and only if its intersection with any line is convex. Similarly, a set is affine if and only if its intersection with any line is affine.

Proof. (convex)
$(\implies)$ Let C be a convex set in $\mathbb{R}^n$ for some n, and let L be a line in $\mathbb{R}^n$. If $C \cap L = \{\}$ or $C \cap L = L$, then $C \cap L$ is convex. Otherwise $C \cap L$ is a subset of the line L, so either $C \cap L$ is a ray, a singleton, or a line segment, all of which are convex sets. So in all cases $C \cap L$ is convex.

$(\impliedby)$ Let $C \subseteq \mathbb{R}^n$ for some n, and suppose that $C \cap L$ is convex for any line L in $\mathbb{R}^n$. Assume, seeking a contradiction, that C is not convex. Then there exists some $x_1, x_2 \in C$ so that the line segment between them is not contained in C. i.e.

$$\{\theta x_1 + (1 - \theta) x_2 \mid 0 \leq \theta \leq 1\} \not\subseteq C.$$

Now let L be the line between x_1 and x_2, then

$$C \cap L = C \cap \{\theta_1 x_1 + \theta_2 x_2 \mid \theta_1, \theta_2 \in \mathbb{R}\}$$

We know that this intersection does not contain the line segment between x_1 and x_2, and we know that x_1, x_2 are in the intersection. Therefore $C \cap L$ is not convex, which is a contradiction. So C is convex. $\square$

Proof. (affine)
$(\implies)$ Let C be an affine set in $\mathbb{R}^n$ for some n, and let L be a line in $\mathbb{R}^n$. If $C \cap L = \{\}$, then $C \cap L$ is affine. Suppose $C \cap L \neq \{\}$. If the intersection is a singleton, then it is affine. Otherwise, $C \cap L$ is a subset of the line L, and thus $C \cap L = L$ as C is affine. So in call cases $C \cap L$ is affine.

($\Longleftarrow$) Let $C \subseteq \mathbb{R}^n$ for some n, and suppose that $C \cap L$ is affine for any line L in $\mathbb{R}^n$. Suppose, seeking a contradiction, that C was not affine. Then there exits some points $x_1, x_2 \in C$ such that C does not contain the line between them. Now let L be this line, so the intersection $C \cap L$ is a subset of the line L which does not contain the entire line, thus it is not affine, which is a contradiction. So C is affine. $\square$

Theorem 2. If C is a convex set, then any convex combinations of points in C is also in C.

Proof. Let $C \subseteq \mathbb{R}^n$ be a convex set, with $x_i \in C, \theta_i \geq 0, i = 1, \ldots, k$, and $\sum_{i=1}^k \theta_i = 1$. We must show that

$$\theta_1 x_1 + \cdots + \theta_k x_k \in C.$$

By definition of convexity, we know that for any $x_1, x_2 \in C$, for any $\theta_1, \theta_2 \geq 0$ such that $\theta_1 + \theta_1 = 1$, we have that $\theta_1 x_1 + \theta_2 x_2 \in C$. Define S to be the set of integers $n \geq 2$ such that any convex combination is in C.

$$S = \left\{ n \in \mathbb{Z}_{\geq 2} \mid \left(x_i \in C, \theta_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^n \theta_i = 1 \right) \implies \sum_{i=1}^n \theta_i x_i \in C \right\}.$$

Here $\mathbb{Z}_{\geq 2} = \{n \mid n \in \mathbb{Z} \text{ and } n \geq 2.\}$ So $2 \in S$ by the definition of a convex set. Now assume that $n \in S$ for some $n \geq 2$. We wish to prove $n + 1 \in S$.

Let $x_i \in C, \theta_i \geq 0, i = 1, 2, \ldots, n+1$ such that $\sum_{i=1}^{n+1} \theta_i = 1$. We can assume that $\theta_{n+1} \neq 1$. Then we note that

$$\theta_1 x_1 + \cdots + \theta_n x_n + \theta_{n+1} x_{n+1} = (1 - \theta_{n+1}) \left(\frac{\theta_1 x_1}{1 - \theta_{n+1}} + \cdots + \frac{\theta_n x_n}{1 - \theta_{n+1}} \right) + \theta_{n+1} x_{n+1}.$$

We note that $\sum_{i=1}^n \theta_i = 1 - \theta_{n+1}$, therefore

$$\frac{\theta_1}{1 - \theta_{n+1}} + \cdots + \frac{\theta_n}{1 - \theta_{n+1}} = 1.$$

Let

$$\frac{\theta_1 x_1}{1 - \theta_{n+1}} + \cdots + \frac{\theta_n x_n}{1 - \theta_{n+1}} = x'.$$

Since $n \in S$ by assumption, $x' \in C$, therefore $(1 - \theta_{n+1}) x' + \theta_{n+1} x_{n+1} \in C$ by convexity, and thus $n + 1 \in S$. So, by mathematical induction, $S = \{n \in \mathbb{Z} \mid n \geq 2\}$. $\square$

Definition 7 (*Convex Hull*)**.** Given a set C, the *convex hull* of C, denoted as $\mathbf{conv}\, C$ is the set of all convex combinations of points in C.

$$\mathbf{conv}\, C = \left\{ \sum_{i=1}^k \theta_i x_i, x_i \in C, \theta_i \geq 0, i = 1, 2, \ldots, k, \sum_i^k \theta_i = 1 \right\}.$$

The convex hull of C is always convex, and it is the smallest convex set which contains C.

2.4 More Convex and Affine Sets

2.4.1 Cones

Definition 8 (*Cone*)**.** A set C is called a *cone*, or *non-negative homogeneous*, if for every $x \in C$ and every $\theta \geq 0$, it follows that $\theta x \in C$.

Definition 9 (*Convex Cone*)**.** A set C is a *convex cone* if it is both convex and a cone.

In other words, for any $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have that

$$\theta_1 x_1 + \theta_2 x_2 \in C.$$

Definition 10. (*conic combination*) Given a set C, a point of the form $\theta_1 x_1 + \theta_2 x_2 + \cdots \theta_k x_k$ such that $x_1, \dots, x_k \in C, \theta_i \geq 0, i = 1, 2, \dots k$, is a *conic combination* of $x_1, \dots, x_k$.

Lastly we may consider the *conic hull* of a set C.

Definition 11. Given a set C, the *conic hull* of C is the set

$$\{\theta_1 x_1 + \cdots + \theta_k x_k \mid x_i \in C, \theta_i \geq 0, i = 1, \dots, k\}.$$

The set is all of the conic combinations of points in C, and thus forms a cone. Furthermore, the conic hull of C is the smallest convex cone which contains C.

Example 5 (*Hyperbolic Sets*)**.** Consider the set $S = \{x \in \mathbb{R}^2_+ \mid x_1 x_2 \geq 1\}$, where $\mathbb{R}^2_+ = \{(x, y) \mid x \geq 0 \text{ and } y \geq 0\}$. Let $x, y \in S$. Then $x_1 x_2 \geq 1, y_1 y_2 \geq 1$ for $x = (x_1, x_2)$ and $y = (y_1, y_2)$. If $x \succcurlyeq y$, (if $x_i \geq y_i$) then for any convex combination, $z = \theta x + (1-\theta) y$, it follows that $z_1 z_2 \geq y_1 y_2 \geq 1$.

If $x \not\succcurlyeq y$, then we make use of the fact that for $a, b \geq 0$, and $\theta \in [0, 1]$, it follows

$$a^\theta b^{1-\theta} \leq \theta a + (1-\theta) b.$$

So

$$\begin{aligned} \theta x_1 + (1-\theta) y_1 &\geq x_1^\theta y_1^{1-\theta}, \\ \theta x_2 + (1-\theta) y_2 &\geq x_2^\theta y_2^{1-\theta}, \end{aligned}$$

which implies that $z_1 z_2 \geq (x_1 x_2)^\theta (y_1 y_2)^{1-\theta} \geq 1$. We can extend this result in a similar fashion to show the set

$$S = \{x \in \mathbb{R}^n_+ \mid x_1 x_2 \cdots x_n \geq 1\},$$

is convex for any $n \in \mathbb{N}$.

2.4.2 Hyperplanes and Halfspaces

Definition 12 (*Hyperplane*)**.** A *hyperplane* is a set of the form

$$\{x \mid a^T x = b\},$$

where $a \in \mathbb{R}^n, a \neq 0$, and $b \in \mathbb{R}$.

Geometrically, we can interpret a hyperplane as a translation of the vectors in $\mathbb{R}^n$ which are normal to a, or a hyperplane with a normal vector a. With this interpretation, we can represent the hyperplane equivalently as

$$\{x \mid a^T (x - x_0) = 0\},$$

where x_0 is any point that satisfies $a^T x_0 = b$ (any point in the hyperplane). This means that anything in the set $x_0 + a^\perp$ where $a^\perp = \{v \mid a^T v = 0\}$.

Definition 13. (*Halfspace*) A (closed) *halfspace* is a set of the form

$$\{x \mid a^T x \leq b\} \text{ where } a \neq 0.$$

A hyperplane divides a space into two halfspaces. The boundary of a halfspace is the hyperplane which defines it, $\{x \mid a^T x = b\}$. The interior of the halfspace is the the halfspace minus the boundary, and equals the set $\{x \mid a^T x < b\}$. This is also referred to as an *open halfspace*.

Theorem 3. The distance between two parallel hyperplanes $H_1 = \{x \in \mathbb{R}^n \mid a^T x = b_1\}$ and $H_2 = \{x \in \mathbb{R}^n \mid a^T x = b_2\}$ in $\mathbb{R}^n$ is

$$\frac{|b_2 - b_1|}{\|a\|_2},$$

where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^n$.

Proof. The distance between the two hyperplanes can be found by calculating the distance between any two corresponding points in each plane which are connected by a line normal to the planes. To simplify our work, we will take the line through the origin which is normal to both the planes. This can be parameterized by

$$L(t) = a\frac{t}{\|a\|_2^2},\ t \in \mathbb{R}.$$

This line will intersect the H_1 when $t = b_1$, and it will intersect H_2 when $t = b_2$. So we find the distance between these two points:

$$\left\| a\frac{b_1}{\|a\|_2^2} - a\frac{b_2}{\|a\|_2^2} \right\|_2 = \|a\|_2 \left| \frac{b_1}{\|a\|_2^2} - \frac{b_2}{\|a\|_2^2} \right| = \frac{|b_1 - b_2|}{\|a\|_2}.$$

□

Example 6. *Voronoi description of halfspace.* Let a and b be distinct points in $\mathbb{R}^n$. We can show that the set of all points that are closer (in Euclidean norm) to a than b, i.e.,

$$\{x \mid \|x - a\|_2 \leq \|x - b\|_2\},$$

is a halfspace.

We know that the midpoint of a and b will be equidistant from both points, so we find this point first:

$$m_{a,b} = \frac{a + b}{2}.$$

The line that connects a to b will contain this midpoint, and this vector from a to b will describe the normal to the hyperplane that we want to find. So we have a normal line to our hyperplane, and a point which is in the hyperplane, which is enough to describe the hyperplane explicitly. Namely, if we set $c = b - a$, and $d = c^T m_{a,b}$, we have that

$$\{x \mid \|x - a\|_2 \leq \|x - b\|_2\} = \{x \mid c^T x \leq d\}.$$

To see this explicitly, we notice that

$$\begin{aligned}
&\|x - a\|_2 \leq \|x - b\|_2 \implies (x - a)^T (x - a) \leq (x - b)^T (x - b) \implies \\
&x^T x - 2a^T x + a^T a \leq x^T x - 2b^T x + b^T b \implies \\
&2(b - a)^T x \leq b^T b - a^T a \implies (b - a)^T x \leq \frac{b^T b - a^T a}{2},
\end{aligned}$$

and finally notice that

$$c^T m_{a,b} = (b - a)^T (a + b)/2 = \frac{b^T b - a^T a}{2}.$$

2.4.3 Euclidean Balls and Ellipsoids

Definition 14. (*Euclidean ball*) A *Euclidean ball* in $\mathbb{R}^n$ has the form

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \le r\} = \{x \mid (x - x_c)^T(x - x_c) \le r^2\},$$

for $x_c \in \mathbb{R}^n$ and $r > 0$.

The vector x_c is the *center* of the ball, and r is its *radius*. $B(x_c, r)$ consists of all the points that are at most distance r from x_c.

Proposition. The Euclidean ball is a convex set.

Proof. Let $x_1, x_2 \in B(x_c, r)$. Then

$$\begin{aligned}
\|\theta x_1 + (1-\theta)x_2 - x_c\|_2 &= \|\theta(x_1 - x_c) + (1-\theta)(x_2 - x_c)\|_2 \\
&\le \theta\|(x_1 - x_c)\|_2 + (1-\theta)\|(x_2 - x_c)\|_2 \\
&\le r.
\end{aligned}$$

So $\theta x_1 + (1-\theta)x_2 \in B(x_c, r)$. □

Definition 15. (Ellipsoid) An *ellipsoid* in $\mathbb{R}^n$ is a set

$$\varepsilon(x_c, P) = \{x \mid (x - x_c)^T P^{-1}(x - x_c) \le 1\}, \tag{9}$$

where P is a symmetric positive definite $n \times n$ matrix and $x_c \in \mathbb{R}^n$.

Theorem 4. Ellipsoids are convex.

Proof. Let $x, y \in \varepsilon$, let $x_c \in \mathbb{R}^n$ and P be a positive definite $n \times n$ matrix. Then we have

$$\begin{aligned}
(x - x_c)^T P^{-1}(x - x_c) \le 1 \\
(y - x_c)^T P^{-1}(y - x_c) \le 1.
\end{aligned}$$

Now take $\theta \in [0, 1]$ and consider

$$\begin{aligned}
(\theta x + (1-\theta)y - x_c)^T P^{-1}(\theta x + (1-\theta)y - x_c) &= (\theta(x - x_c) + (1-\theta)(y - x_c))^T P^{-1}(\theta(x - x_c) + (1-\theta)(y - x_c)) \\
&= \theta^2(x - x_c)^T P^{-1}(x - x_c) + (1-\theta)^2(y - y_c)^T P^{-1}(y - y_c) \\
&\le \theta^2 + (1-\theta)^2 \\
&\le 1.
\end{aligned}$$

□

To unpack what this is saying, we'll consider the case in $\mathbb{R}^2$. Then

$$P = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

where $\lambda_1, \lambda_2 > 0$. So $P^T = P$ and $P \succ 0$. Here $P \succ 0$ is used to denote the fact that $x^T P x > 0$ for all $x \in \mathbb{R}^n$, *i.e.* $P \succ 0 \Leftrightarrow P$ is positive definite. So P is symmetric and positive definite. Then

$$P^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{bmatrix},$$

and we see that

$$\varepsilon = \{x \mid (x - x_c)^T P^{-1}(x - x_c) \le 1\} = \left\{x : \frac{(x_1 - x_{c_1})^2}{\lambda_1} + \frac{(x_2 - x_{c_2})^2}{\lambda_2} \le 1\right\}.$$

Also, if $P = P^T, P \succcurlyeq 0$, (where $P \succcurlyeq 0 \Leftrightarrow P$ is positive semidefinite) we may extend to possibly degenerate ellipsoids. In the 2-dimensional case we could flatten to a line segment or a point. Degenerate ellipsoids are also convex.

2.4.4 Norm Balls and Norm Cones

Considering the Euclidean ball and the ellipsoid examples, we remark that both of these sets could be realized as unit balls with respect to some norm. In the case of the Euclidean ball we used the norm induced by the Euclidean metric. In the case of the non-degenerate ellipsoids, for $P \succ 0$ the function

$$\|\cdot\|_{P^{-1}} : \mathbb{R}^n \to \mathbb{R}$$

defined by

$$\|x\|_{P^{-1}} = x^T P^{-1} x$$

indeed defines a norm on $\mathbb{R}^n$.

Given any norm, $\|\cdot\|$, on $\mathbb{R}^n$, the *norm ball* of *radius* r *centered around* x_c , denoted $B(x_c, r)$, is the set

$$B(x_c, r) = \{x \mid \|x - x_c\| \le r\} \,.$$

Theorem 5. For any norm $\|\cdot\|$ on $\mathbb{R}^n$, $x_c \in \mathbb{R}^n$ and $r \in \mathbb{R}$, the set

$$\{x \mid \|x - x_c\| \le r\}$$

is convex.

Proof. The proof is identical to that of the Euclidean ball in $\mathbb{R}^n$, as the proof only made use of norm properties. □

Proof. □

Indeed, we made use of $\mathbb{R}^n$ only in that it was a normed space. Thus the theorem above holds if $\mathbb{R}^n$ is replaced by a general normed space $(X, \|\cdot\|)$. This property of norms is very nice, and we will see in Section (3) that any norm will be what we will call a convex function.

Definition 16. The *norm cone* associated with this norm is the set

$$C = \{(x, t) \mid \|x\| \le t\} \subseteq \mathbb{R}^n.$$

The norm cone is indeed a cone, since for any $(x, t) \in C$ we have that for $\theta \ge 0$

$$\theta(x, t) = (\theta x, \theta t)$$

and

$$\|\theta x\| = \theta\|x\| \le \theta t.$$

2.4.5 Polyhedra

Definition 17. (*Polyhedra*) A *polyhedron* is defined as the solution set of a finite number of equalities and inequalities:

$$\mathcal{P} = \left\{x \mid a_i^T x \le b_i, i = 1, \ldots, m, \quad c_j^T x = d_j, j = 1, \ldots, n\right\}$$

By this definition, we see that a polyhedron is the finite intersection of closed halfspaces and hyperplanes. Affine sets, ray, line segments, and halfspaces are all polyhedra. In more compact notation, we can refer to a polyhedron in the following way:

$$\mathcal{P} = \{x \mid Ax \preccurlyeq b, Cx = d\},$$

where $A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix}$, and C $= \begin{bmatrix} c_1^T \\ \vdots \\ c_n^T \end{bmatrix}$, where $Ax \preccurlyeq b$ means that $a_i^T x_i \le b_i$ for all $i = 1, \ldots, m$.

It is worth noting that polyhedra are closed by definition.

Definition 18. (*Simplices*) A *simplex* is a polyhedra formed by taking the convex hull of $k+1$ affinely independent points in $\mathbb{R}^n$. i.e.

$$C = \mathbf{conv}\{v_0, \ldots, v_k\} = \left\{\theta_0 v_0 + \cdots + \theta_k v_k, \theta \succcurlyeq 0, \sum_{i=1}^{k} \theta_i = 1\right\},$$

where $v_1 - v_0, v_2 - v_1, \ldots, v_k - v_{k-1}$ are linearly independent.

The affine dimension of a simplex formed by $k+1$ affinely independent points has affine dimension k, and thus is commonly referred to as a k-dimensional simplex or k-simplex.

2.4.6 The Positive Semidefinite Cone

Let $\mathbf{S}^n$ denote the set of symmetric $n \times n$ matrices. *i.e.*

$$\mathbf{S}^n = \{X \in \mathbb{R}^{n \times n} \mid X = X^T\}.$$

This is a vector space with dimension $\dfrac{n(n+1)}{2}$. The notation $\mathbf{S}^n_+$ will be used to denote the set of all positive semidefinite matrices:

$$\mathbf{S}^n_+ = \{X \in \mathbf{S}^n \mid X \succcurlyeq 0\},$$

and $\mathbf{S}^n_{++}$ will denote the set of all symmetric positive definite matrices:

$$\mathbf{S}^n_{++} = \{X \in \mathbf{S}^n \mid X \succ 0\}.$$

Theorem 6. The set $\mathbf{S}^n_+$ is a convex cone.

Proof. Let $A, B \in \mathbf{S}^n_+$, and let $\theta_1, \theta_2 \geq 0$. Then

$$\theta_1 A \in \mathbf{S}^n_+, \theta_2 B \in \mathbf{S}^n_+,$$

and $\mathbf{S}^{\mathbf{n}}_+$ is closed under addition. Thus

$$\theta_1 A + \theta_2 B \in \mathbf{S}^n_+.$$

$\square$

Example 7. If we represent a general element of $\mathbf{S}^n_+$ as a matrix, we can then describe the element as a set of inequalities regarding these matrix coefficients. The cases for $n = 1, 2, 3$ are illustrated below:

- $\mathbf{S}^1_+$: A general element in this set is $x_1 \in \mathbb{R}_+$. So we have that $x_1 \geq 0$.
- $\mathbf{S}^2_+$: In this case, a general element $S \in \mathbf{S}^2_+$ takes the form

  $$A = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix}.$$

  In order for $A \succcurlyeq 0$, it must be that $\det A = x_1 x_3 - x_x^2 \geq 0$ since A must have non-negative eigenvalues. Also, we know that $e_1^T A e_1 = x_1 \geq 0$, and $e_2^T A e_2 = x_3 \geq 0$, thus $x_1, x_3 \geq 0$.

- $\mathbf{S}^3_+$: Here, $A \in \mathbf{S}^3_+$ can be written

$$A = \begin{bmatrix} x_1 & x_2 & x_3 \\ x_2 & x_4 & x_5 \\ x_3 & x_5 & x_5 \end{bmatrix}.$$

The vectors e_1, e_2, e_3 give that $x_1, x_4, x_6 \geq 0$ by the same argument as above. Also, each principal minor must be non-negative, and thus we see that

$$\det \begin{bmatrix} x_1 & x_2 \\ x_2 & x_4 \end{bmatrix} = x_1x_4 - x_2^2 \geq 0, \det \begin{bmatrix} x_1 & x_3 \\ x_3 & x_6 \end{bmatrix} = x_1x_6 - x_3^2 \geq 0, \det \begin{bmatrix} x_4 & x_5 \\ x_5 & x_6 \end{bmatrix} = x_4x_6 - x_5^2 \geq 0.$$

Lastly, non-negative eigenvalues are still required, so we get $\det A \geq 0$, thus

$$x_1x_4x_6 + 2x_2x_3x_5 - x_1x_5^2 - x_6x_2^2 - x_4x_3^2 \geq 0.$$

2.5 Operations Preserving Convexity

2.5.1 Intersection

Theorem 7. Given any collection of convex sets, the intersection of the sets is also convex.

Proof. Let $\{C_\lambda\}_{\lambda \in \Lambda}$ be a collection of convex sets. If the intersection is empty, or contains a single point, then it is convex. Suppose the intersection contains more than one point, and let $c_1, c_2 \in \bigcap_{\lambda \in \Lambda} C_\lambda$. Since each C_λ is convex, and $c_1, c_2 \in C_\lambda$ for every $\lambda \in \Lambda$, we know that for any $\theta \in [0,1]$ we have $\theta c_1 + (1-\theta)c_2 \in C_\lambda$ for every $\lambda \in \Lambda$. So

$$\theta c_1 + (1-\theta)c_2 \in \bigcap_{\lambda \in \Lambda} C_\lambda,$$

and thus $\bigcap_{\lambda \in \Lambda} C_\lambda$ is convex. □

This fact will be very helpful in creating new convex sets, or finding ways to prove the convexity of other sets by decomposing them into intersections of sets we know to be convex. The following result should help see this utility.

Corollary 1. Given any norm $\|\cdot\|$ on $\mathbb{R}^n$, any intersection of norm balls

$$\{B(x_\lambda, r_\lambda) \mid \lambda \in \Lambda, x_\lambda \in \mathbb{R}^n, r_\lambda \in \mathbb{R}\},$$

is convex.

Proof. This is immediate by Theorem (7) and Theorem (5). □

We remark again that the only requirement of $\mathbb{R}^n$ used is that it was a normed space, so the above result also holds for a general normed space $(X, \|\cdot\|)$.

As another example of this, we consider the positive semidefinite cone $\mathbf{S}^n_+$, which is the set of the matrices

$$\mathbf{S}^n_+ = \left\{ X \in \mathbf{S}^n \mid x^T X x \geq 0 \text{ for all } x \in \mathbb{R}^n \right\}.$$

Theorem 8. $\mathbf{S}^n_+$ is convex.

Proof. We see that $\mathbf{S}_+^n$ is the intersection of halfspaces in $\mathbf{S}^n$,

$$\bigcap_{x \in \mathbb{R}^n} \left\{ X \in \mathbf{S}^n \mid x^T X x \geq 0 \right\},$$

so by Theorem (7) it is convex. $\square$

Example 8. More Examples of Convex Sets

- A slab of the form $\{x \in \mathbb{R}^n \mid \alpha \leq a^T x \leq \beta\}$ is convex because it can be rewritten as the intersection of two half spaces, as follows:

$$\{x \in \mathbb{R}^n \mid \alpha \leq a^T x \leq \beta\} = \{x \in \mathbb{R}^n \mid a^T x \geq \alpha\} \cap \{x \in \mathbb{R}^n \mid a^T x \leq \beta\}.$$

- A rectangle, or hyperrectangle, of the form

$$\{x \in \mathbb{R}^n \mid \alpha_i \leq x_i \leq \beta_i, i = 1, 2, \ldots, n\},$$

is convex as well. In vector notation, we can rewrite this set as

$$\{x \in \mathbb{R}^n \mid \alpha_i \leq x_i \leq \beta_i, i = 1, 2, \ldots, n\} = \{x \in \mathbb{R}^n \mid \alpha \leq \mathbf{1}^T x \leq \beta\},$$

where $x = [x_1, \ldots, x_n]^T, \alpha = [\alpha_1, \ldots, \alpha_n]^T$, and $\beta = [\beta_1, \ldots, \beta_n]^T$. This set is convex by the previous example.

- A wedge, or a set of the form

$$\{x \in \mathbb{R}^n \mid a_1^T x \leq b_1, a_2^T x \leq b_2\},$$

can also be written as an intersection of two halfspaces and thus is convex.

- The set of points closer to a given point than to a set $S \subseteq \mathbb{R}^n$, is a convex set. In set notation we can write this as

$$\{x \in \mathbb{R}^n \mid \|x - x_0\|_2 \leq \|x - y\|_2 \text{ for all } y \in S\}.$$

This is the intersection

$$\bigcap_{y \in S} \{x \in \mathbb{R}^n \mid \|x - x_0\|_2 \leq \|x - y\|_2\},$$

of which each set in the intersection is a halfspace by the Voronoi Description of a halfspace, and thus is convex.

- The set of points which are closer to one set than another is not necessarily convex. A counterexample is a set containing any two distinct points on the real line, and a point in between them. In particular, consider the set $S = \{-1, 1\}$, and the point at 0. The set of points closer to S than to 0 is not convex because the line between -1 and 1 is not contained in the set, though -1 and 1 are.

- The set of points whose distance to a point a does not exceed a fixed scale θ of the distance to b,

$$\{x \in \mathbb{R}^n \mid \|x - a\|_2 \leq \theta \|x - b\|_2\},$$

is a convex set because it can be rewritten in the form of a norm ball (if $\theta \neq 0$), which is convex. If $\theta = 1$, then this is just a halfspace by the previous example.

2.5.2 Affine Functions

Recall the definition of an affine function.

Definition 19 (*Affine Function*)**.** A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is an *affine function* if it is a sum of linear function and a constant. *i.e.*

$$f(x) = Ax + b,$$

for $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.

Theorem 9. If $S \subseteq \mathbb{R}^n$ is convex, and $f : \mathbb{R}^n \to \mathbb{R}^m$ is an affine function, then the image of S under f,

$$f(S) = \{f(x) \mid x \in S\},$$

is a convex set in $\mathbb{R}^m$.

Similarly, if $f : \mathbb{R}^k \to \mathbb{R}^n$ is an affine function, then the inverse image of S under f,

$$f^{-1}(S) = \{x \mid f(x) \in S\},$$

is a convex set in $\mathbb{R}^k$.

Proof. Suppose $S \subseteq \mathbb{R}^n$ is convex. Consider an affine function $f : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$f(x) = Ax + b,$$

for some $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Let $\xi_1, \xi_2 \in f(S)$, $\theta \in [0, 1]$, and consider the point given by

$$\theta\xi_1 + (1 - \theta)\xi_2.$$

Because $\xi_1, \xi_2 \in f(S)$, we can find $s_1, s_2 \in S$ so that $f(s_1) = \xi_1$, and $f(s_2) = \xi_2$. Then

$$\begin{aligned}
\theta\xi_1 + (1 - \theta)\xi_2 &= \theta f(s_1) + (1 - \theta)f(s_2) \\
&= \theta(As_1 + b) + (1 - \theta)(As_2 + b) \\
&= \theta As_1 + (1 - \theta)As_2 + b \\
&= A(\theta s_1 + (1 - \theta)s_2) + b \\
&= f(\theta s_1 + (1 - \theta)s_2).
\end{aligned}$$

Since S is convex, and $\theta \in [0, 1]$, we see that $\theta s_1 + (1 - \theta)s_2 \in S$, whence $\theta\xi_1 + (1 - \theta)\xi_2 \in f(S)$. So $f(S)$ is convex and is a subset of $\mathbb{R}^m$.

Now consider an affine function $f : \mathbb{R}^k \to \mathbb{R}^n$ given by

$$f(x) = Ax + b,$$

for some $A \in \mathbb{R}^{n\times k}$, $b \in \mathbb{R}^k$. Let $x_1, x_2 \in f^{-1}(S)$, and let s_1, s_2 be two elements in S such that $f(x_1) = (s_1$ and $f(x_2) = s_2$. Then let $\theta \in [0, 1]$, and consider the vector in $\mathbb{R}^k$ given by

$$\theta x_1 + (1 - \theta)x_2.$$

Then we notice that

$$\begin{aligned}
f(\theta x_1 + (1 - \theta)x_2) &= f(\theta x_1) + f((1 - \theta)x_2) \\
&= \theta f(x_1) + (1 - \theta)f(x_2) \\
&= \theta s_1 + (1 - \theta)s_2,
\end{aligned}$$

where the first two equalites hold because f is affine. The final result is in S since S is convex, thus $\theta x_1 + (1 - \theta)x_2 \in f^{-1}(S)$, and thus $f^{-1}(S)$ is convex. $\square$

In particular, we see immediately that scaling, translation, and projection all preserve convexity.

2.5.3 Linear-fractional and perspective functions

To define the linear-fractional function, we first define the perspective function.

Definition 20 (*Perspective function*)**.** The *perspective function* is the function

$$P : \mathbb{R}^{n+1} \to \mathbb{R}^n,$$

with domain $\mathbf{dom}\, P = \mathbb{R}^n \times \mathbb{R}_{++}$ defined as

$$P(z,t) = z/t.$$

So the last component is dropped, and the vector is scaled by the reciprocal of the dropped value. It is helpful to think of this function as looking at the convex set through a pin-hole. The proof that this function indeed takes convex sets to convex sets is derived from the fact that it takes line segments to line segments.

Theorem 10. The image of a convex set and the inverse image of a convex set under the perspective function are both convex.

Proof. Suppose C is a convex set in $\mathbb{R}^{n+1}$. If we let $x = (x', x_{n+1})$, and $y = (y', y_{n+1})$, be two points in C where $x_{n+1}, y_{n+1} > 0$, and $\theta \in [0,1]$, then we see

$$P(\theta x + (1-\theta)y) = \frac{\theta x' + (1-\theta)y'}{\theta x_{n+1} + (1-\theta)y_{n+1}}.$$

Now set

$$\theta' = \frac{\theta x_{n+1}}{\theta x_{n+1} + (1-\theta)y_{n+1}}.$$

Then we see

$$\frac{\theta x' + (1-\theta)y'}{\theta x_{n+1} + (1-\theta)y_{n+1}} = \theta' P(x) + (1-\theta')P(y).$$

Moreover, θ' varies monotonically as θ and draws out the line between $P(x)$ and $P(y)$. Therefore the image of the line segment between any two points $x, y \in C$ is the corresponding line segment between $P(x)$ and $P(y)$. Thus $P(C)$ is convex.

Similarly, let $C \subseteq \mathbb{R}^n$ be convex. Let $(x,t), (y,s) \in P^{-1}(C)$, where

$$P^{-1}(C) = \{(x,t) \in \mathbb{R}^n \times \mathbb{R}_{++} \mid x/t \in C\},$$

and let $\theta \in [0,1]$. So $t, s > 0$. We wish to show that

$$\theta(x,t) + (1-\theta)(y,s) \in P^{-1}(C).$$

First, we note that this is the same as showing

$$\frac{\theta x + (1-\theta)y}{\theta t + (1-\theta)s} \in C.$$

We note that the denominator is in $\mathbb{R}_{++}$. Let

$$\theta' = \frac{\theta t}{\theta t + (1-\theta)s},$$

then it follows $\theta' \in [0,1]$, and

$$\frac{\theta x + (1-\theta)y}{\theta t + (1-\theta)s} = \theta' x + (1-\theta')y \in C \implies \theta(x,t) + (1-\theta)(y,s) \in P^{-1}(C).$$

So the perspective function preserves convexity under its image and preimage. □

A linear-fractional function is a function formed by composing the perspective function with an affine function.

Definition 21 (*Linear-fractional Function*)**.** Suppose $g : \mathbb{R}^n \to \mathbb{R}^{m+1}$ is affine and given by

$$g(x) = \begin{bmatrix} A \\ c^T \end{bmatrix} x + \begin{bmatrix} b \\ d \end{bmatrix},$$

where $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n, d \in \mathbb{R}$.

Then the function $f = P \circ g$, where P is the perspective function on $\mathbb{R}^{n+1}$ given by

$$f(x) = (Ax + b)/(c^T + d), \quad \mathbf{dom} f = \{x \mid c^t x + d > 0\},$$

is a *linear-fractional*, or *projective*, function.

If $c = 0, d > 0$, then the domain of f is $\mathbb{R}^n$ and f is an affine function. So we realize affine functions as a special case of linear-fractional functions. Linear-fractional functions are more general than affine functions, and they still preserve convexity of sets.

Corollary 2. Linear-fractional functions preserve convexity under image and pre-image.

Proof. This follows immediately from convexity being preserved under affine functions and the perspective function. If we have a convex set $C \subseteq \mathbb{R}^n$, then under the affine function the image of C is convex, which implies the image under the perspective image is convex. Similarly, if $C \subseteq \mathbb{R}^m$, then its inverse image, or pre-image, under the perspective function is convex, and thus its preimage under the affine map is also convex. $\square$

2.6 Generalized Inequalities

To disucss the generalized inequality, we first define a proper cone.

Definition 22 (*Proper Cone*)**.** A cone $K \subseteq \mathbb{R}^n$ is called a *proper cone* if it satisfies all of the following:

- K is convex,
- K is closed,
- K is solid, or has non-empty interior, and
- K is pointed, which means it contains no line (equivalently, $x \in K, -x \in K \implies x = 0$).

A proper cone is used to define a generalized inequality, which is a partial ordering on $\mathbb{R}^n$. The partial ordering associated with a proper cone K is given by

$$x \preccurlyeq_K y \Leftrightarrow y - x \in K.$$

We interpret the generalized inequality $x \succcurlyeq_K y$ as the statement $y \preccurlyeq_K x$, and define a strict ordering by the relation

$$x \prec_K y \Leftrightarrow y - x \in \mathbf{int}\, K,$$

and, finally, write $x \succ_K y$ for $y \prec_K x$. Note that when $K = \mathbb{R}_+$, the partial ordering $\preccurlyeq_K$ is the usual ordering $\leq$ on $\mathbb{R}$, and the strict partial ordering $\prec_K$ is the usual $<$ ordering on $\mathbb{R}$.

2.6.1 Properties of generalized inequalities

Let K be a proper cone and $\preccurlyeq_K$ be the associated partial ordering on K. Then the following conditions are satisfied:

- $\preccurlyeq_K$ is preserved under addition: if $x \preccurlyeq_K y$ and $u \preccurlyeq_K v$, then $x + u \preccurlyeq y + v$.
- $\preccurlyeq_K$ is transitive.
- $\preccurlyeq_K$ is preserved under non-negative scaling.
- $\preccurlyeq_K$ is reflexive.
- $\preccurlyeq_K$ is antisymmetric.
- $\preccurlyeq_K$ is preserved under limits: if $x_i \preccurlyeq_K y_i$ for $i = 1, \dots, x_i \to x, y_i \to y$ as $i \to \infty$, then $x \preccurlyeq_K y$.

The strict generalized inequality $\prec_K$ satisfies the following properties:

- if $x \prec_K y$ then $x \preccurlyeq y$.
- if $x \prec y$ and $u \preccurlyeq_K v$ then $x + u \prec_K y + v$.
- if $x \prec y$ and $\alpha > 0$

2.6.2 Minimal and Minimum Elements

The generalized inequality disallows us to be able to compare any two points in our space, and so our idea of minimums and maximums becomes more complicated. We define some notion of this in the following way:

Definition 23 (*Minimum/Maximum Element*)**.** Given a set S, and a generalized inequality $\preccurlyeq_K$ of a some proper cone containing S, we say that $x \in S$ is a *minimum element* of S if $x \preccurlyeq_K y$ for every $y \in S$, and, that x is a *maximum element* of S if $x \succcurlyeq_K y$ for all $y \in S$.

If a set has a minimum (or maximum) element, then it is unique. However, we also define the notion of minimal and maximal when we do not necessarily have unique extrema with respect to the generalized inequality.

Definition 24 (*Minimal/Maximal Element*)**.** We say that $x \in S$ is a *minimal element* of S if for any $y \in S$, $y \preccurlyeq_K x$ if and only if $y = x$. We say that x is a *maximal element* of S if for any $y \in S$, $y \succcurlyeq_K x$ if and only if $y = x$.

Lemma 1. Let S be a set and $\preccurlyeq_K$ be a generalized inequality on S. Then

(a) x is a minimum element of S if and only if $S \subseteq x + K$.

(b) x is a minimal element if and only if $(x - K) \cap S = \{x\}$.

Proof.

(a) ($\implies$) Assume that x is a minimum element of S with respect to the generalized inequality $\preccurlyeq_K$. Thus for any $y \in S$ we have $x \preccurlyeq_K y \implies y - x \in K$, and so we have $y = x + (y - x) \in x + K$. ($\impliedby$). Conversely, if $S \subseteq x + K$ then for all $y \in S$ we have $y = x + k$ for some $k \in K$ which implies $y - x \in K \implies x \preccurlyeq_K y$.

(b) ($\implies$) Let x be minimal in S with respect to $\preccurlyeq_K$ and suppose $y \in (x - K) \cap S$. This implies $y \in S$ and $x \in y + K$ so that $x - y \in K \Leftrightarrow y \preccurlyeq_K x$. Thus $y = x$ as x is minimal. ($\impliedby$) Now suppose $(x - K) \cap S = \{x\}$. Let $y \in S$ so that $y \preccurlyeq_K x$. Then $y \in S$ and $x - y \in K \implies y \in x - K$. Thus $y = x$ and so x is minimal in S with respect to $\preccurlyeq_K$.

$\square$

Example 9. Consider the cone $\mathbb{R}^2_+$ which induces component wise inequality in $\mathbb{R}^2$. Then for some set S, if $x \in S$ is the minimum element of S, it must be that all other points in S are above and to the right of x. If x is a minimal element of S, then it is the case that no point lies to the left of and below x.

2.6.3 Separating and Supporting Hyperplanes

Theorem 11 (**Separating Hyperplane Theorem**)**.** Given two disjoint convex sets $C, D \subseteq \mathbb{R}^n$, there exists some $a \in \mathbb{R}^n, b \in \mathbb{R}$ so that $a^T x \leq b$ for all $x \in C$, and $a^T x \geq b$ for all $x \in D$. The hyperplane given by $\{x \mid a^T x = b\}$ is called a *separating hyperplane* for the sets C and D, or is said to *separate* C and D. *Strict separation* is established when a, b can be found so that the halfspaces described above are open halfspaces.

Proof. Consider the set $K = C - D$, which is convex since $C, -D$ are both convex and the sum of convex sets is convex. Now take the closure of K, $\overline{K}$. and let v be the unique element of $\overline{K}$ with minimum (Euclidean) norm (i.e. $\|v\| < \|k\|$ for any $k \in \overline{K}$, $k \neq v$). Such an element exists because $\overline{K}$ is a closed subset of $\mathbb{R}^n$. Then by the convexity of $\overline{K}$, for any $k \in K$ we know

$$v + t(k - v) \in \overline{K} \text{ for } t \in [0, 1].$$

Thus
$$\|v\|^2 \leq \|v + t(k - v)\|^2 = \|v\|^2 + 2t\langle v, k - v\rangle + t^2\|k - v\|^2.$$
So we get the inequality

$$\begin{aligned}
&0 \leq -2t\|v\|^2 + 2t\langle v, k\rangle + t^2\|k - v\|^2 \\
&0 \leq 2\langle v, k\rangle - 2\|v\|^2 + t\|k - v\|^2 \\
&0 \leq \langle v, k\rangle - \|v\|^2,
\end{aligned}$$

where the second step uses the fact that $t \in [0, 1]$, and the third step is for the limiting case as $t \to 0$. Since k was an arbitrary element of K, this inequality gives

$$\langle v, x - y\rangle \geq \|v\|^2,$$

for all $x \in C, y \in D$. Supposing $v \neq 0$, this implies that $v^T x \geq v^T y + \|v\|^2$ for all $x \in C$, and for all $y \in D$. So we see:
$$\inf_{x \in C} v^T x \geq \sup_{y \in D} v^T y,$$
To cover the case for $v = 0$, we construct a more general argument. Let K be defined as above. First, in the case that K has an empty interior it follows that the dimension of $\textbf{aff}\, K$ is less than the dimension of the whole space, which implies K is contained in some hyperplane, and thus there exists some $c \in \mathbb{R}^n$ so that $v^T x \geq c$ for all $x \in K$.

Now suppose that K has non-empty interior. Then we can construct an increasing sequence $K_1 \subset K_2 \subset K_3 \subset \cdots$ of compact convex subsets of $\textbf{int}\, K$. Since $0 \notin K$ (as C and D are disjoint) we have that each K_n contains a minimum non-zero element v_n, so that by our previous argument $v_n^T x \geq 0$ for all $x \in K_n$. If we define our sequence $(\frac{v_1}{\|v_1\|}, \frac{v_2}{\|v_2\|}, \dots)$ then we have a sequence in the unit ball in $\mathbb{R}^n$, which is compact and thus contains a convergent subsequence (v_{n_k}). This converges to give us an element $v \in \overline{K}$. Now for any $x \in \textbf{int}\, K$, we may find an $N \in \mathbb{N}$ large enough such that $x \in K_N$ and so $v^T x \geq 0$. So we have shown that $v^T x \geq 0$ for all $x \in \textbf{int}\, K$. By the continuity of $\langle \cdot, v\rangle$ this holds for all $x \in K$. Thus by our argument above we again have

$$\inf_{x \in C} v^T x \geq \sup_{y \in D} v^T y,$$

and the proof is complete.

$\square$

Definition 25 (*Supporting Hyperplane*)**.** Consider a subset C of $\mathbb{R}^n$, and a point $x_0 \in \mathbf{bd}\, C$. Suppose $a \in \mathbb{R}^n$ is non-zero and satisfies $a^T x \leq a^T x_0$ for all $x \in C$. Then we say that the hyperplane

$$\{x \mid a^T x = a^T x_0\}$$

is a *supporting hyperplane* to C and the point x_0.

We remark that a supporting hyperplane to C and a given point x_0 is a separating hyperplane for C and $\{x_0\}$. Geometrically, we can think of supporting hyperplanes as tangent planes to C passing through x_0.

Theorem 12 (**Supporting Hyperplane Theorem**)**.** Given any non-empy convex set $C, x_0 \in \mathbf{bd}\, C$, there exists a supporting hyperplane to C at x_0.

Proof. We will consider two cases. First, suppose C has a non-empty interior. Then by the separating hyperplane theorem we know there exists a hyperplane separating the convex sets $\{x_0\}, \mathbf{int}\, C$, which gives some non-zero $a \in \mathbb{R}^n, b \in \mathbb{R}$ so that $a^T x \leq a^T x_0 = b$ for all $x \in C$.

Now supposing that C has empty interior, as in the proof of Separating Hyperplan Theorem, we note that $\mathbf{aff}\, C \subset \mathbb{R}^n$, and thus any hyperplane containing this affine set will contain C and x_0, which gives a trivial supporting hyperplane to C at x_0. $\square$

2.7 Dual Cones and Generalized Inequalities

Definition 26 (*Dual Cone*)**.** Let K be a cone. The set

$$K^* = \{y \mid x^T y \geq 0 \text{ for all } x \in K\}$$

is called the *dual cone* of K.

Theorem 13. K^* is a cone, and is always convex, even if K is not convex.

Proof. Let $y \in K^*$ and $t > 0$. Then $x^T(ty) \geq 0$ for all $x \in K$. So K^* is a cone. To show that K^* is convex, we observe

$$K^* = \bigcup_{x \in K} \{y \mid x^T y \geq 0\}.$$

Thus K^* is a union of hyperplanes and therefore is convex. $\square$

Geometrically, we can think of K^* as the set of y such that the hyperplane with normal $-y$ supports K at the origin. In fact, we can characterize it this way.

Example 10. Find the dual cone of $K = \{Ax \mid x \succcurlyeq 0\}$, where $A \in \mathbb{R}^{m \times n}$.

$$K^* = \{y \mid x^T y \geq 0 \text{ for all } x \in K\} = \{y \mid (Ax)^T y \geq 0 \text{ for all } x \succcurlyeq 0.\}.$$

Rewriting $(Ax)^T y = x^T A^T y = x^T(A^T y) = (A^T y)^T x$, we get that

$$K^* = \{y \mid (A^T y)^T x \geq 0 \text{ for all } x \succcurlyeq 0.\}.$$

Example 11. The *monotone non-negative cone* in $\mathbb{R}^n$ is defined by

$$K_{n+} = \{x \in \mathbb{R}^n \mid x_1 \geq x_2 \geq \cdots \geq x_n\}.$$

We note that K_{n+} is convex since for any $x, y \in K_{n+}, \theta \in [0, 1]$ we get $\theta x, (1-\theta)y \in K_{n+}$, so $\theta x + (1-\theta)y$ will also monotonic and non-negative. For any $x \in K_{n+}$ such that $x \neq 0$, it is clear that

$-x \notin K_{n+}$, thus K_{n+} contains no lines. This cone is also closed, as it is defined by a finite number of inequalities, and is in fact a closed polyhedral cone. Lastly, the element $(n, n-1, n-2, \ldots, 2, 1)$ satisfies the strict inequalities, which is enough to show that the interior is non-empty. Thus K_{n+} is a proper cone, so we can consider its dual.

The dual of K_{n+} is defined by

$$K_{n+}^* = \{y \mid x^T y \geq 0 \text{ for all } x \in K_{n+}\} = \left\{ y \mid \sum_{i=1}^n x_i y_i \geq 0 \right\}.$$

Invoking the identity

$$\sum_{i=1}^n x_i y_i = (x_1 - x_2)y_1 + (x_2 - x_3)(y_1 + y_2) + \cdots + (x_{n-1} - x_n)(y_1 + \cdots + y_{n-1}) + x_n(y_1 + \cdots + y_n),$$

we see that

$$K_{n+}^* = \left\{ y \mid \sum_{i=1}^k y_i \geq 0 \text{ for } k = 1, \ldots, n \right\}.$$

Theorem 14. Let K be a cone. Then $y \in K^*$ if and only if $-y$ is the normal of a hyperplane that supports K at the origin.

Proof. ($\Longrightarrow$) Let $y \in K^*$. So $x^T y \geq 0$ for any $x \in K$. So $-y$ supports K as the origin.
($\Longleftarrow$) Now let y be such that $-y$ is the normal of a hyperplane that supports K at the origin. So for any $x \in K$ we have $x^T(-y) \leq 0 \implies x^T y \geq 0 \implies y \in K^*$. $\square$

Let K be a proper cone and K^* be its dual. Then the dual satisfies several properties:

- K^* is closed and convex.
- $K_1 \subseteq K_2 \implies K_2^* \subseteq K_1^*$.
- If K has nonempty interior, then K^* is pointed.
- If the closer of K is pointed, then K^* has nonempty interior.
- K^{**} is the closure of the convex hull of K. So if K is convex and closed, $K^{**} = K$.

Corollary 3. If K is a proper cone, then so is K^*, and $K^{**} = K$.

Proof. If K is proper then K is convex, closed, non-empty, and pointed. Thus by th the properties of the dual cone it is immediate that K^* is also convex, closed, and pointed with non-empty interior. So K^* is a proper cone. Furthermore, $K^{**} = K$ since K is convex and closed. $\square$

2.7.1 Dual Generalized Inequalities

With the preceding result, we can now consider the generalized inequality induced by K^* whenever K itself is proper. The generalized inequality $\preccurlyeq_{K^*}$ is the *dual* of the generalized inequality $\preccurlyeq_K$. Two important properties relating a generalized inequality and its dual are the following:

- $x \preccurlyeq_K y$ if and only if $\lambda^T x \leq \lambda^T y$ for all $\lambda \succcurlyeq_{K^*} 0$.
- $x \prec_K y$ if and only if $\lambda^T x \leq \lambda^T y$ for all $\lambda \succcurlyeq_{K^*} 0, \lambda \neq 0$.

Also, because $K^{**} = K$, we see that the dual generalized inequality associated with $\preccurlyeq_{K^*}$ is $\preccurlyeq_K$, so the above properties also if we swap the role of K and K^* above.

2.7.2 Characterization of Minimum and Minimal Elements

With the notion of a dual inequality, we are able to characterize minimum and minimal elements of a set $S \subseteq \mathbb{R}$ with respect to the generalized inequality induced by some proper cone K.

Theorem 15. x is the minimum element of S, with respect to the generalized inequality $\preccurlyeq_K$, if and only if, for all $\lambda \succ_{K^*} 0$, x is the unique minimizer of $\lambda^T z$ for all $z \in S$.

Proof. ($\Longrightarrow$) Suppose that x is the minimum element of S with respect to $\preccurlyeq_K$. So we have that $x \preccurlyeq_K z$ for all $z \in S$. Let $z \in S$ such that $z \neq x$. Thus $z - x \preccurlyeq_K 0$. Let $\lambda \succ_{K^*} 0$. So, since $z - x \neq 0$, we have that $\lambda^T(z - x) > 0$. Since $z \in S$ was arbitrary, this shows that x is the unique minimizer of $\lambda^T z$ over $z \in S$.
($\Longleftarrow$) Suppose, conversely, that x is the unique minimizer of $\lambda^T z$ over $z \in S$ for all $\lambda \succ_{K^*} 0$. Then for any $z \in S$, $z \neq x$, and for any $\lambda \succ_{K^*} 0$, we have that $\lambda^T z > \lambda^T x \implies \lambda^T(z - x) > 0$. Therefore $z - x \preccurlyeq_K 0$, and thus $x \preccurlyeq_K z$. So x is the minimum element of the set S. ☐

We remark that S in the previous theorem was not assumed to be convex. In the case of minimal elements of a set S, a similar theorem exists, but is unable to characterize minimal elements without stricter conditions on the set S. We state the following two theorems without proof.

Theorem 16. (1) If $\lambda \succ_{K^*} 0$ and x minimizes $\lambda^T z$ over $z \in S$, then x is minimal in S.

Theorem 17. (2) If x is minimal in a convex set S, then x is a minimizer of $\lambda^T z$ for any $\lambda \succcurlyeq_{K^*} 0$.

3 Convex Functions

Our interest in convex sets in the previous chapter is due to the fact that they behave very nicely with particular classes of functions defined on those sets (and, lets be honest, we are just curious people in general). Perhaps unsurprisingly, convex functions are exactly the class of functions we will work with in convex optimization problems, and thus, in many cases, these are the functions we will be able to optimize. This chapter explores their structure in detail and identifies several sufficient conditions for the existence of a global optimal solution.

3.1 Convex Functions

Definition 27 (*Convex function*)**.** A function $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ is *convex* if $C = \mathbf{dom}\, f$ is a convex set and if for all $x, y \in \mathbf{dom}\, f$ and $\theta \in [0,1]$, the inequality

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) \tag{10}$$

holds.

This definition is equivalent to the geometric idea that the line segment between any two points on the graph of f lies above f. We say that a function is *strictly convex* if the strict inequality above holds for $x \neq y$. We say that f is *concave* if $-f$ is convex, and f is *strictly concave* if $-f$ is strictly convex.

Example 12. Consider the quadratic $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$. First we note that $\mathbb{R}$ is convex. We know the shape of this function and we also know that it has a global minimum at $x = 0$. So let $x, y \in \mathbb{R}, \theta \in (0,1)$, then we set

$$\begin{aligned}
f(\theta x + (1-\theta)y) &\leq \theta f(x) + (1-\theta)f(y) && \Leftrightarrow \\
(\theta x + (1-\theta)y)^2 &\leq \theta x^2 + (1-\theta)y^2 && \Leftrightarrow \\
\theta^2 x^2 + 2\theta xy - 2\theta^2 xy + y^2 - 2\theta y^2 + \theta^2 y^2 &\leq \theta x^2 + y^2 - \theta y^2 && \Leftrightarrow \\
\theta^2(x^2 - 2xy + y^2) - \theta(x^2 - 2xy + y^2) &\leq 0 && \Leftrightarrow \\
\theta(\theta - 1)(x-y)^2 &\leq 0.
\end{aligned}$$

If $\theta = 0$ or $\theta = 1$ then f clearly satisfies (10). Thus $f(x) = x^2$ is a convex function.

Example 13. Affine functions are convex. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an affine function, defined by

$$f(x) = Ax + b.$$

Then equality in (10) holds for any function of this form. So all affine functions are both convex and concave. Conversely, any functions that is both convex and concave is an affine function.

We can characterize convex functions in a very helpful way:

Theorem 18. A function is convex if and only if it is convex when restricted to any line that intersects its domain.

Proof. ($\Longrightarrow$) Suppose f is convex. Then let L be a line intersecting the domain of f. Since $\mathbf{dom}\, f$ is convex, then any two points on the line segment of intersection will satisfy the inequality $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$. So f is convex when restricted to any line intersecting its domain.

($\Longleftarrow$) Conversely, assume f is convex when restricted to any line intersecting its domain. Assume, seeking a contradiction, that f is not convex on its domain. So there must exist some $x, y \in \mathbf{dom}\, f$ such that

$$f(\theta x + (1-\theta)y) > \theta f(x) + (1-\theta)f(y),$$

for some $\theta \in [0,1]$. Now consider the line that passes through x and y. The function f is convex when restricted to this line by assumption, so we get that

$$f(\theta x + (1-\theta)y) > \theta f(x) + (1-\theta)f(y),$$

for any $\theta \in [0,1]$, which is a contradiction. $\square$

3.1.1 Extended Value Extensions

Definition 28. (*Extended-value extension*) If f is a convex function, then $\tilde{f} : C \subseteq \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is the *extended-value extension* defined by

$$\tilde{f} = \begin{cases} f(x) & \text{if } x \in \mathbf{dom}\, f \\ \infty & \text{if } x \notin \mathbf{dom}\, f. \end{cases}$$

Because the extension is defined on all of $\mathbb{R}^n$, it is useful to simplify notation and limit the references made to $\mathbf{dom}\, f$, while still preserving the domain of f, which is given by the set $\{x \mid \tilde{f}(x) < \infty\}$. It is seen easily that the convexity of f implies convexity of $\tilde{f}$.

3.1.2 First Order Conditions

Theorem 19. Let f be differentiable (i.e. ∇f exists for all $x \in \mathbf{dom}\, f$). Then f is convex if and only if $\mathbf{dom}\, f$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) \tag{11}$$

holds for all $x, y \in \mathbf{dom}\, f$.

The affine function of y given by $f(x) + \nabla f(x)^T(y-x)$ is the first-order Taylor approximation of f near x. So the inequality states that the first-order Taylor approximation of a convex function is a *global* underestimator of the function, and that if the function is underestimated globally by the first-order Taylor approximation, then the function is convex.

Similarly, we can characterize strict convexity by making the inequality in (11) strict.

Proof. ($n=1$) We will first prove the theorem in the case where $n=1$.
($\implies$) Assume that $f : \mathbb{R} \to \mathbb{R}$ is convex and differentiable. Let $x, y \in \mathbf{dom}\, f$. Then, since the domain of f is convex (an interval), we see that for any $t \in (0,1]$ we have

$$ty + (1-t)x = x + t(y-x) \in \mathbf{dom}\, f.$$

By convexity we see

$$f(x + t(y-x)) \leq (1-t)f(x) + tf(y) \implies f(y) \geq f(x) + \frac{f(x+t(y-x)) - f(x)}{t}.$$

The limit in the final expression as $t \to \infty$ is the inequality in (11).

Conversely, assuming (11) holds for a differentiable function $f : \mathbb{R} \to \mathbb{R}$, let x, y be any two elements in $\mathbf{dom}\, f$, and let $t \in [0,1]$. Define $z = tx + (1-t)y$. By (11) we see

$$f(x) \geq f(z) + f'(z)(x-z), \text{ and } f(y) \geq f(z) + f'(z)(y-z).$$

Multiplying the first expression by t, and the second by $(1-t)$ yields the equations

$$tf(x) \geq tf(z) + tf'(z)(x-z), \text{ and } (1-t)f(y) \geq (1-t)f(z) + (1-t)f'(z)(y-z).$$

Finally, summing the inequalities we have

$$\begin{aligned}
tf(x) + (1-t)f(y) &\geq (tf'(z)x - tf'(z)z + tf(z)) \\
&\quad + ((1-t)f'(z)y - (1-t)f'(z)z + (1-t)f(z)) \\
&= f(z) + tf'(z) + tf'(z)x + (1-t)f'(z)y - f'(z)z \\
&= f(z).
\end{aligned}$$

So f is convex. $\square$

Now we will prove the conditions for a general convex $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$.

Proof. Let $f : C \subset \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Let $x, y \in \mathbb{R}^n$, and consider f restricted to the line passing through them. This is the function

$$g(t) = f(ty + (1-t)x) \implies g'(t) = \nabla f(ty + (1-t)x)^T (y - x).$$

Assume f is convex. So g is convex, and thus by above we have $g(1) \geq g(0) + g'(0)$, which gives

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

Now assume this inequality holds for any x and y in the domain of f. Since $\mathbf{dom}\, f$ is a convex set we have, for $t \in [0,1]$, $z = ty + (1-t)x \in \mathbf{dom}\, f$. The inequality tells us

$$f(y) \geq f(z) + \nabla f(z)^T (y - z), \text{ and } f(x) \geq f(z) + \nabla f(z)^T (x - z).$$

Following the same pattern we saw for $n = 1$, we now multiply the above inequalities by t and $1 - t$ respectively to yield

$$tf(y) + (1-t)f(x) \geq f(z) = f(ty + (1-t)x),$$

and thus f is convex. $\square$

Example 14. With the above condition in mind, consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = x^T P x + q^T x + b,$$

with $P \in \mathbb{R}^{n \times n}$, $q \in \mathbb{R}^n$. Then we have that f is convex if and only if

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

for all $x, y \in \mathbb{R}^n$. This implies

$$\begin{aligned}
y^T P y + q^T y + b &\geq x^T P x + q^T x + b + y^T P x - q^T x + q^T y - x^T P x \\
&= y^T P x + q^T y + b,
\end{aligned}$$

so we have that f is convex if and only if

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

which implies $P \succcurlyeq 0$.

3.1.3 Second Order Conditions

We can further characterize convex functions by assuming a function f is twice differentiable. i.e. if the Hessian, $\nabla^2 f$, exists at each point in the domain of f (an open set). This is the idea of describing convexity by non-negative curvature everywhere, and has a very nice geometric interpretation, as well as being invaluable in determining the convexity, or lackthereof, of many functions.

Theorem 20. Let $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function, $C = \mathbf{dom}\, f$ be open, and suppose f is twice differentiable. Then f is convex if and only if $\mathbf{dom}\, f$ is convex and $\nabla^2 f \succcurlyeq 0$.

Proof. First, suppose $f : C \to \mathbb{R}$ is a twice differentiable function, and $C = \mathbf{dom}\, f$ is open.
($\Longrightarrow$) Assume $\nabla^2 f \succcurlyeq 0$ and $\mathbf{dom}\, f$ is convex. Then let $x, y \in \mathbf{dom}\, f$ and define a function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(t) = f(tx + (1-t)y).$$

We see then that

$$g'(t) = \nabla f(tx + (1-t)y)(x-y)$$

and

$$g''(t) = (x-y)^T \nabla^2 f(tx + (1-t)y)(x-y).$$

Since $\nabla^2 f$ is positive semidefinite, we see that $g''(t) \geq 0$ for all t such that $tx + (1-t)y \in \mathbf{dom}\, f$. Now we use Taylor's Theorem to say that around t_0 in the domain of g, we can find a t^* so that the following equality holds:

$$g(t) = g(t_0) + g'(t_0)(t-t_0) + g''(t^*)/2(t-t_0)^2.$$

Since $g''(t) \geq 0$ we have the following inequality:

$$g(t) \geq g(t_0) + g'(t_0)(t-t_0).$$

By using this fact for $t = 0$, $t = 1$, and approximating around the point $t_0 = t$, we have

$$\begin{aligned} g(0) &\geq g(t) + g'(t)(0-t) \\ g(1) &\geq g(t) + g'(t)(1-t). \end{aligned}$$

Then, combining these inequalities we see that

$$\begin{aligned} tg(1) + (1-t)g(0) &\geq t(g(t) + g'(t)(1-t)) + (1-t)(g(t) + g'(t)(0-t)) \\ &= g(t) \\ &\Longrightarrow \\ tf(x) + (1-t)f(y) &\geq f(tx + (1-t)y), \end{aligned}$$

and thus f is convex.
($\Longleftarrow$) Now suppose that f is a convex function. So $\mathbf{dom}\, f$ is convex. Thus the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(t) = f(tx + (1-t)y).$$

is also convex, and is defined for (at least) $t \in [0, 1]$. Let $x, y \in \mathbf{dom}\, f$. By the convexity of g and the first order condition of convexity, we have the following inequalities for $t > t'$:

$$g(t) \geq g(t') + g'(t')(t-t') \tag{12}$$
$$g(t') \geq g(t) + g'(t)(t'-t). \tag{13}$$

Combining these inequalites gives the result that

$$g'(t')(t-t') \leq g(t) - g(t') \leq g'(t)(t-t').$$

This implies

$$\frac{g'(t) - g'(t')}{t - t'} \geq 0 \text{ for } t \neq t'.$$

So we as $t' \to t$ we get

$$g''(t) \geq 0.$$

But this shows that for any $x \in \mathbf{dom}\, f, v \in \mathbb{R}^n$, and α such that $\alpha x + (1-\alpha)v \in \mathbf{dom}\, f$, we have

$$g''(\alpha) = v^T \nabla^2 f(x + \alpha v)v \geq 0,$$

thus $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in \mathbf{dom}\, f$. □

Example 15. Classifying Some Convex Functions via Second Order Conditions

- Consider again the quadratic function $f(x) = x^T P x + q^T x$. Then we have

  $$\nabla f(x) = Px + q, \text{ and } \nabla^2 f(x) = P,$$

  thus $f(x)$ is quadratic whenever $P \succcurlyeq 0$.

- Let $f(x) = \|Ax - b\|$, Then we have

  $$\nabla f(x) = 2A^T(Ax - b), \text{ and } \nabla^2 f(x) = 2A^T A,$$

  and $A^T A \succcurlyeq 0$ for all A, thus the least squares objective function is always convex.

3.1.4 Sublevel Sets

Given any function $f : \mathbb{R}^n \to \mathbb{R}$, we can define sublevel and superlevel sets corresponding to some fixed value $\alpha \in \mathbb{R}$. When working with convex functions, these sets will have distinct properties that will often be taken advantage of. Let $f : \mathbb{R}^n \to \mathbb{R}$. Then we have the following definitions:

Definition 29. (α-*sublevel set*) The α-*sublevel set* is defined as

$$C_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \leq \alpha\}.$$

Definition 30. (α-*superlevel set*) The α-*superlevel set* is defined as

$$S_\alpha = \{x \in \mathbf{dom}\, f \mid f(x) \geq \alpha\}.$$

If f is a convex function, then each C_α is also a convex set, but converse of this statement is false. Indeed, consider the function $f(x) = e^{-x}$ and its sublevel sets. Similarly, if f is concave then each S_α is convex, and the converse is false. These properties prove useful to establish the convexity of certain sets when studying convex or concave functions.

3.1.5 Epigraph

Definition 31 (*Epigraph, Hypograph*)**.** The graph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\{(x, f(x)) \mid x \in \mathbf{dom}\, f\},$$

which is a subset of $\mathbb{R}^{n+1}$. The *epigraph* of f is defined as

$$\mathbf{epi}\, f = \{(x, t) \mid x \in \mathbf{dom}\, f, f(x) \leq t\},$$

which is also a subset of $\mathbb{R}^{n+1}$.

Similarly, we can define the *hypograph* of a function to be the set

$$\mathbf{hypo}\, f = \{(x, t) \mid t \leq f(x)\}.$$

Proposition. Given a convex function $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$,

$$\mathbf{epi}\, f = \mathbf{hypo}\ - f.$$

Proof. □

Here we find a link between convex sets and convex functions in the following theorem:

Theorem 21.

1. A function is convex if and only if its epigraph is a convex set.
2. A function is concave if and only if its hypograph is a convex set.

Proof. (1) ($\Longrightarrow$) First, assume that f is convex. Then let $(x_1, t_1), (x_2, t_2) \in \mathbf{epi}\, f$. So $f(x_1) \leq t_1$ and $f(x_2) \leq t_2$. Now let $\theta \in [0, 1]$. Since f is convex we have

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x) \leq \theta t_1 + (1-\theta)t_2 = t.$$

So $(\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2) \in \mathbf{epi}\, f$.

($\Longleftarrow$) Conversely, assume $\mathbf{epi}\, f$ is a convex set. Then let $x_1, x_2 \in \mathbf{dom}\, f$. So $(x_1, f(x_1)), (x_2, f(x_2)) \in \mathbf{epi}\, f$. Since $\mathbf{epi}\, f$ is convex, we see immediately that, for $\theta \in [0, 1]$,

$$\begin{aligned}\theta(x_1, f(x_1)) + (1-\theta)(x_2, f(x_2)) = (\theta x_1 + (1-\theta)x_2, \theta f(x_1) + (1-\theta) f(x_2)) \in \mathbf{epi}\, f \\ \Longrightarrow f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2).\end{aligned}$$

Thus f is convex, as desired.

The second statement follows in a similar fashion.

(2) ($\Longrightarrow$) Suppose f is concave, so $-f$ is convex. So for any $(x_1, t_1), (x_2, t_2) \in \mathbf{hypo}\, f$, we have

$$f(x_1) \geq t_1 \text{ and } f(x_2) \geq t_2.$$

Thus $-f(x_1) \leq -t_1$ and $-f(x_2) \leq -t_2$. So from the convexity of $-f$ we have, for $\theta \in [0, 1]$

$$-f(\theta x_1 + (1-\theta)x_2) \leq -f(\theta x_1) - (1-\theta) f(x_2),$$

so that

$$f(\theta x_1 + (1-\theta)x_2) \geq \theta f(x_1) + (1-\theta) f(x_2) \geq \theta t_1 + (1-\theta)t_2.$$

From this we conclude

$$(\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2) \in \mathbf{hypo}\, f,$$

and therefore $\mathbf{hypo}\, f$ is convex.

($\Longleftarrow$) Finally, suppose that $\mathbf{hypo}\, f$ is convex. Then $(x_1, f(x_1))$ and $(x_2, f(x_2))$ are in $\mathbf{hypo}\, f$ and for any $\theta \in [0, 1]$, we have

$$(\theta x_1 + (1-\theta)x_2, \theta f(x_1) + (1-\theta) f(x_2)) \in \mathbf{hypo}\, f.$$

Therefore

$$f(\theta x_1 + (1-\theta)x_2) \geq \theta f(x_1) + (1-\theta) f(x_2),$$

and thus f is concave. □

3.2 Operations Preserving Convexity

Classifying specific operations which preserve convexity (or concavity) of functions provides us with new ways of constructing convex (or concave) functions and broadens the type of functions we can discuss.

3.2.1 Non-negative Weighted Sums

If f is a convex function and $\alpha > 0$, then the function αf is also convex. To see a use of epigraphs from the previous section, we state and prove the following theorem.

Theorem 22. A non-negative scaling of a convex function is convex.

Proof. Suppose f is convex and $w \geq 0$. Then

$$\textbf{epi}\, wf = \begin{bmatrix} I & 0 \\ 0 & w \end{bmatrix} \textbf{epi}\, f.$$

Thus $\textbf{epi}\, wf$ is convex since $\textbf{epi}\, f$ is convex and the matrix is a linear map. $\square$

Extending this idea, along with the fact that the sum of convex functions is convex, we see, for convex functions $f_1, f_2, \ldots, f_n$, and non-negative weights $w_1, w_2, \ldots, w_n$, the function f defined by

$$f = \sum_{i=1}^{n} w_i f_i$$

is convex. Similarly, a non-negative weighted sum of concave functions is concave. Moreover, a non-negative, nonzero sum of weighted strictly convex, or strictly concave, functions is strictly convex or strictly concave respectively.

Corollary 4. The set of convex functions $f : \mathbb{R}^n \to \mathbb{R}$ is a convex cone.

Proof. Let $\mathcal{F} = \{f : \mathbb{R}^n \to \mathbb{R} \mid f \text{ is convex }\}$. Let $f_1, f_2 \in \mathcal{F}$, and let $w_1, w_2 \geq 0$. Then by the previous theorem we see $w_1 f_1 + w_2 f_2 \in \mathcal{F}$. $\square$

This can also be extended to infinite sums and integrals. As an example, if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, and $w(y) \geq 0$ for each $y \in \mathcal{A}$, then the function g defined by

$$g(x) = \int_{\mathcal{A}} w(y) f(x, y)\, dy$$

is convex in x as long as the integral exists.

3.2.2 Composition with Affine Mappings

Theorem 23. Given a function $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^n$, and $g : \mathbb{R}^m \to \mathbb{R}$ defined by

$$g(x) = f(Ax + b), \quad \textbf{dom}\, g = \{x \mid Ax + b \in \textbf{dom}\, f\},$$

g is convex if f is convex, and g is concave if f is concave.

Proof. If f is twice differentiable, this follows immediately as we have

$$\nabla^2 g = A \nabla^2 f A^T \succcurlyeq 0 \text{ as } \nabla^2 f \succcurlyeq 0.$$

Without assuming anything futher about f, let $x, y \in \textbf{dom}\, g$ and $\theta \in [0, 1]$, then we have

$$\begin{aligned} g(\theta x + (1 - \theta y) &= f(A(\theta x + (1 - \theta)y) + b) \\ &= f(\theta x + (1 - \theta)Ay + b) \\ &= f(\theta x + \theta b + (1 - \theta)Ay + b - \theta b) \\ &= f(\theta(Ax + b) + (1 - \theta)(Ay + b)) \\ &\leq \theta f(Ax + b) + (1 - \theta) f(Ay + b) \\ &= \theta g(x) + (1 - \theta) g(y). \end{aligned}$$

$\square$

3.2.3 Pointwise Maximum, Supremum

Theorem 24. Let $f_1, f_2, \ldots, f_m$ be convex functions. Then

$$f(x) = \max\{f_1(x), f_2(x), \ldots, f_m(x)\}$$

is convex (with $\mathbf{dom}\, f = \bigcap_i^m \mathbf{dom}\, f_i$).

Proof. Note that $\mathbf{dom}\, f$ is an intersection of convex sets, so it is convex. We proceed by induction. Let $m = 2$, and let $x, y \in \mathbf{dom}\, f, \theta \in [0, 1]$. Then it follows

$$\begin{aligned} f(\theta x + (1-\theta)y) &= \max\{f_1(\theta x + (1-\theta)y), f_2(\theta x + (1-\theta)y)\} \\ &\le \max\{\theta f_1(x) + (1-\theta)f_1(x), \theta f_2(x) + (1-\theta)f_2(x)\} \\ &\le \theta \max\{f_1(x), f_2(x)\} + (1-\theta)\max\{f_1(y), f_2(y)\} \\ &= \theta f(x) + (1-\theta)f(y). \end{aligned}$$

Now assume the theorem holds for $m \ge 2$. Then consider

$$f(x) = \max\{f_1(x), f_2(x), \ldots, f_m(x), f_{m+1}(x).\}$$

We observe that

$$f(x) = \max\{\max\{f_1(x), f_2(x), \ldots, f_m(x)\}, f_{m+1}(x)\},$$

and thus since $\max\{f_1(x), f_2(x), \ldots, f_m(x)\}$ is convex by hypothesis, we conclude f is convex. □

Alternatively, we can use another epigraph argument.

Proof. *(Alternate)* We know that for each f_i, $\mathbf{epi}\, f_i$ is a convex set, and that

$$\mathbf{epi}\, f = \{(x, t) \mid f_i(x) \ge t, \text{ for } i = 1, \ldots, n\}.$$

Thus $\mathbf{epi}\, f = \bigcap_i^n \mathbf{epi}\, f_i$ is convex. □

Theorem 25. Let $f(x, y)$ be a function convex in x for all $y \in \mathcal{A}$. Then the function g defined by

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y), \text{ with } \mathbf{dom}\, g = \{x \mid (x, y) \in \mathbf{dom}\, f \text{ for all } y \in \mathcal{A}, \sup_{y \in \mathcal{A}} f(x, y) < \infty\},$$

is convex in x.

Proof. The epigraph of g will correspond to the intersection of all epigraphs of $f(x, y)$ for $y \in \mathcal{A}$. Thus we have

$$\mathbf{epi}\, g = \bigcap_{y \in \mathcal{A}} \mathbf{epi}\, f(\cdot, y).$$

Using the fact that the intersection of an arbitrary family of convex sets is convex, we see g is convex. □

Example 16. Let $C \subseteq \mathbb{R}^n$, and let $\|\cdot\|$ be any norm on $\mathbb{R}^n$, then the function

$$f(x) = \sup_{y \in C} \|x - y\|$$

is convex.

3.2.4 Composition

Given two functions $h : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^k$, we would like to be able to say when the composition $f = h \circ g$ is convex or concave.

3.2.5 Scalar Composition

We'll first consider the case of scalar composition (when $k = 1$). Assuming h and g are both twice differentiable, and also that $n = 1$ since convexity is equivalent to convexity restricted to all lines, we'll suppose $\mathbf{dom}\, h = \mathbb{R}$. So the convexity of f implies $f''(x) \geq 0$, which gives the inequality

$$f''(x) = h''(g(x))(g(x))^2 + h'(g(x))g''(x) \geq 0. \tag{14}$$

We see that if h is convex and non-decreasing, then $h''(x) \geq 0$, and $h'(g(x)) \geq 0$. Then if g is convex, $g''(x) \geq 0$ and so the entire inequality will hold. All similar restrictions for h and g can be derived from similar conditions for convexity/concavity and non-decreasing/non-increasing. Before presenting these, we'll first state that we obtain similar results without assuming $\mathbf{dom}\, g = \mathbb{R}^n, \mathbf{dom}\, h = \mathbb{R}$, and dropping the requirement of twice differentiability of g and h. In this more general setting we have the following result:

Theorem 26. Let $A \subseteq \mathbb{R}$, $B \subseteq \mathbb{R}^n$, and $h : A \to \mathbb{R}$, and $g : B \to \mathbb{R}$, and define $f : \mathbb{R}^n \to \mathbb{R}$ to be the composition

$$f(x) = (h \circ g)(x) = h(g(x)).$$

- f is convex if h is convex, $\tilde{h}$ is non-decreasing, and g is convex
- f is convex if h is convex, $\tilde{h}$ is non-increasing, and g is concave
- f is concave if h is concave, $\tilde{h}$ is non-increasing, and g is convex
- f is concave if h is concave, $\tilde{h}$ is non-decreasing, and g is concave,

where $\tilde{h}$ refers to the extended-value extension of h,

$$\tilde{h}(x) = \begin{cases} h(x) & x \in \mathbf{dom}\, h \\ \infty & \text{otherwise} \end{cases}$$

Proof. We will prove the first statement only. First let $x, y \in \mathbf{dom}\, f$. This implies $a, b \in \mathbf{dom}\, g$, and that $g(a), g(b) \in \mathbf{dom}\, h$. Assuming that h is convex, $\tilde{h}$ non-decreasing, and that g is convex, we see for some $\theta \in [0, 1]$:

$$g(\theta x + (1 - \theta)y) \leq \theta g(x) + (1 - \theta)y.$$

The convexity of h gives us that the right hand side of this inequality is in $\mathbf{dom}\, h$, and by $\tilde{h}$ non-decreasing we see it must be that $g(\theta x + (1 - \theta)y) \in \mathbf{dom}\, h$ as well. But this means that $\theta x + (1 - \theta)y \in \mathbf{dom}\, f$. So $\mathbf{dom}\, f$ is convex. To show f is convex, we use the fact that $\tilde{h}$ is non-decreasing and notice

$$h(g(\theta x + (1 - \theta)y)) \leq h(\theta g(x) + (1 - \theta)y).$$

Lastly, by the convexity of h we have

$$h(\theta g(x) + (1 - \theta)y) \leq \theta(h(g(x)) + (1 - \theta)(h(g(y)),$$

so we see that f is indeed convex. ☐

Example 17.

- If g is a convex function then $e^{g(x)}$ is convex.
- if g is a convex function, $g \geq 0$, and $p \geq 1$, then $(g(x))^p$ is convex.

3.2.6 Vector Composition

Now we will consider the case for $k \geq 1$. In this case we have the composition

$$f(x) = h(g(x)) = h(g_1(x), g_2(x), \ldots, g_k(x)),$$

where each $g_i : \mathbb{R}^n \to \mathbb{R}$. In the case that the functions are twice differentiable and $\mathbf{dom}\, h = \mathbb{R}^k, \mathbf{dom}\, g = \mathbb{R}$ (where we again assume $n = 1$, considering restricting g to the intersection of a line and its domain), we have:

$$f''(x) = g'(x)^T \nabla^2 h(g(x)) g'(x) + \nabla h(g(x))^T g''(x), \tag{15}$$

the analogous equation to 14. Again, for f to be convex we need $f''(x) \geq 0$, which gives the following result:

- f is convex if h is convex, h is non-decreasing in each argument, and g_i is convex for each i.
- f is convex if h is convex, h is non-increasing in each argument, and g_i is concave for each i.
- f is concave if h is concave, h is non-increasing in each argument and g_i is convex for each i.
- f is concave if h is concave, h is non-decreasing in each argument and g_i is concave for each i.

We can also produce a similar result for $n > 1$, dropping the differentiablility conditions on g, h, and the domain restriction by forcing the monotonicity condition to apply for $\tilde{h}$.

3.2.7 Minimization

Given a family of convex functions, we saw that a maximum or supremum of the functions is convex. In the case of minimums and infimums, we have a different result.

Theorem 27. Let f be a convex function in (x, y), and let $C \neq \{\}$ be a convex set. Then the function

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex in x if $g(x) > -\infty$ for all x.

Proof. The domain of g is the set $\mathbf{dom}\, g = \{x \mid (x, y) \in \mathbf{dom}\, f, y \in C\}$. Let $x_1, x_2 \in \mathbf{dom}\, g$. Since g is the infimum of f over $y \in C$ at x_1, x_2 respectively, we know that for any $\varepsilon > 0$ there exists some $y_1, y_2 \in C$ so that $g(x_1) + \varepsilon \geq f(x_1, y_1)$, and $g(x_2) + \varepsilon \geq f(x_2, y_2)$. Let $\theta \in [0, 1]$, then by the convexity of f

$$f(\theta x_1 + (1 - \theta)x_2, \theta y_1 + (1 - \theta)y_2) \leq \theta f(x_1, y_1) + (1 - \theta) f(x_2, y_2) \leq \theta g(x_1) + (1 - \theta) g(x_2) + \varepsilon.$$

Now consider $g(\theta x_1 + (1 - \theta)x_2) = \inf_{y \in C} f(\theta x_1 + (1 - \theta)x_2, y)$. By the inequality above we see

$$g(\theta x_1 + (1 - \theta)x_2) \leq \theta g(x_1) + (1 - \theta) g(x_2) + \varepsilon$$

for all $\varepsilon > 0$, and thus $g(\theta x_1 + (1 - \theta)x_2) \leq \theta g(x_1) + (1 - \theta) g(x_2)$. So g is convex. $\square$

Example 18. The distance to a point x in a set $S \subseteq \mathbb{R}^n$ with any norm $\|\cdot\|$ on $\mathbb{R}^n$ can be written as

$$\mathbf{dist}(x, S) = \inf_{y \in S} \|x - y\|.$$

The norm is a convex function, and thus if S is convex, this function is convex as well.

3.2.8 Perspective of a function

Definition 32 (*Perpsective*)**.** If $f : \mathbb{R}^n \to \mathbb{R}$, then we say the *perspective* of f is the function $g : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by

$$g(x,t) = tf(x/t), \text{ with } \mathbf{dom}\, g = \{(x,t) \mid x/t \in \mathbf{dom}\, f, t > 0\}.$$

Lemma 2. If f is a convex function, then the perspective function of f is also convex.

Proof. Let g denote the perspective function of f. For some $(x,t,z) \in \mathbf{epi}\, g$, we have that $tf(x/t) \leq z \Leftrightarrow f(x/t) \leq z/t$, which happens if and only if $(x/t, z/t) \in \mathbf{epi}\, f$. Thus $\mathbf{epi}\, g$ is convex because it is the inverse image of $(x,z)/t$ under the mapping $(x,t,z) \mapsto (x,z)/t$, and so g is convex. (See 10). $\square$

3.3 The Conjugate Function

Definition 33. Let $f : \mathbb{R}^n \to \mathbb{R}$. The *conjugate* of the function f, $f^* : \mathbb{R}^n \to \mathbb{R}$, defined as

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} y^T x - f(x), \tag{16}$$

with $\mathbf{dom}\, f^* = \{x \in \mathbf{dom}\, f \text{ so that } f^*(y) \in \mathbb{R}\}$.

Note that f^* *is always a convex function* since $y^T x - f(x)$ is an affine function in y for each y, and we are taking a pointwise supremum of these functions. Notice also that f itself need not be convex for this to be true.

Example 19.

- Let $f(x) = ax + b$. Then $f^*(y) = \sup_{x \in \mathbb{R}} yx - (ax+b)$ is only bounded above when $y = a$, and thus is only defined on the domain $x = a$, in which case $f^*(a) = -b$.
- Consider the convex function $f(x) = -\log(x)$. In this case we have

  $$f^*(y) = \sup_{x \in \mathbb{R}_{++}} yx + \log(x),$$

  where $\frac{d}{dx}\left(yx + log(x)\right) = y + 1/x = 0 \implies x = -1/y$, and thus we have

  $$f^*(y) = -1 - \log(-y)$$

  for $y < 0$.
- Let $f(x) = (1/2)x^T Q x$ for $Q \in \mathbf{S}^n_{++}$. Then the function

  $$y^T x - (1/2)x^T Q x$$

  is bounded above as a function of x for any y, and attains its maximum at $x = Q^{-1}y$, thus we see

  $$f^*(y) = y^T Q^{-1} y - (1/2) y^T (Q^{-1})^T Q Q^{-1} y = (1/2) y^T Q^{-1} y.$$

Example 20. Consider a business which consumes n resources and produces a product to be sold. Let $r = (r_1, \ldots, r_n)$ denote the vector of resources consumed, and $S(r)$ denote the sales revenue derived from the product (which is a function of the resources consumed). Let p_i denote the price per unit of resource i, so that the total amount paid for resources can be written as $p^T r$. The profit derived by the firm can then be written as the function $P(r) = S(r) - p^T r$. We would like to know the optimal amount of reach resource r_i to consume in order to maximize our profit. We can denote the maximum profit function as

$$M(p) = \sup_r (S(r) - p^T r),$$

which can be realized in terms of the conjugate function of $S(r)$ by

$$M(p) = (-S)^*(-p).$$

Example 21. Let $f(x) = x^p$ on $\mathbb{R}_{++}$, for $p > 1$. Then we see that for $y \leq 0$, the function $yx - x^p$ achieves its maximum at $x = 0$, thus $f^*(y) = 0$. For $y > 0$, the function attains its maximum at $x = (y/p)^{1/(p-1)}$, giving a value of

$$y(y/p)^{1/(p-1)} - (y/p)^{p/(p-1)}.$$

Letting $q = p/(p-1)$ we see $q - 1 = 1/(p-1)$ and the above equation becomes

$$(y/p)^q(p-1).$$

Thus our conjugate function $f^*(y)$ is given by

$$f^*(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ (y/p)^q(p-1) & \text{if } y > 0. \end{cases}$$

3.3.1 Basic Properties

From the definition of the conjugate function, we have

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} y^T x - f(x),$$

which leads immediately to the inequality

$$f(x) + f^*(y) \geq x^T y \quad \forall x, y.$$

This is called *Fenchel's Inequality.*

Example 22. Considering again the quadratic function $f(x) = (1/2)x^T Qx$ for some $Q \in \mathbf{S}^n_{++}$, we have the inequality

$$(1/2)x^T Qx + (1/2)y^T Q^{-1} y \geq x^T y,$$

for all $x, y \in \mathbb{R}^n$.

In general it is typical to associate the conjugate of a mathematical object with the property that the conjugate of the conjugate is the object itself. The following theorem characterizes exactly when this occurs.

Theorem 28. If f is convex and $\mathbf{epi}\, f$ is closed, then $f = f^{**}$.

If f is a differentiable function then we refer to f^* as the *Legendre transform* of f.

Theorem 29. Suppose f is convex, differentiable, and $\mathbf{dom}\, f = \mathbb{R}^n$. Then any maximizer x^* of $y^T x - f(x)$ satisfies $y = \nabla f(x^*)$, and if x^* satisfies $y = \nabla f(x^*)$, then x^* maximizes $y^T x - f(x)$. Thus if $y = \nabla f(x^*)$ then

$$f^*(y) = \nabla f(x^*)^T x^* - f(x^*).$$

Proof. Let f be a convex, differentiable function with $\mathbf{dom}\, f = \mathbb{R}^n$. Then suppose x^* maximizes $y^T x - f(x)$. So

$$y - \nabla(f(x^*) = 0 \implies y = f(x^*).$$

Conversely, suppose that x^* satisfies $y = f(x^*)$. Then $y - \nabla f(x^*) = 0$ which implies that x^* is a maximizer of the function $y^T x - f(x)$. $\square$

This result allows us to determine $f^*(y)$ for any $y \in \mathbb{R}^n$ for which we can solve the equation $y = \nabla f(z)$ for z. To see this more clearly, suppose $z \in \mathbb{R}^n$ and define $y = \nabla f(z)$. Then

$$f^*(y) = z^T \nabla f(z) - f(z).$$

3.4 Quasiconvex Functions

Definition 34 (*Quasiconvex, Quasiconcave, Quasilinear*)**.**

- A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is called *quasiconvex* if its domain and all of its sublevel sets, C_α, $\alpha \in \mathbb{R}$ are convex.
- A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is called *quasiconcave* if its domain and all of its superlevel sets, S_α, $\alpha \in \mathbb{R}$ are convex.
- If a function $f : \mathbb{R}^n \to R$, is both quasiconvex and quasiconcave, we say f is *quasilinear* and note that every level set $\{x \mid f(x) = \alpha\}$ is convex.

Example 23. The ceiling function $ceil(x) = \inf\{z \in \mathbb{Z} \mid z \geq x\}$ is quasilinear. See:

$$C_\alpha = \{x \mid \mathrm{ceil}(x) = \alpha\} = \begin{cases} \{\} & \text{if } \alpha \notin \mathbb{Z} \\ \{\alpha\} & \text{if } \alpha \in \mathbb{Z} \end{cases}.$$

Example 24. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x, y) = xy,$$

with $\mathbf{dom}\, f = \mathbb{R}^2_{++}$.

Then

$$\nabla^2 f(x) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which is neither positive nor negative semidefinite. So f is not convex nor concave.

For $\alpha \in \mathbb{R}$ we have $C_\alpha = \{(x, y) \in \mathbb{R}^2_{++} \mid y \leq \alpha/x\}$ which is concave, so f is not quasiconvex but f is quasiconcave.

Example 25. Let $f : \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}$ be defined by $f(x, y) = x^2/y$.

Then

$$\nabla^2 f(x) = 2 \begin{bmatrix} 1/y & -x/y^2 \\ -x/y^2 & x^2/y^3 \end{bmatrix},$$

which is positive semidefinite since $1/y, x^2/y^3 \geq 0$ and $\det\big(\nabla^2 f(x)\big) \geq 0$. So f is convex.

Example 26. (Approximation Width) Let $f_0, f_1, dots, f_n : \mathbb{R}^n \to \mathbb{R}$ be continuous functions. For some $x \in \mathbb{R}^n$ we say that $f = x_1 f_1 + x_2 f_2 + \cdots + x_n f_n$ approximates f_0 with tolerance $\varepsilon > 0$ over $[0, T]$ if

$$|f(t) - f_0(t)| \leq \varepsilon$$

for every $t \in [0, T]$. Fixing $\varepsilon > 0$ we say the *approximation width* as the largest T so that f approximates f_0 over the interval $[0, T]$. We can define this function as follows:

$$W_\varepsilon(x) = \sup_T \{|x_1 f_1 + x_2 f_2 + \cdots + x_n f_n - f_0| \leq \varepsilon \text{ for all } t \in [0, T]\}.$$

It follows that W_ε is quasiconcave.

Proof. Let $\alpha \in \mathbb{R}$. Then we see

$$\{x \mid W_\varepsilon(x) \geq \alpha\} = \bigcap_{0<t<\alpha} \{x \mid |x_1 f_1(t) + x_2 f_2(t) + \cdots + x_n f_n(t) - f_0(t)| \leq \varepsilon\}.$$

Since this function is convex in x is follows each set in the intersection is convex and thus W_ε is quasiconvex. □

3.4.1 Basic Properties

Theorem 30. A function f is quasiconvex if and only if $\mathbf{dom}\, f$ is convex and for any $x, y \in \mathbf{dom}\, f$ and $\theta \in [0, 1]$ it follows,

$$f(\theta x + (1-\theta)y) \le \max\{f(x), f(y)\}. \tag{17}$$

This says that the value of the function on a segment does not exceed the maximum of its values on the end points.

Example 27. Let $A, B \in \mathbf{S}_+^n$ and consider the **rank** function. We see that

$$\mathbf{rank}(A+B) \ge \min\{\mathbf{rank}(A), \mathbf{rank}(B)\},$$

and thus is quasiconcave.

Analogous to convexity, we can characterize quasiconvexity by quasiconvexity restricted to $\mathbb{R}$.

Theorem 31. A function f is quasiconvex if and only if its restriction to any line intersecting its domain is convex.

3.4.2 First and Second Order Conditions

We state first and second order conditions for quasiconvexity without proof.

Theorem 32 (First Order Conditions)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then f is quasiconvex if and only if $\mathbf{dom}\, f$ is convex and

$$f(y) \le f(x) + \nabla f(x)^T (y - x) \le 0.$$

Theorem 33. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. Then if f is quasiconvex, then for all $x \in \mathbf{dom}\, f$, $y \in \mathbb{R}^n$

$$y^T f(x) = 0 \implies y^T f(x) y \ge 0.$$

3.5 Convexity with Respect to Generalized Inequalities

3.5.1 Monotonicity with Respect to a generalized inequality:

Suppose K is a proper cone with associated generalized inequality $\preccurlyeq_K$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *K-non-decreasing* if

$$x \preccurlyeq_K y \implies f(x) \le f(y),$$

and *K-increasing* if

$$x \preccurlyeq_K y,\, x \ne y \implies f(x) < f(y).$$

K-non-increasing and *K-decreasing* are defined similarly.

Example 28. $\operatorname{tr}(X^{-1})$ is decreasing on $\mathbf{S}_{++}^n$.

To see this, let X and Y be two positive definite $n \times n$ matrices so that $X \preccurlyeq_{\mathbf{S}_{++}^n} Y$. So $\operatorname{tr}(X) \le \operatorname{tr}(Y)$. and, since both X^{-1} and Y^{-1} exists and are positive definite, $\operatorname{tr}(X^{-1}) \le \operatorname{tr}(Y^{-1})$.

Theorem 34 (Gradient conditions for monotonicity)**.** Let f be a differential function with a convex domain. Then f is K-non-decreasing if and only if

$$\nabla f(x) \succcurlyeq_{K^*} 0, \tag{18}$$

for all $x \in \mathbf{dom}\, f$.

Proof. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable with a convex domain.

($\Longleftarrow$) First suppose that (18) holds for all $x \in \mathbf{dom}\, f$. Then, seeking a contradiction, assume that f is not K-non-decreasing. This implies that there exists some $x, y \in \mathbf{dom}\, f$ so that $x \preccurlyeq_K y$ and $f(x) > f(y)$. Since f is differentiable we know that there exists some $t \in [0, 1]$ so that

$$\frac{d}{dt} f(x + t(x - y)) = \nabla f(x + t(y - x))^T (y - x) < 0.$$

Also, since $y - x \in K$, this implies

$$\nabla f(x + t(x - y)) \notin K^*,$$

which contradicts our assumption. Thus f is K-non-decreasing.

($\Longrightarrow$) Now suppose that f is K-non-decreasing. Again seeking a contradiction, assume that

$$\nabla f(x) \not\succcurlyeq_{K^*} 0$$

for some $x \in \mathbf{dom}\, f$. Choose such an x. So there exists some $v \in K$ so that

$$\nabla f(x)^T v < 0.$$

Now define $h(t)$ by

$$h(t) = f(x + tv).$$

Then $h'(t) = \nabla f(x)^T v < 0$. Thus there exists some $t > 0$ so that

$$h(t) = f(x + tv) < h(0) = f(x),$$

thus f is not K-non-decreasing, a contradiction. $\square$

3.5.2 Convexity with Respect to a Generalized Inequality

Definition 35 (*K-convex*)**.** Let $K \subseteq \mathbb{R}^m$ be a proper cone with the associated generalized inequality $\preccurlyeq_K$. We say that a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *K-convex* if for all x, y, and $\theta \in [0, 1]$,

$$f(\theta x + (1 - \theta)y) \preccurlyeq_K \theta f(x) + (1 - \theta) f(y).$$

We also say that the function if *strictly K-convex* if it satisfies the strict general inequality above.

Example 29. Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$. f is convex with respect to the component wise inequality (the generalized inequality induced by $\mathbb{R}^m_+$) if and only if for all x, y, and $\theta \in [0, 1]$, it follows

$$f(\theta x + (1 - \theta)y) \prec \theta f(x) + (1 - \theta) f(y).$$

Notice that this will be true if and only if f is convex in each component. i.e. if

$$f(x) = (f_1(x_1), f_2(x_2), \ldots, f_m(x_m)),$$

then each f_i must be convex.

Example 30. The function $f(X) = X^{-1}$ is matrix convex on $\mathbf{S}^n_{++}$.

4 Convex Optimization Problems

There is something to be said for rewarding hard work. For us, the fruit of our labor to understand convex sets and functions comes in the form of convex optimization problems. These are a particular family of optimization problems for which we will be able to fully discuss solvability criteria. In many cases we will be able to prove the existence of a unique global minimum solution. Of course, the existence of a solution does not provide the solution itself, so we will still be left with some hard work to do. But remember, hard work means good payoff. This chapter defines a convex optimization problem and discusses what can be said about this family of problems. It also includes several examples of well known convex problems for which we can provide a solution.

4.1 Optimization Problems

As we have seen before, a general optimization problem is described as follows:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, i = 0, \ldots, n, \\ & h_i(x) = 0, i = 0, \ldots, m. \end{aligned} \tag{19}$$

This describes the problem of finding an x that minimizes $f_0(x)$ while satisfying the conditions $f_i(x) \leq 0, h_i(x) = 0$ for all i. The variable $x \in \mathbb{R}^n$ is the *optimization variable*, and the function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the *objective function* or *cost function*. We refer to the inequalities $f_i(x) \leq 0$ as the *inequality constraints*, and the functions f_i as the corresponding *inequality constraint functions.* Lastly, the equations $h_i(x) = 0$ are referred to as the *equality constraints* and the functions given by h_i are the *equality constraint functions.* If an optimization problem has no constraints (if $n = m = 0$) then we say the problem is *unconstrained.*

A particular x is said to be *feasible* if $x \in \mathcal{D}$, the domain of the problem, and x satisfies all the constraints. We define the domain as you would expect:

$$\mathcal{D} = \left(\bigcap_{i=0}^{n} \mathbf{dom}\, f_i \right) \cap \left(\bigcap_{i=0}^{m} \mathbf{dom}\, h_i \right)$$

The optimization problem is said to be *feasible* if there is at least one feasible point. Otherwise it is said to be *infeasible.* The set of all feasible points is called the *feasible set* or *constraint set.*

The *optimal value*, $p^* \in \overline{\mathbb{R}}$, of an optimization problem is defined as

$$\inf\{f_0(x) \mid f_i(x) \leq 0, i = 1, \ldots, n,\ h_i(x) = 0, i = 1, \ldots, m\}.$$

If the feasible set is empty then $p^* = \infty$, and if $p^* = \infty$ then the problem is *unbounded below.*

An *optimal point*, x^*, is a feasible point such that $f_0(x^*) = p^*$. An optimal point is said to be a solution to the optimization problem. We denote the set of all optimal points as

$$X_{\text{opt}} = \{x \mid f_i(x) \leq 0, i = 1, \ldots, n,\ h_i(x) = 0, i = 1, \ldots, m,\ f_0(x) = p^*\}.$$

If X_{opt} is non-empty then we say the optimal value is *achieved* and that the problem is *solvable.* If X_{opt} is empty then the optimal value is not achieved. For example, this will occur when the problem us infeasible and when the problem is unbounded below.

Lastly, we will often be interested in (19) where $f_0(x)$ is identically zero. This is referred to as the *feasibility problem* because any feasible point will also be an optimal point, thus the solutions to this problem are exactly the points in the feasible set.

4.2 Convex Optimization Problems

4.2.1 Standard Form

Definition 36 (*Convex Optimization Problem*)**.** A *convex optimization problem*, in standard form, is written as:

$$\begin{aligned}
\underset{x}{\text{minimize}} \quad & f_0(x) \\
\text{subject to} \quad & f_i(x) \leq 0, \quad i = 0, \ldots, m, \\
& a_i^T x = b_i, \quad i = 0, \ldots, p,
\end{aligned} \tag{20}$$

where $f_0, f_1, \ldots, f_m$ are all convex functions.

A convex optimization problem differs from a standard optimization problem in that the objective function and the inequality constraint functions must all be convex, and in that the equality constraint functions must be affine.

An immediate consequence of this definition is that the feasible set of a convex optimization problem is convex.

Lemma 3. Given the standard convex optimization problem (20), the domain of the problem $\mathcal{D}$, and the feasible set are both convex sets.

Proof. The domain of this problem is the intersection

$$\mathcal{D} = \left(\bigcap_{i=0}^{n} \mathbf{dom}\, f_i \right) \cap \left(\bigcap_{i=0}^{m} \mathbf{dom}\, h_i \right).$$

Since each f_i is convex, and each h_i is affine, this is an intersection of convex sets and thus is convex.

We now note that feasible set of this problem can be written as the intersection

$$\mathcal{D} \cap \left(\bigcap_{i=1}^{m} \{x \mid f_i(x) \leq 0\} \right) \cap \left(\bigcap_{i=1}^{p} \left\{x \mid a_i^T x = b_i \right\} \right),$$

which is the intersection of $\mathcal{D}$, 0-sublevel sets of convex functions, and hyperplanes, all of which are convex sets. □

Therefore to solve a convex optimization problem we must minimize a convex function (the objective function) over a convex set (the feasible set). We now show that any local minimum in the feasible set must be a global minimum.

Theorem 35. Give the standard convex optimization problem (20), any local minima $\tilde{x}$ of f_0 in the feasible set is a global minima of f_0 over the feasible set.

Proof. We will assume $\tilde{x}$ is feasible and is a local minima of f_0 with some radius ρ. So for all x such that $\|\tilde{x} - x\| \leq \rho$ we have $f_0(x) > f_0(\tilde{x})$. Then, for the sake of contradiction, suppose that $x^* \in X_{\text{opt}}$ is such that

$$f_0(\tilde{x}) > f_0(x^*).$$

Since $\tilde{x}$ and x^* are both feasible, and the feasible set is convex per the preceding lemma, we have that

$$\theta\tilde{x} + (1-\theta)x^*$$

is feasible for any $\theta \in [0, 1]$. Since f_0 is convex we have that

$$\begin{aligned}
f_0(\theta\tilde{x} + (1-\theta)x^*) &\leq \theta f_0(\tilde{x}) + (1-\theta) f_0(x^*) \\
&\leq f_0(\tilde{x}) \text{ for any } \theta \in [0, 1].
\end{aligned}$$

So choose θ to be such that $\|\tilde{x} - (\theta\tilde{x} + (1-\theta)x^*)\| \leq \rho$. Then it follows that

$$f_0(\theta\tilde{x} + (1-\theta)x^*) > f_0(\tilde{x}),$$

which is a contradiction. Therefore it must be that $f_0(\tilde{x}) \leq f_0(x^*)$. Since x^* is a global minimizer of f_0 over the feasible set, we conclude

$$f_0(\tilde{x}) = f_0(x^*),$$

so that $\tilde{x} \in X_{\text{opt}}$. □

Remark. The preceding result is informs our definition of a convex optimization problem. In order to show that any local minima is a global minima, we needed the feasible set to be convex, which is why we require each inequality constraint function to be convex, and each equality constraint function to be affine. The importance of the convexity of the objective function f_0 is also very clear.

If the objective function f_0 is strictly convex and feasible, then the problem has a unique global minimum solution.

Theorem 36. If f_0 is strictly convex then X_{opt} contains at most one point.

Proof. Suppose X_{opt} is non-empty and let $x^* \in X_{\text{opt}}$. Then for any feasible x such that $x \neq x^*$, and for any $\theta \in [0,1]$, it follows

$$\begin{aligned} f_0(\theta x^* - (1-\theta)x) &< \theta f_0(x^*) + (1-\theta)f_0(x) \\ &\leq f_0(x), \end{aligned}$$

and therefore

$$f_0(x^*) < f_0(x).$$

□

We may also talk about *concave maximization problems*, which are of the form:

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 0, \ldots, n, \\ & h_i(x) = 0, \quad i = 0, \ldots, m, \end{aligned} \tag{21}$$

where f_0 is concave. By forming the equivalent problem of minimizing $-f_0(x)$ we have the same results as above for concave maximization problems.

If we allow f_0 to be quasiconvex, then we call the optimization problem a *quasiconvex optimization problem*.

4.2.2 Common Convex Optimization Problems

When the objective and constraint functions are all affine, we call this a *Linear Program*, or *LP*. e.g., the optimization problem given by

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & c^T x + d \\ \text{subject to} \quad & Gx \preccurlyeq 0, \\ & Ax = b, \end{aligned} \tag{22}$$

is the standard form for a LP.

Linear Programs are arguably the best understood convex optimization problems. There are various robust methods for solving these sort of problems in a number of variables that is quite appalling to look at.

A well known method for solving linear programs is the *Simplex Method*, which takes advantage of the fact that the feasible set of a linear program forms a polyhedron and is optimized on one of the vertices of this polyhedron.

Theorem 37. Given the standard linear program, (22), the feasible set is a polyhedron and, if the set is non-empty and bounded, there exists an optimal solution x^*. Furthermore, any optimal solution x^* can be assumed to lie on a vertex of the feasible set.

Proof. The feasible set $\mathcal{D}$ is the set given by

$$\mathcal{D} = \{x \mid Ax = b, \quad Gx \preccurlyeq 0\},$$

which is an intersection of hyperplanes and closed halfspaces, and thus a polyhedron. If $\mathcal{D}$ is non-empty and bounded then it is compact, and thus $f(x) = c^T x + d$ attains an absolute minimum x^* for some $x \in \mathcal{D}$. Now suppose that $x^* \in X_{opt}$. First, we note that the objective function has a constant gradient given by

$$\nabla(c^T + d) = c,$$

and thus we have no interior extrema. This proves that $x^* \in \mathbf{bd}\,\mathcal{D}$. Now suppose that x^* does not lie on a vertex of $\mathcal{D}$, and let $\hat{x}$ be a vector along the boundary of $\mathcal{D}$ such that $x^* + \hat{X} \in \mathcal{D}$. Here we have several cases.

Case 1: If $\hat{x}$ is parallel to c then

$$c^T(x^* + \alpha\hat{x}) + d = c^T x^* + d = p^*.$$

for any α such that $x^* + \alpha\hat{x} \in \mathcal{D}$. Thus we may assume our optimal value is $x^* + a\hat{x}$ where this new point is moved as far in the $\hat{x}$ direction in $\mathcal{D}$ as possible.

Case 2: If $\hat{x}$ is not parallel to c, then either $c^T\hat{x} > 0$ or $c^T\hat{x} < 0$. We may assume $c^T\hat{x} < 0$. Then for some $\alpha > 0$ it follows that $x^* + \alpha\hat{x} \in \mathcal{D}$ and that

$$c^T(x^* + \alpha\hat{x}) < c^T x^* + d = p^*,$$

so $x^* \notin X_{opt}$, a contradiction. □

4.2.3 Examples of Optimization Problems

Example 31. Consider the optimization problem

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & f_0(x_1, x_2) \\ \text{subject to} \quad & 2x_1 + x_2 \geq 1, \\ & x_1 + 3x_2 \geq 1, \\ & x_1, x_2 \geq 0 \end{aligned}$$

We will sketch the feasible region and find the optimal value for each of the following objective functions:

1. $f_0(x_1, x_2) = x_1 + x_2$,
2. $f_0(x_1, x_2) = -x_1 - x_2$,
3. $f_0(x_1, x_2) = x_1$,
4. $f_0(x_1, x_2) = \max\{x_1, x_2\}$,
5. $f_0(x_1, x_2) = x_1^2 + 9x_2^2$.

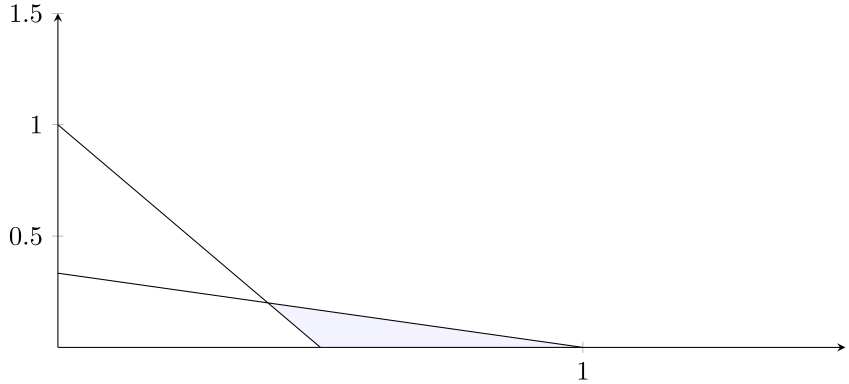


Figure 3: Feasible region of the the optimization problem, shaded in blue.

The first 3 objective functions are linear, and each constraint is linear as well, so the optimal value, if attainable, will lie on a vertex of the feasible region. In the case of (1), $X_{\text{opt}} = \{(2/5, 1/5)\}$ and thus the optimal value is 3/5. For (2) the function is unbounded below and therefore the optimal value is $-\infty$ and is not achieved. For (3) $X_{\text{opt}} = \{(0, x_2) \mid x_2 \geq 0\}$ and $p^* = 1$. For the max function we are concerned with the smallest pair (x, x) so that (x, x) is feasible. This occurs at $(1/3, 1/3)$ thus $X_{\text{opt}} = \{(1/3, 1/3)\}$ and $p^* = 1/3$. The last objection function is quadratic, and we maximize this over the feasible region by finding where the gradient is perpendicular to the curve defining the region. First we see $\nabla f_0 = (2x_1, 18x_2)$. This is perpendicular to the line $2x_1 + x_2 = 1$, or $x_2 = 1 - 2x_1$, when

$$\begin{aligned} 2x_1 &= 1 \\ 18x_2 &= 1/2 \\ &\implies x_1 = 1/2, x_2 = 1/36. \end{aligned}$$

But the point $(1/2, 1/36)$ is not in the feasible region as $1/2 + 1/12 < 1$. The gradient is perpendicular to the line $x_1 + 3x_2 = 1$, or $x_2 = 1/3 - x_1/3$, when

$$\begin{aligned} 2x_1 &= 1 \\ 18x_2 &= 3 \\ &\implies x_1 = 1/2, x_2 = 1/6. \end{aligned}$$

This point is in the feasible region and thus $X_{\text{opt}} = \{(1/2, 1/6)\}$ and $p^* = 1/2$.

Example 32 (Symmetries and Convex Optimization)**.** Suppose that $G = \{Q_1, Q_2, \ldots, Q_k\} \subseteq \mathbb{R}^{n \times n}$ is a group. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is G-invariant, or symmetric with respect to G, if $f(Q_i x) = f(x)$ for all $i = 1, \ldots, k$. We define $\overline{x} = 1/k \sum_{i=1}^{k} Q_i x$, the average of x over its G-orbit, and label the fixed point set of G as $\mathcal{F}$. Recall

$$\mathcal{F} = \{x \in \mathbb{R}^n \mid Q_i x = x \text{ for } i = 1, \ldots, k\}.$$

We first note that for any $x \in \mathbb{R}^n$ it is the case that $\overline{x} \in \mathcal{F}$.

Proof. Let $x \in \mathbb{R}^n$. Then for any $Q_j \in G$ we have

$$
\begin{aligned}
Q_j\overline{x} &= Q_j\left(1/k\sum_{i=1} kQ_ix\right)\\
&= 1/k\sum_{i=1} kQ_jQ_ix\\
&= 1/k\sum_{i=1} kQ_ix\\
&= \overline{x}.
\end{aligned}
$$

□

Further, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex and G-invariant, then it follows that $f(\overline{x}) \leq f(x)$.

Proof. We proceed directly and see

$$
\begin{aligned}
f(\overline{x}) &= f(1/k\sum_{i=1}^{k} Q_ix),\\
&\leq 1/k\sum_{i=1}^{k} f(Q_ix),\\
&= 1/k\sum_{i=1}^{k} f(x),\\
&= f(x),
\end{aligned}
$$

where the first inequality comes from convexity, and the following equality comes from G-invariance.

□

An optimization problem,

$$
\begin{array}{ll}
\underset{x}{\text{minimize}} & f_0(x)\\
\text{subject to} & f_i(x) \leq 0, k = 0, \dots, m,
\end{array}
\tag{23}
$$

is said to be G-invariant if the objective function, f_0, is G-invariant and the feasible set is G-invariant. That is to say $f(x) = f(Q_ix)$ and

$$f_1(x) \leq 0,\ \dots, f_m(x) \leq 0 \Leftrightarrow f_1(Q_ix) \leq 0, \dots, f_m(Q_ix) \leq 0$$

for all $Q_i \in G$. If an optimization problem is convex and G-invariant, and there exists an optimal point, then there exists an optimal point in $\mathcal{F}$. This implies we may optimize f_0 over the fixed point set of G, rather than the feasible set, without loss of generality.

Proof. Suppose the above optimization problem is a convex and G-invariant optimization problem with $X_{\text{opt}} \neq \{\}$. Let $x \in X_{\text{opt}}$. By the first proof in this example we see that $\overline{x} \in \mathcal{F}$. Then, by the previous proof we have $f_0(\overline{x}) \leq f_0(x)$. Because the problem is G-invariant and convex we have that

$$f_j(\overline{x}) \leq 1/k\sum_{i=1}^{k} f_j(Q_ix) \leq 0,$$

where the last inequality comes from x feasible and the feasible set being G-invariant. Thus $\overline{x}$ is a feasible point, therefore

$$f(\overline{x}) \leq f(x) \implies f(\overline{x}) = f(x),$$

and we conclude $\overline{x} \in X_{\text{opt}} \cap \mathcal{F}$. □

As an example, consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(Px) = f(x)$ for every permutation P (in this case we say f is symmetric). If f has a minimizer then it has a minimizer of the form $\alpha\mathbb{1}$, i.e. we can solve the problem by minimizing $f(t\mathbb{1})$ over $t \in \mathbb{R}$. This follows immediately from the previous result by noticing $G = S_n$ and $\overline{x} = 1/n!(\sum_{i=1}^{n!} x_i, \ldots, \sum_{i=1}^{n!} x_i)$.

Example 33. (Analytic Solutions To Linear Programs)

(a) Minimizing over an affine set.

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & Ax = b. \end{array}$$

Let's consider three cases

(i) If $b \notin \mathcal{R}(A)$ then the problem is infeasible and thus $p^* = -\infty$.

(ii) Rewriting c as $c = A^T\lambda + \hat{c}$ for $c \in \mathcal{R}(A)^\perp$, then if $\hat{c} \neq 0$ the problem is unbounded below. To see this, let $x(t) = x_0 - t\hat{c}$ and let $t \to \infty$. Then $x(t)$ is feasible for all t, and as $t \to \infty$ $c^T x \to \infty$, so $p^* = -\infty$.

(iii) Lastly, if $\hat{c} = 0$ then the optimal value is λb for any feasible x.

In conclusion, the solution to this problem is

$$p^* = \begin{cases} +\infty, & b \notin \mathcal{R}(A), \\ \lambda^T b, & c = A^T\lambda \text{ for some } \lambda, \\ -\infty, & \text{otherwise.} \end{cases}$$

(b) Minimizing a linear function over a halfspace.

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & a^T x \leq b \end{array}$$

for $a \neq 0$. Again, we have a few cases to consider. First, we will rewrite c as $c = a\lambda + \hat{c}$ for some $\lambda \in \mathbb{R}$ and $\hat{c}$ orthogonal to a.

(i) If $\lambda > 0$, by setting $x(t) = -ta$, we see $c^T x = -t\lambda a^T a$ approaches $-\infty$ for large t and eventually $-ta^T a - b \leq 0$, thus $x(t)$ is feasible for sufficiently large t. Therefore $p^* = -\infty$.

(ii) If $\hat{c} \neq 0$ then we again see, for $x(t) = -ba - t\hat{c}$, as we let $t \to \infty$, $c^T x = -c^T(ba) - tc^T\hat{c}$, which is unbounded below.

(iii) If $c = a\lambda$ for $\lambda < 0$, we want to minimize $c^T x$ such that $a^T x = 1/\lambda c^T \leq b$, or $c^T x \geq b$. Thus the optimal value is $c^T x = \lambda b$.

In conclusion, we have

$$p^* = \begin{cases} \lambda b, & c = a\lambda,\ \lambda < 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

(c) Minimizing a linear function over a rectangle.

$$\begin{array}{ll} \underset{x}{\text{minimize}} & c^T x \\ \text{subject to} & l \preccurlyeq x \preccurlyeq u. \end{array}$$

To solve this problem, for each entry x_i in x we let

$$x_i = \begin{cases} l_i, & c_i > 0, \\ u_i, & c_i < 0, \\ a \in [l_i, u_i], & c_i = 0. \end{cases}$$

Any x in the form above is feasible and we see $p^* = l^T c^+ + u^T c^-$ where $c_i^+ = \max\{c_i, 0\}$ and $c_i^- = \max\{-c_i, 0\}$.

(d) Minimizing a linear function over the probability simplex

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & c^T x \\ \text{subject to} \quad & \mathbb{1}^T x = 1, \\ & x \succcurlyeq 0. \end{aligned}$$

Let $c_{min} = \min\{c_1, c_2, \ldots, c_n\}$. Then the optimal value of this problem is $p^* = c_{min}$ which is achieved for any x so that the sum of the components corresponding to c_{min} is 1, while the rest are zero. For example, if the components of c were ordered as such:

$$c_1 = c_2 = c_3 = \cdots = c_k < c_{k+1} \leq \cdots \leq c_n$$

for some $k \geq 1$ then we see for any feasible $x, c^T x \geq c_1^T x = c_{min}$, and we have equality if and only if $\sum_{i=1}^k x_i = 1$, and $x_{k+1} = \cdots = x_n = 0$. This problem can be thought of as the optimal portfolio investment strategy, where $-c_i$ corresponds to the return of investing in asset i, and x_i is the fraction of our total fund invested in asset i. Hence we seek to maximize $-c^T x$. In the case above we simply invest in the asset, or possibly assets, with the best return.

If the equality constraint is replaced by the inequality constraint $\mathbb{1}^T x \preccurlyeq 0$ then, in the portfolio interpretation, we are no longer required to invest all of our funds. Now, if $c_{min} \geq 0$ we have that $p^* = 0$ corresponding to $x = 0$. Otherwise $c_{min} < 0$ and we have the same solution as before.

Example 34. (Optimal Activity Levels) Consider the vector $x \in \mathbb{R}^n$ as a set of activity levels $x_1, x_2, \ldots, x_n$. These activities consume m limited resources. The amount of resource i consumed by activity j is given by $A_{ij} x_j$ for some known $A \in \mathbb{R}^{m \times n}$. We will allow A_{ij} to be negative and interpret $A_{ij} < 0$ to mean that activity j generates some amount of resource i. Because each resource is limited we also have $c_i \leq c_i^{\max}$, where $c_i^{\max}$ is known for each resource. Each activity generates revenue as a piece-wise concave function of the activity, given below:

$$r_j(x_j) = \begin{cases} p_j x_j & 0 \leq x_j \leq q_j \\ p_j q_j + p_j^{\text{disc}}(x_j - q_j) & x_j \geq q_j, \end{cases}$$

where $p_j \in \mathbb{R}^n$ is the price for activity j and $q_j \in \mathbb{R}^n$ is the quantity discount level, and $0 < p_j^{\text{disc}} < p_j$ is the quantity discount price for activity j.

The goal is to maximize the total revenue,

$$R(x) = \sum_{j=1}^n r_j(x_j)$$

under the given constraints.

Because the objective function is concave, we can write this as a convex optimization problem as follows:

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & R(x) \\ \text{subject to} \quad & Ax \preccurlyeq c^{\max}, \\ & x \succcurlyeq 0. \end{aligned}$$

Furthermore we notice that we can express each $r_j(x_j)$ in the objective function by the minimum function

$$r_j(x_j) = \min\{p_j x_j, p_j q_j + p_j^{\text{disc}}(x_j - q_j)\},$$

which is valid since r_j is concave. Then $r_j(x_j) \geq u_j$ if and only if $p_j x_j \geq u_j$ and $p_j q_j + p_j^{\text{disc}}(x_j - q_j) \geq u_j$. Thus we can form the problem

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & \mathbb{1}^T u \\ \text{subject to} \quad & Ax \preccurlyeq c^{\max}, \\ & x \succcurlyeq 0, \\ & p_j x_j \geq u_j, \\ & p_j q_j + p_j^{\text{disc}}(x_j - q_j) \geq u_j, \quad j = 1, \ldots, n, \end{aligned}$$

which is a linear program.

For a concrete instance, lets suppose that a certain yard service company is attempting to increase revenue–something most companies wish to do. Consider each activity level to be the average number of hours employee j works per day, so that $A_{ij}x_j$ represents the average amount of resource c_i used by employee j on a given day. In this way q_j is the number of hours after which employee j works for a discounted wage, and p_j^{disc} represents the discounted hourly price for employee j's labor. Thus the optimal solution is the average number of hours each employee should work per day in order for the company to maximize revenue. This assumes each employee is taking and / or providing resources from the same pool and that each employee has a distinct hourly wage, discounted wage, and threshold for discounted work.

4.2.4 Solving Common Convex Optimization Problems

Example 35. Let's consider the LP from (34) for the given parameters:

$$A = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 3 & 1 \\ 0 & 3 & 1 & 1 \\ 2 & 1 & 2 & 5 \\ 1 & 0 & 3 & 2 \end{bmatrix}, \; c^{\max} = \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ 100 \end{bmatrix}, \; p = \begin{bmatrix} 3 \\ 2 \\ 7 \\ 6 \end{bmatrix}, \; p^{\text{disc}} = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 2 \end{bmatrix}, \; q = \begin{bmatrix} 4 \\ 10 \\ 5 \\ 10 \end{bmatrix}.$$

We can interpret this problem as follows: We have exactly 100 of each of our 5 resources c_i^{max}, $i = 1, \ldots, 5$ available. For each activity x_i we consume some amount of resource c_j given by $A_{ij}x_i$. So the first activity operating at level x_1 consumes x_1 of resource c_1, $2x_1$ of resource c_2, etc.. We want to find the activity levels that maximize the revenue while using only as much resource as we are allotted. This problem can be solved easily using a wide range of software. We will use the `cvxpy` package in python.

```
import cvxpy as cvx
import numpy as np

def optimalActivity(A, cmax, p, pdisc, q):
```

```

(m,n) = A.shape
x = cvx.Variable(n)
rev = cvx.sum( cvx.minimum( p*x, p*q + pdisc*(x - q) ) )
objective = cvx.Maximize( rev )
constraints = [x >= np.zeros(n), cvx.matmul(A,x) <= cmax]

prob = cvx.Problem(objective, constraints)
prob.solve()

return (x.value)
```

Listing 1: `cvxpy` Optimal Activity LP

Inputting the variables defined above gives for

$$\begin{aligned} x &= [7.69,\ 23.08,\ 30.77,\ 0] \\ r &= [19.38,\ 33.01,\ 138.08,\ 0] \\ \text{Total Revenue } &= 190.54. \end{aligned}$$

So our highest activity level is activity 3, which has the highest price point associated and also the highest discounted price point. In contrast, activity 4 has a discounted price much lower than the base price as well as consuming a lot of resource, and thus contributes negligibly to the revenue.

Remark. In the preceding problem it we may wish to confine the optimal solution to be integer valued, so that the hours assigned to each employee are integers. The field of study in which such constraints are imposed is that of Integer Programming [**integerProgramming**], which is not discussed in this work.

4.3 Limitations

In the the preceding example (35), we shifted to a solver in `cvxpy` in order to solve our convex optimization problem. Solving linear programs is well understood, and there are many techniques available, but at this point in time we still do not have the tools to be able to understand these solvers nor to be able to write our own. The next chapter explores the dual problem of an optimization problem, which is precisely what will be necessary to understand a variety of different methods that can be used to solve these problems numerically.

5 Duality

This chapter focuses on the dual problem of an optimization problem. In mathematics, we associate duality with induced structure of a mathematical object. The hope is that given some object X, there is some related object, X^*, which we call the dual of X, that can be derived from X in such a way that X^* inherits the structure of X and often has more structure that X does not have. This dual object is typically useful in studying the structure of the original object, X. As a concrete example, for any vector space V, there is an associated dual space V^*, that is, the set of all linear functionals on V. It may be shown that V^* is itself a vector space and, in the case of a finite dimensional vector space, a fundamental result in linear algebra shows that

$$V \simeq V^*,$$

so that each vector v in V can be associated to a linear functional $v^* \in V^*$ in a one-to-one correspondence. Thus we may express every linear functional $f : V \to \mathbb{F}$ as the map

$$\langle \cdot, v \rangle : V \to \mathbb{F}$$

for some $v \in V$.

In infinite dimensions a similar construction can be made on a normed vector space V, with the exception that V is no longer required to be isometrically isomorphic to V^* (they need not be the the 'same' space). It can be shown that the dual space of V^* is always complete and hence a Banach space, whereas V need not be.

In mathematical optimization a similar theme follows. We will define and explore the dual problem of an optimization problem and discover that the dual problem is always a convex optimization problem, even if the original is not. This property allows us to use what we understand about convex optimization in order to uncover information about the original, or primal, problem. To continue the vector space analogy, if the primal problem is itself convex, then its relationship to its dual will be even more intimate.

5.1 The Lagrangian and the Lagrange Dual Function

We begin our exploration of duality by considering the Lagrangian of an optimization problem. Recall the standard optimization problem (19)

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, i = 0, \ldots, m, \\ & h_i(x) = 0, i = 0, \ldots, p. \end{aligned}$$

with $x \in \mathbb{R}^n$. We will assume the domain, $\mathcal{D}$, is non-empty and let p^* be the optimal value of the problem. Note that this is not necessarily convex.

Definition 37 (*Lagrangian*)**.** The *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the optimization problem above is the function

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

with $\mathbf{dom}\, L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. We refer to λ_i as the *Lagrange multiplier* associated with the i^{th} inequality constraint, and ν_i as the *Lagrange multiplier* associated with the i^{th} equality constraint. The vectors λ, ν are called the *dual variables* or *Lagrange multiplier vectors* associated with the optimization problem.

Definition 38 (*Lagrange Dual Function*)**.** The *Lagrange Dual Function*, or *dual function* associated with (19) is the function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ defined to be the infimum of the Lagrangian over x for all $\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p$. i.e.:

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

If the Lagrangian is unbounded below in x then the dual function takes on the value $-\infty$. We notice that the dual function g is the pointwise infimum of a family of affine functions of (λ, ν), and thus g is concave (see 25). Remember that (19) is not convex in general, so we see that the dual function of any optimization problem is concave.

5.2 The Dual Function and Lower Bounds

Recall that p^* denotes the optimal value of (19). A very interesting aspect of the dual function is that for any $\lambda \succcurlyeq 0$ and any $v \in \mathbb{R}^p$ it follows that

$$g(\lambda, \nu) \leq p^*. \tag{24}$$

Proof. Let $\tilde{x}$ be a feasible point for (19) and let $\lambda \succcurlyeq 0$. Then $f_i(\tilde{x}) \leq 0$ for $i = 1, \ldots, m$ and $h_i(\tilde{x}) = 0$ for $i = 1, \ldots, p$. We see immediately,

$$\sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \leq 0,$$

since $\sum_{i=1}^{p} \nu_i h_i(x) = 0$ and $\sum_{i=1}^{m} \lambda_i f_i(x) 0$ as $\lambda \succcurlyeq 0$ and each $f_i(\tilde{x}) \leq 0$. Thus

$$L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x}).$$

By definition of the dual function we see

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \mu) \leq L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x}).$$

We have just shown that $g(\lambda, \nu) \leq f(\tilde{x})$ holds for any feasible $\tilde{x}$, so, in particular, (24) holds. $\square$

Example 36. Consider the problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & x^T x \\ \text{subject to} \quad & Ax = b, \end{aligned}$$

for some $A \in \mathbb{R}^{p \times n}$. The Lagrangian of this problem is

$$L(x, \nu) = x^T x + v^T (Ax - b)$$

with domain $\mathbb{R}^n \times \mathbb{R}^p$. From this, noticing that $L(x, \nu)$ is a convex quadratic, we can find the dual function $g(\nu) = \inf_x \{x^T x + \nu^T (Ax - b)\}$ by noting

$$\nabla_x L(x, \nu) = 2x + A^T \nu,$$

thus the minimum of $L(x, \nu)$ occurs at $x = (-1/2) A^T \nu$ and

$$g(\nu) = L((-1/2) A^T \nu, \nu) = (-1/4) \nu^T A A^T \nu - b^T \nu.$$

Since we have no inequality constraints, the above result shows $g(\nu) \leq p^*$ for any $\nu \in \mathbb{R}^p$.

5.3 A Connection between the Dual Function and Conjugate Functions

We recall the conjugate function f^* of a function $f : \mathbb{R}^n \to R$ is defined as

$$f^*(y) = \sup_{x \in \mathbf{dom}\, f} (y^T x - f(x)).$$

To see how this function may be related to the dual function, consider the optimization problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f(x) \\ \text{subject to} & x = 0. \end{array}$$

We note the optimal value of this problem is $f(0)$ for the only optimal point $x = 0$. But that is not what we are interested in. Consider the Lagrangian $L(x, \nu) = f(x) + \nu^T x$, and the dual function

$$g(\nu) = \inf_x (f(x) + \nu^T x = -\sup_x(-\nu^T x - f(x)) = -f^*(\nu).$$

More generally, for any optimization problem with linear constraint functions,

$$\begin{array}{ll} \underset{x}{\text{minimize}} & f_0(x) \\ \text{subject to} & Ax \preccurlyeq b., \\ & Cx = d. \end{array}$$

The dual function for this problem is

$$\begin{aligned} g(\lambda, \nu) &= \inf_x (f(x) + \lambda^T(Ax - b) + \nu^T(Cx - d) \\ &= -b^T\lambda - d^T\nu + \inf_x (f_0(x) + (A^T\lambda + C^T\nu)^T x) \\ &= -b^T\lambda - d^T\nu - f_0^*(-A^T\lambda - C^T\nu). \end{aligned}$$

We see we can now write the domain of the dual function in terms of the conjugate function as follows:

$$\mathbf{dom}\, g = \{(\lambda, \nu) \mid -A^T\lambda - C^T\nu \in \mathbf{dom}\, f^*\}.$$

Example 37. Consider the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\| \\ \text{subject to} & Ax = b, \end{array}$$

for any norm $\|\cdot\|$. Recall that

$$f^*(x) = \begin{cases} 0 & \|y\|_* \leq 1 \\ \infty & \text{otherwise,} \end{cases}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ and is defined by $\|y\|_* = \sup_x (y^T x \mid \|x\| \leq 1)$. Thus f_0^* is the indicator function on the dual norm unit ball. Then we see, from the previous result,

$$g(\nu) = -b^T\nu - f_0^*(-A^T\nu) = \begin{cases} -b^T\nu & \|A^T\nu\|_* \leq 1 \\ -\infty & \text{otherwise.} \end{cases}$$

5.4 The Lagrange Dual Problem

We have seen that, given a general optimization problem, the Lagrange dual function produces a lower bound on the optimal value p^* for any pair (λ, ν) so long as $\lambda \succcurlyeq 0$. From this result we would very much like to know the best lower bound that we can achieve, which leads us naturally to form the *Lagrange Dual Problem*, or *dual problem*, associated with (19),

$$\begin{aligned} &\text{maximize} \quad g(\lambda, \nu) \\ &\text{subject to} \quad \lambda \succcurlyeq 0. \end{aligned} \tag{25}$$

We often call (19) the *primal problem*, and we use the term *dual feasible* to refer to a point (λ, ν) with $\lambda \succcurlyeq 0$ and $g(\lambda, \nu) > -\infty$, as a point of this description is a feasible point of the dual problem. The optimal solution to the dual problem will be denoted (λ^*, ν^*) and referred to as *dual optimal* or as the *optimal Lagrange Multipliers*.

Remark. The Lagrange Dual Problem is a convex optimization problem because the objective function is concave and the constraint function is convex.

Though there is only one constraint function associated with the dual problem, we may often make implicit constraints explicit. This is easiest to see when the dimension of $\mathbf{dom}\, g$ is less than $m + p$ (for $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$). This may allow us to form equivalent problems that are more familiar to us.

5.4.1 Making Dual Constraints Explicit

Example 38. Given the primal problem,

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad c^T x \\ &\text{subject to} \quad Ax = b, \\ &\qquad\qquad\quad\;\; x \succcurlyeq 0, \end{aligned}$$

with dual function

$$g(\lambda, \nu) = \begin{cases} -b^T \nu & A^T \nu - \lambda + c = 0 \\ -\infty & \text{otherwise,} \end{cases}$$

we form the dual problem (strictly from definition) as follows:

$$\begin{aligned} &\text{maximize} \quad g(\lambda, \nu) = \begin{cases} -b^T \nu & A^T \nu - \lambda + c = 0 \\ -\infty & \text{otherwise,} \end{cases} \\ &\text{subject to} \quad \lambda \succcurlyeq 0. \end{aligned}$$

But we notice that g is only finite when $A^T \nu - \lambda + c = 0$. So we form an equivalent problem

$$\begin{aligned} &\underset{\nu}{\text{maximize}} \quad -b^T \nu \\ &\text{subject to} \quad A^T \nu - \lambda + c = 0, \\ &\qquad\qquad\qquad\qquad\quad\; \lambda \succcurlyeq 0, \end{aligned}$$

which is again equivalent to

$$\begin{aligned} &\underset{\nu}{\text{maximize}} \quad -b^T \nu \\ &\text{subject to} \quad A^T \nu + c \succcurlyeq 0. \end{aligned}$$

This is a LP in standard inequality form. We often, abusively, refer to the previous two problems as the Lagrange Dual Problem of the standard LP form we began with.

5.5 Weak Duality

The optimal value of the Lagrange Dual Problem, d^*, is, by definition, the best lower bound on p^* that can be obtained by the Lagrange Dual Function. In particular, (24) tells us

$$d^* \leq p^*. \tag{26}$$

This property is called *weak duality* and it holds in every case. Indeed, we see that if $p^* = -\infty$ we must have $d^* = -\infty$ which implies the Lagrange Dual Problem is infeasible. Similarly, if we have that $d^* = \infty$ then it must be that $p^* = \infty$ and thus the primal problem is infeasible.

The difference $p^* - d^*$ is called the *duality gap* of the primal problem since it gives the distance between optimal values. It is clear that the duality gap is always non-negative. We make heavy use of the preceding facts to find lower bounds on difficult optimization problems.

Example 39. The Dual of the Unconstrained Least Squares Problem

Consider the unconstrained least squares problem given below:

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|_2^2 = \sum_{i=1}^{k} (a_i^T x - b_i)^2$$

for $A \in \mathbb{R}^{k \times n}$, $k \geq n$, $b \in \mathbb{R}^k$, $x \in \mathbb{R}^n$.

Introducing an auxiliary variable $z \in \mathbb{R}^n$, we can form an equivalent minimization problem

$$\begin{aligned} \underset{x,\, z}{\text{minimize}} \quad & \|z - b\|_2^2 = \sum_{i=1}^{k} (z_i - b_i)^2 \\ \text{subject to} \quad & z = Ax, \end{aligned}$$

in which case we have the Lagrangian

$$L(z, x, \nu) = \|z - b\|_2^2 + \nu^T (z - Ax).$$

To find the Lagrange Dual function we minimize with respect to z and x by solving the system

$$\begin{aligned} \nabla_z L &= 2(z - b) + \nu = 0 \\ \nabla_x L &= -\nu^T A = 0 \end{aligned}$$

$$\begin{aligned} 2(z - b) + \nu = 0 &\implies z = b - \frac{1}{2}\nu \\ -\nu^T A = 0 &\implies \nu \in \mathcal{N}(A^T), \end{aligned}$$

where $\mathcal{N}(A^T)$ denotes the null space of A^T. Thus we have

$$\begin{aligned} g(\nu) &= L(b - \frac{1}{2}\upsilon, x, \nu) \\ &= \frac{1}{4}\|\nu\|_2^2 + \nu^T (b - \frac{1}{2}\nu - Ax) \\ &= \frac{1}{4}\|\nu\|_2^2 + \nu^T (b - \frac{1}{2}\nu) \\ &= \nu^T (b - \frac{1}{4}\nu) \end{aligned}$$

so the Lagrange Dual problem is given by

$$\underset{\nu \in \mathcal{N}(A^T)}{\text{maximize}} \quad \nu^T(b - \frac{1}{4}\nu).$$

By weak duality we now have for any feasible ν, x

$$\nu^T(b - \frac{1}{4}\nu) \leq \|Ax - b\|_2^2.$$

But check it out, we also notice from above that

$$\begin{aligned} g(\nu) &= \frac{1}{4}\|\nu\|_2^2 + \nu^T(b - \frac{1}{2}\nu) \\ &= \frac{1}{4}\|\nu\|_2^2 - \frac{1}{2}\|\nu - b\|_2^2 + \frac{1}{2}\|b\|_2^2, \end{aligned}$$

and so we can again form the Lagrange Dual problem below,

$$\underset{\nu \in \mathcal{N}(A^T)}{\text{maximize}} \quad \frac{1}{4}\|\nu\|_2^2 - \frac{1}{2}\|\nu - b\|_2^2 + \frac{1}{2}\|b\|_2^2.$$

Since the first and third terms are always positive, we can form another equivalent minimization problem:

$$\begin{aligned} &\underset{x,\,z}{\text{minimize}} && \|\nu - b\|_2^2 \\ &\text{subject to} && A^T\nu = 0. \end{aligned}$$

This above problem can be interpreted with respect to the original least squares problem by finding the vector ν in $\mathcal{R}(\mathcal{A})$ that is closest to b, which is exactly what we sought to do with the primal problem, and thus, when the primal problem is feasible the dual and the primal problem will coincide. Thus our duality gap is 0 and so we conclude the primal and dual problem are equivalent, or that

$$d^* = p^*.$$

5.6 Strong Duality and Slater's Constraint Qualification

If the equality holds in (26) then we say *strong duality* holds for the primal problem. It is clear already from previous examples that strong duality does not always hold. However, if the primal problem is convex we usually have strong duality. This is not at all a rigorous statement, and thus we are very much interested in the conditions that must be satisfied for strong duality to hold. The conditions that ensure strong duality of a convex optimization problem are called *constraint qualifications*. One simple constraint qualification is *Slater's Condition* which states that there exists some $x \in \textbf{relint}\, \mathcal{D}$ such that

$$f_i(x) < 0, i = 1, 2, \ldots, m, \; Ax = b.$$

This is sometimes referred to as a *strictly feasible* point. If the first k inequality constraints are affine, then we can refine the conditions to state that

$$f_i(x) \leq 0, i = 1, 2, \ldots, k, \; f_i(x) < 0, i = k+1, k+2, \ldots, m, \; Ax = b.$$

Remark. Per the above result, Slater's Condition reduces to feasibility when all constraint functions are affine. For example, strong duality holds for (38), therefore one could solve either the primal or dual problems to find the optimal value.

Example 40. Recall the problem

$$\begin{array}{ll} \underset{x}{\text{minimize}} & x^T x \\ \text{subject to} & Ax = b. \end{array}$$

The dual problem is

$$\underset{\nu}{\text{maximize}} \quad (-1/4)\nu^T AA^T \nu - b^T \nu.$$

In this case Slater's Condition is simply that the primal problem is feasible, thus $p^* = d^*$ as long as $b \in \mathcal{R}(A)$ or $p^* < \infty$. In fact, in this particular case, if $p^* = \infty$ then the dual problem is infeasible as well and thus $d^* = \infty$. So strong duality holds completely in this instance.

5.6.1 A Geometric View of Duality and Proof of Slater's Condition

We will now work towards a proof that Slater's condition indeed ensures strong duality. To do this we'll first we will discuss a geometric interpretation of weak and strong duality. Working with the standard convex problem (20) we construct the following set:

$$\mathcal{G} = \{(f_1(x), f_2(x), \ldots, f_m(x), h_1(x), h_2(x), \ldots, h_p(x), f_0(x)) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid x \in \mathcal{D}\}.$$

This is simply the set of values taken on by the objective and constraint functions. We can now express p^* in terms of this set in the following way:

$$p^* = \inf\{t \mid (u, v, t) \in G, u \preccurlyeq 0, v = 0\}.$$

We can also immediately express the dual function as

$$\begin{aligned} g(\lambda, \nu) &= \inf\{(\lambda, \nu, 1)^T (u, v, t) \mid (u, v, t) \in \mathcal{G}\} \\ &= \inf \left\{ \sum_{i=1}^m \lambda_i u_i + \sum_{i=1}^p \nu_i v_i + t \mid (u, v, t) \in \mathcal{G} \right\}. \end{aligned}$$

Now, if $g(\lambda, \nu)$ is finite then we define a supporting hyperplane to $\mathcal{G}$ by

$$(\lambda, \nu, 1)^T (u, v, t) \geq g(\lambda, \nu).$$

If we impose $\lambda \succcurlyeq 0$ then we see that, if $u \preccurlyeq 0$ and $v = 0$, $t \geq (\lambda, \nu, 1)^T (u, v, t)$. We now establish weak duality as follows:

$$\begin{aligned} p^* &= \inf\{t \mid (u, v, t) \in G, u \preccurlyeq 0, v = 0\} \\ &\geq \inf\{(\lambda, \nu, 1)^T (u, v, t) \mid (u, v, t) \in G, u \preccurlyeq 0, v = 0\} \\ &\geq \inf\{(\lambda, \nu, 1)^T (u, v, t) \mid (u, v, t) \in G\} \\ &= g(\lambda, \nu). \end{aligned}$$

Lastly, before we prove Slater's Condition we will establish a sort of epigraph of $\mathcal{G}$. Define the set $\mathcal{A} \subseteq \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}$ by

$$\mathcal{A} = \mathcal{G} + (\mathbb{R}_+^m \times \{0\} \times \mathbb{R}_+).$$

Geometrically, this can be thought of as a sort of epigraph of $\mathcal{G}$ because it contains everything in $\mathcal{G}$ as well as those which are 'worse', or have a larger objective values or inequality constraint function values. In terms of $\mathcal{A}$ we see

$$p^* = \inf\{t \mid (0,0,t) \in \mathcal{A}\}.$$

Now, similarly to $\mathcal{G}$, we can evaluate the dual function at (λ, ν) with $\lambda \succcurlyeq 0$ by minimizing $(u,v,t)^T(\lambda,\nu,1)$ over $\mathcal{A}$. For $\lambda \succcurlyeq 0$ we have

$$g(\lambda,\nu) = \inf\{(\lambda,\nu,1)^T(u,v,t) \mid (u,v,t) \in \mathcal{A}\},$$

and again, if $g(\lambda,\nu)$ is finite we define a supporting hyperplane to $\mathcal{A}$ by

$$(\lambda,\nu,1)^T(u,v,t) \geq g(\lambda,\nu).$$

In particular, since $(0,0,p^*) \in \textbf{bd}\,\mathcal{A}$ we see that

$$p^* = (\lambda,\nu,1)^T(0,0,p^*) \geq g(\lambda,\nu),$$

again showing weak duality. In this case we have strong duality if and only if the strict equality is satisfied in the inequality above, i.e. if there is a supporting hyperplane to $\mathcal{A}$ at $(0,0,p^*)$. With the above results in mind, we now prove that Slater's Condition provides constraint qualificiations.

Proof. Let $f_0, f_1, f_2, \ldots, f_m$ be convex functions and assume Slater's condition holds. i.e. assume there exists some $\tilde{x} \in \textbf{relint}\,\mathcal{D}$ so that $f_i(\tilde{x}) < 0$ for $i = 1,2,\ldots,m$ and $A\tilde{x} = b$. We also make two other simplifying assumptions. The first is that $\mathcal{D}$ is non-empty and thus $\textbf{relint}\,\mathcal{D} = \textbf{int}\,\mathcal{D}$, and the second is that rank $A = p$. We also presume p^* is finite since $p^* \neq \infty$ as the problem is feasible, and $p^* = -\infty$ implies $d^* = -\infty$ by weak duality. Now, since the problem itself is convex we see that $\mathcal{A}$ is a convex set. We also define another convex set $\mathcal{B}$ by

$$\mathcal{B} = \{(0,0,s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid s < p^*\}.$$

Lemma 4. $\mathcal{A} \cap \mathcal{B} = \{\}$

Proof. To prove the above claim, suppose $(u,v,t) \in \mathcal{A} \cap \mathcal{B}$. We first note that $(u,v,t) \in \mathcal{B}$ implies $u = 0, v = 0$ and $t < p^*$. Also, since $(u,v,t) \in \mathcal{A}$ it follows that there exists some $x \in \mathcal{D}$ so that $f_i(x) \leq 0, i = 1,\ldots,m, Ax = b$, and $f_0(x) \leq t < p^*$. But this contradicts p^* optimal, and thus the claim is proved. □

Since we have two disjoint convex sets, using the separating hyperplane theorem we have some non-zero $(\tilde{\lambda}, \tilde{\nu}, \mu)$ and $\alpha \in \mathbb{R}$ so that

$$(u,v,t) \in \mathcal{A} \implies \tilde{\lambda}^T u + \tilde{\nu}^T v + \mu t \geq \alpha \tag{27}$$

$$(u,v,t) \in \mathcal{B} \implies \tilde{\lambda}^T u + \tilde{\nu}^T v + \mu t \leq \alpha. \tag{28}$$

Now, in the case that $(u,v,t) \in \mathcal{A}$ it must be that $\tilde{\lambda} \succcurlyeq 0$ and $\mu \succcurlyeq 0$, otherwise $\tilde{\lambda} + \mu t$ is unbounded below. The second condition above says simply that $\mu t \leq \alpha$ for $t < p^*$ which implies $\mu p^* \leq \alpha$. Thus, for any $x \in \mathcal{D}$, we have that,

$$\sum_{i=1}^{m} \tilde{\lambda}^T f_i(x) + \tilde{\nu}^T(Ax - b) + \mu f_0(x) \geq \alpha \geq \mu p^*. \tag{29}$$

Assuming now that $\mu > 0$ we have that $L(x, \tilde{\lambda}/\mu, \tilde{\nu}/\mu) \geq p^*$ for all $x \in \mathcal{D}$, and thus, minimizing over x,

$$g(\tilde{\lambda}, \tilde{\nu}) \geq p^*.$$

Along with weak duality we now have $g(\tilde{\lambda}, \tilde{\nu}) = p^*$. So, for $\mu > 0$, strong duality holds and the dual optimal is obtained at $(\lambda^*, \nu^*) = (\tilde{\lambda}, \tilde{\nu})$.

We lastly consider the case where $\mu = 0$. In this case we have, by (29), that for all $x \in \mathcal{D}$

$$\sum_{i=1}^{m} \tilde{\lambda}^T f_i(x) + \tilde{\nu}^T (Ax - b) \geq 0.$$

Now take a point $\tilde{x} \in \mathcal{D}$ satisfying Slater's Condition. So we have,

$$\sum_{i=1}^{m} \tilde{\lambda}^T f_i(\tilde{x}) \geq 0.$$

Notice that $f_i(\tilde{x}) < 0$ for $i = 1, 2, \ldots, m$, and $\tilde{\lambda} \succcurlyeq 0$, which implies $\tilde{\lambda} = 0$. But we are working with the assumption that $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$ thus $\tilde{\nu} \neq 0$. Further, for all $x \in \mathcal{D}$ we have,

$$\tilde{\nu}^T (Ax - b) \geq 0,$$

and for $\tilde{x}$ we know $\tilde{\nu}^T (A\tilde{x} - b) = 0$. Since $\tilde{x} \in \mathbf{int}\, \mathcal{D}$, there exists some $x \in \mathcal{D}$ satisfying $\tilde{\nu}^T (Ax - b) < 0$ unless $A^T \tilde{\nu} = 0$. But this can only happen if $\tilde{\nu} = 0$ which gives a contradiction since we assumed $\operatorname{rank} A = p$. $\square$

5.7 Optimality Conditions

Recall that our results about weak duality hold for any optimization problem, not just those that are convex. For this reason we are very interested in the dual problem as a tool for finding lower bounds on p^* that will act as *certificates* of suboptimality. In other words, these are *proof* that $p^* \geq g(\lambda, \nu)$. If strong duality holds for a problem then these certificates can be made arbitrarily close to p^*

Notice that since $p^* \geq g(\lambda, \nu)$ it follows that, for any feasible point x,

$$f_0(x) - p^* \leq f_0(x) - g(\lambda, \nu),$$

so the dual feasible point (λ, ν) ensures that x is ε-suboptimal where $\varepsilon = f_0(x) - g(\lambda, \nu)$. For any primal feasible x and dual feasible (λ, ν), we call the gap $f_0(x) - g(\lambda, \nu)$ between these points the *duality gap* associated with x and (λ, ν). This duality gap allows us to bound the interval in which p^* and d^* will be found, namely,

$$p^* \in [g(\lambda, \nu), f_0(x)], \qquad d^* \in [g(\lambda, \nu), f_0(x)].$$

Not surprisingly, these facts are incredibly useful for providing stopping criterion for many optimization algorithms.

5.8 Complementary Slackness

Let's consider the case where both the primal optimal and dual optimal values to (19) are obtained and equal ($p^* = d^*$). In other words, strong duality holds and we can find some x^* primal optimal point and $g(\lambda^*, \nu^*)$. Then the following string of inequalities holds by strong duality:

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\ &\leq f_0(x^*). \end{aligned}$$

The last inequality follows from the fact that $x^*, (\lambda^*, \nu^*)$ are primal feasible and dual feasible respectively. Thus $\lambda_i^* \succcurlyeq 0, f_i(x^*) \leq 0, h_i(x^*) = 0$. Therefore equality must hold for all the statements above. This leads to two interesting observations. The first is that x^* minimizes $L(x, \lambda^*, \nu^*)$ (that is not to say that x^* is the only minimizer). The second, again following from $\lambda_i^* \succcurlyeq 0, f_i(x^*) \leq 0, h_i(x^*) = 0$, is that

$$\lambda_i^* f_i(x^*) = 0,\ i = 1, \ldots, m, \tag{30}$$

which will be referred to as *complementary slackness*. Complementary slackness, as we have just seen, holds for any primal optimal x^* and dual optimal (λ^*, ν^*) when strong duality holds. This condition can also be expressed as

$$\lambda_i^* > 0 \implies f_i(x^*) = 0,$$

or, equivalently, as

$$f_i(x^*) > 0 \implies \lambda_i^* = 0.$$

5.9 KKT Optimality Conditions

We will now discuss another set of optimality conditions, and we will assume that the functions $f_0, f_1, \ldots, f_m$, $h_1, \ldots, h_p$ are all differentiable and therefore have open domains. We will not yet assume the problem is convex.

5.9.1 KKT Conditions for Non-Convex Problems

Suppose, as before, that x^* and (λ^*, ν^*) are primal and dual optimal respectively, and that $p^* = d^*$. Since x^* minimizes $L(x, \lambda^*, \nu^*)$ over x it follows that its gradient vanishes at x^*, i.e.

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0.$$

Then all in all we have the following conditions:

$$\begin{aligned} f_i(x^*) &\leq 0,\ i = 1, \ldots, m \\ h_i(x^*) &= 0,\ i = 1, \ldots, p \\ \lambda_i^* &\geq 0,\ i = 1, \ldots, m \\ \lambda_i^* f_i(x^*) &= 0,\ i = 1, \ldots, m \\ \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) &= 0. \end{aligned} \tag{31}$$

These are the *Karush-Kuhn-Tucker*, or *KKT* conditions.

5.9.2 KKT Conditions for Convex Problems

If the given optimization problem is convex then we have the result that the KKT conditions are sufficient for the points $x^*, (\lambda^*, \nu^*)$ to be primal and dual optimal respectively. i.e. for any points $\tilde{x}, (\tilde{\lambda}, \tilde{\nu})$ which satisfy

$$
\begin{aligned}
f_i(\tilde{x}) \leq 0,\ i = 1, \ldots, m \\
h_i(\tilde{x}) = 0,\ i = 1, \ldots, p \\
\tilde{\lambda}_i \geq 0,\ i = 1, \ldots, m \\
\tilde{\lambda}_i f_i(\tilde{x}) = 0,\ i = 1, \ldots, m
\end{aligned}
$$

$$
\nabla f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{x}) = 0,
$$

it follows that $\tilde{x}, (\tilde{\lambda}, \tilde{\nu})$ are primal and dual optimal.

Proof. Since $f_i(\tilde{x}) \leq 0$, $i = 1, \ldots, m$, $h_i(\tilde{x}) = 0$, $i = 1, \ldots, p$, $\tilde{x}$ is primal feasible. Then as $\tilde{\lambda}_i \geq 0$ for $i = 1, \ldots, m$, it follows that $L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ is convex in x, and the last condition shows that $\tilde{x}$ minimizes L over x. From that we can conclude that

$$
\begin{aligned}
g(\tilde{\lambda}, \tilde{\nu}) &= L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \\
&= f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{\nu}_i h_i(\tilde{x}) \\
&= f(\tilde{x}),
\end{aligned}
$$

where the last step in the equality comes from the second and fourth KKT conditions. Thus the duality gap for $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$ is zero, which implies $\tilde{x}, (\tilde{\lambda}, \tilde{\nu})$ are primal and dual optimal. $\square$

Remark. If a convex problem with differentiable objective and constraint functions satisfies Slater's condition, the KKT conditions provide necessary and sufficient conditions for optimality. This is because Slater's condition implies that the optimal duality gap is zero and thus there is a dual optimal point. This means that given some primal optimal x^*, we have a dual feasible and dual optimal pair (λ^*, ν^*), which together with x^* satisfy the KKT conditions.

6 Interior Point Methods

We've come a long way, and at this point I think it is fitting to preface our next step forward with a joke that I heard second-hand from some peers.

A Joke: *An engineer, a physicist, and a mathematician all find themselves in separate hotels on one particularly dry evening. By chance, a poorly placed candle and a gust of wind start a fire in the engineer's hotel room. Fast on his feet, he finds the fire extinguisher and puts out the fire quickly. In the physicist's hotel room a similar event transpires, and a fire breaks out. She also quickly locates the fire extinguisher, but spends a few moments considering the most economical way to douse the flame. After a time, she elegantly and efficiently puts the fire out. In the mathematician's hotel room there is also a fire. He spots it, and then quickly locates the fire extinguisher. Relieved, he lets out a deep sigh. 'Good,' he says, putting on his hat and coat and heading out to dinner, 'I found a solution.'*

In a sense, we are now leaving the realm of the mathematician described above. At this point in time we have developed an underlying understanding of convexity, and should be ready to see how to exploit the structure of a convex optimization problem to actually solve some of the various convex problems we have explored. Henceforth we will be concerned with this exploitation, although we will not cover these methods in their entirety. Instead, we will explore a specific interior point method known as the *barrier method*. First, however, we must understand some of the basics of these methods.

6.1 Descent Methods

First we will develop a modest understanding of descent methods. Lets begin by considering the unconstrained minimization problem

$$\text{minimize} \quad f(x), \tag{32}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and f is convex and twice differentiable. This implies $\mathbf{dom}\, f$ is open. We will also assume the problem is solvable and thus there exists $x^* \in \mathbf{dom}\, f$, i.e. we have a minimizer x^* of (32) so that

$$\inf_x f(x) = f(x^*) = p^*.$$

Since f is differentiable and convex it must be the case that

$$\nabla f(x^*) = 0. \tag{33}$$

Furthermore, we will show that x^* is unique, which means that solving (32) is the same as finding a solution x^* satisfying (33). In some cases it is possible to solve (33) analytically, as it is a system of n equations in n variables, namely $x_1, x_2, \ldots, x_n$. However, most of the time we solve these problems in an iterative fashion that will require of us a well behaved starting position $x^{(0)}$. In particular, we will require two things:

$$x^{(0)} \in \mathbf{dom}\, f \tag{34}$$

$$S = \{x \in \mathbf{dom}\, f \mid f(x) \le f(x^{(0)})\} \text{ is closed,} \tag{35}$$

where S is the $f(x^{(0)})$-sublevel set.

The goal of a descent method is to construct a *minimizing sequence* $x^{(0)}, \ldots, x^{(k)}, x^{(k+1)}, \cdots \in \mathbf{dom}\, f$ so that $f(x^{(k)}) < f(x^{(k+1)})$ and $f(x^{(k)}) \to p^*$ as $k \to \infty$. In practice this is terminated when $f(x^{(k)}) - p^* < \varepsilon$ for some specified tolerance ε.

A general descent method can be described in the following way:

Algorithm 1: General Descent Method

given $\varepsilon > 0$, an $x \in \mathbf{dom}\, f$;
while $f(x) - p^* \geq \varepsilon$ **do**
  Determine a descent direction Δx;
  Line search. Choose step size $t > 0$;
  Update $x = x + t\Delta x$
end
return x

6.2 The Line Search

To be more explicit, we will describe what is known as a *backtracking line search.*

Algorithm 2: Backtracking Line Search

given a descent direction Δx for f for $x \in \mathbf{dom}\, f, \alpha \in (0, 0.5), \beta \in (0, 1)$;
$t := 1$;
while $f(x + \Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x, t := \beta t$.

Notice that

$$f(x + t\Delta x) \approx f(x) + t\nabla f(x)^T \Delta x < f(x) + t\alpha \nabla f(x)^T \Delta$$

for some t, because Δx is a descent direction and thus $\nabla f(x)^T < 0$. So, for a small enough t, we see

$$f(x + t\Delta x) < f(x) + t\alpha \nabla f(x)^T \Delta.$$

6.3 Newton's Method

In Newton's Method we construct the minimizing sequence in the following way. For $x \in \mathbf{dom}\, f$, consider the vector

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x).$$

Δx_{nt} is called the *Newton step* for f at x. Since f is convex, $\nabla^2 f(x)$ is positive definite and so

$$\nabla f(x) \Delta x_{\mathrm{nt}} = -\nabla f(x) \nabla^2 f(x)^{-1} \nabla f(x)^T < 0,$$

unless $\nabla f(x) = 0$. Therefore the Newton step is in the gradient descent direction unless x itself is optimal.

One interpretation of the Newton step comes from considering the second order Taylor series approximation of f around x, $\hat{f}$ which is

$$\hat{f}(x + v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v.$$

This is a convex quadratic in v and is minimized when $v = \Delta x_{\mathrm{nt}}$, so in this sense the newton step is exactly what must be added to x to minimize the second order Taylor series approximation of f.

Definition 39 (*Newton decrement*)**.** The *Newton decrement* at $x \in \mathbf{dom}\, f$ is the quantity $\lambda(x)$ given by:

$$\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{1/2}.$$

The Newton decrement is an important quantity in Newton's Method particularly as a stopping criterion. We can relate $\lambda(x)$ to $f(x) - \inf_y \hat{f}(y)$ in the following way:

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x + \Delta x_{\mathrm{nt}}) = \frac{1}{2} \lambda(x)^2,$$

where $\hat{f}$ is the second order Taylor approximation of f at x. Therefore $\lambda^2/2$ provides an estimate of $f(x) - p^*$ based on a quadratic approximation of f.

The Newton decrement can also be expressed as

$$\lambda(x) = (\Delta x_{\mathrm{nt}}^T \nabla^2 f(x) \Delta x_{\mathrm{nt}})^{1/2} = \|\Delta x_{\mathrm{nt}}\|_{\nabla^2 f(x)}.$$

We also notice that

$$\nabla f(x)^T \Delta x_{\mathrm{nt}} = -\lambda(x)^2.$$

This is the constant used in a backtracking line search, and can be seen as the directional derivative of f at x in the direction of the Newton step, as shown below,

$$-\lambda(x)^2 = \nabla f(x)^T \Delta x_{\mathrm{nt}} = \frac{d}{dt} f(x + t\Delta x_{\mathrm{nt}})|_{t=0}.$$

We now give the algorithm for Newton's Method.

Algorithm 3: Newton's Method

Result: $x \in \mathbf{dom}\, f$ & ε-suboptimal
given $x \in \mathbf{dom}\, f$, tolerance $\varepsilon > 0, \lambda = \lambda(x)$;
repeat:

1. Compute Newton step and decrement.

$$\begin{aligned} \Delta x_{\mathrm{nt}} &= -\nabla^2 f(x)^{-1} \nabla f(x), \\ \lambda^2 &= \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \end{aligned}$$

2. Stopping Criterion. **quit** if $\lambda^2/2 \leq \varepsilon$.

3. Line search. Choose size step t by backtracking line search.

4. Update. $x = x + \Delta x_{\mathrm{nt}}$.

6.4 An Equality Constrained Convex Quadratic Problem

Recall from Section (5.9) we have the optimality condition that x^* is optimal for (39) if and only if there exists an $\nu^* \in \mathbb{R}^p$ such that

$$Ax^* = b,\ \nabla f(x^*) + A^T \nu^* = 0. \tag{36}$$

The equations given by $Ax^* = b$ are called the *primal feasibility equations*, and those given by $\nabla f(x^*) + A^T \nu^* = 0$ are the *dual feasibility equations*. The dual feasibility equations are not linear in general. Using this system an equality constrained minimization problem can be reformulated as an unconstrained minimization problem by eliminating the equality constraints. With this approach any equality constrained convex optimization problem can be recast as a unconstrained problem, in which case we may use our existing techniques to solve. However, this may not always be the best approach, and so we will also develop techniques to solve equality constrained problems directly. Consider the minimization problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) = (1/2)x^T P x + q^T x + r \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{37}$$

where $P \in S_n^+$ and $A \in \mathbb{R}^{p \times n}$. Here the optimality conditions are

$$Ax^* = b,\ Px^* + q + A^T \nu^* = 0,$$

which may be written in a matrix form as follows:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}. \tag{38}$$

This system is called a *KKT System* for (37). It is a set of $n+p$ linear equations in x^* and ν^* and it may be solved analytically. More generally, the coefficient matrix above is referred to as the *KKT matrix.* When this matrix is non-singular there exists a primal-dual optimal pair (x^*, ν^*). When the matrix is singular but the system remains solvable, then any solution yields and primal-dual optimal pair. If the system is not solvable then the problem is either unfeasible or is unbounded below.

6.5 Newton's Method with Equality Constraints

We now focus on solving the equality constrained convex minimization problem given by

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad Ax = b. \end{aligned} \tag{39}$$

Now we are concerned with a starting point $x \in \mathbf{dom}\, f$ so that $Ax = b$, and we must redefine our Newton step so that it also takes into account the equality constraint. Specifically, we wish to construct Δx_{nt} to be in a feasible direction, so that $A\Delta x_{\text{nt}} = 0$.

6.5.1 The Equality Constrained Newton Step

For some $x \in \mathbf{dom}\, f$, $Ax = b$, we begin our construction by replacing f in (39) with $\hat{f}$, its second order Taylor approximation near x. So we have the new minimization problem

$$\begin{aligned} &\underset{v}{\text{minimize}} \quad \hat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 f(x) v \\ &\text{subject to} \quad A(x+v) = b. \end{aligned} \tag{40}$$

The above is the convex quadratic optimization problem which we just solved in (6.4). We define the Newton Step Δx_{nt} by

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{\text{nt}} \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}. \tag{41}$$

Here w is the optimal dual variable for the quadratic problem. This description of the Newton Step Δx_{nt} and dual variable w can be considered the solution to the linear approximation of the system

$$Ax^* = b,\ \nabla f(x^*) + A^T \nu^* = 0.$$

If x^* is replaced by $x + \Delta x_{\text{nt}}$, ν^* is replaced by w, and $\nabla f(x)$ is replaced by it's linear approximation near x, $\nabla f(x) + \nabla^2 f(x)\Delta x_{\text{nt}}$, we have the system

$$A(x + \Delta x_{\text{nt}}) = b,\ \nabla f(x) + \nabla^2 f(x)\Delta x_{\text{nt}} + A^T w = 0.$$

Since $Ax = b$ this can be rearranged as

$$A\Delta x_{\text{nt}} = 0,\ \nabla^2 f(x)\Delta x_{\text{nt}} + A^T w = -\nabla f(x),$$

which is precisely the form of the KKT system defining the Newton step. We define and interpret the Newton Decrement in the same way as we did in the unconstrained case,

$$\lambda(x) = (\Delta x_{\text{nt}}^T \nabla^2 f(x) \Delta x_{\text{nt}})^{1/2},$$

as the norm of the Newton step with respect to the Hessian. Again we can note that

$$f(x) - \inf_v \{\hat{f}(x+v) \mid A(x+v) = b\} = \lambda(x)/2,$$

and thus from $\lambda(x)$ we can get an estimate on

$$f(x) - p^*.$$

We can also relate $\lambda(x)$ to the line search again, noticing

$$\frac{d}{dt} f(x + t\Delta x_{\text{nt}})|_{t=0} = -\nabla f(x)^T \Delta x_{\text{nt}} = \lambda(x)^2,$$

as before.

6.5.2 Newton Step at Infeasible Points

Often it is non-trivial or an inconvenience to find a feasible point to initialize the problem. Discussion of the construction of a feasible point $x \in \mathbf{dom}\, f$ is omitted from this work, though is discussed thoroughly by Boyd [**boyd**].

6.6 The Barrier Method

In this section we will consider the optimization problem

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \le 0 \,, i = 1, 2, \ldots, m, \\ & Ax = b, \end{aligned} \tag{42}$$

where $f_0, f_1, \ldots, f_m$ are convex and twice continuously differentiable functions, and $A \in \mathbb{R}^{p \times n}$ with $\operatorname{rank} A = p < n$. We will also assume that this problem is solvable and that there exists a strictly feasible point. These conditions imply that Slater's Condition holds, so we have some optimal x^* and dual optimal (λ^*, ν^*) so that

$$\begin{aligned} f_i(x^*) \le 0,\ i = 1, \ldots, m \\ Ax^* - b = 0 \\ \lambda_i^* \ge 0,\ i = 1, \ldots, m \\ \lambda_i^* f_i(x^*) = 0,\ i = 1, \ldots, m \\ \nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + A\nu^* = 0. \end{aligned} \tag{43}$$

Interior point methods, in general, solve the inequality constrained problem (42), or equivalently the KKT conditions above, by applying Newton's method to a sequence of equality constrained problems or to a sequence of modified KKT conditions.

6.7 Logarithmic Barrier Function

A first attempt at forming an equality constrained problem from (42) would be to make the inequality constraints implicit in the objective function. We can do this by rewriting (42) as

$$\begin{aligned} \text{minimize} \quad & f_0(x) + \sum_{i=1}^m I_-(f_i(x)) \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{44}$$

where $I_- : \mathbb{R} \to \mathbb{R}$ is defined as $I_-(u) = \begin{cases} \infty & u < 0 \\ 0 & \text{otherwise.} \end{cases}$ Now our problem has no inequality constraints, but our objective is no longer differentiable, so Newton's method cannot be applied to this problem.

The next step is to replace the indicator function I_- with a logarithmic function $\hat{I}$ to approximate I_-, as follows: let

$$\hat{I}_-(u,t) = -\frac{1}{t}\log(-u), \text{ for } t > 0.$$

We notice $\hat{I}(u,t)$ gives a good approximation for $I_-(u)$ that gets better as t gets large, as seen in Figure (4). Since $\log(-u)$ is concave, $-\log(-u)$ is convex. Multiplying by $\frac{1}{t}$ is positive scaling and thus preserves convexity. We will further take a composition with a convex function, and so we also note $-\frac{1}{t}\log(-u)$ is non-decreasing. So we have that that $\hat{I}_-(u,t)$ is convex, non-decreasing, and takes on the value ∞ for $x < 0$, just as I_-. This step gives us the equality constrained problem

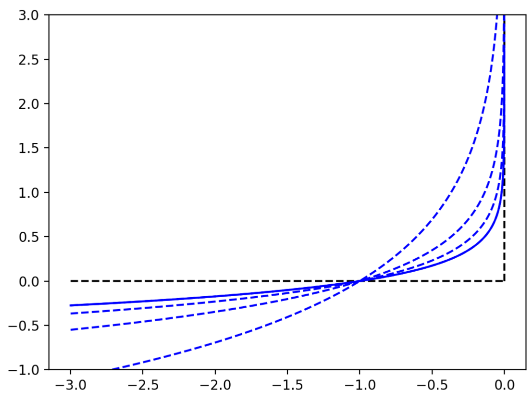


Figure 4: Plots of the $\hat{I}(u)$ for increasing values of t from $t = 1$ to $t = 4$ in blue. The best approximation shown is in solid blue, with $t = 4$. The black dashed line is the indicator function $I_-(u)$.

$$\begin{aligned} &\text{minimize} \quad f_0(x) - \frac{1}{t}\sum_{i=1}^{m}\log(-f_i(x)) \\ &\text{subject to} \quad Ax = b, \end{aligned} \tag{45}$$

which has a twice continuously differentiable objective function, and thus we can solve this problem using Newton's Method. We call the convex function φ defined by

$$\varphi(x) = -\sum_{i=1}^{m}\log(-f_i(x)),$$

the *logarithmic barrier* or *log barrier* of the problem (42). Thus we have, for the gradient and Hessian of $\varphi(x)$,

$$\nabla\varphi(x) = \sum_{i=1}^{m}\frac{\nabla f_i(x)}{-f_i(x)}, \tag{46}$$

$$\nabla^2\varphi(x) = \sum_{i=1}^{m}\frac{\nabla f_i(x)^T\nabla f_i(x)}{f_i(x)^2} + \sum_{i=1}^{m}\frac{\nabla^2 f_i(x)}{-f_i(x)}. \tag{47}$$

Example 41. Consider the linear program

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{subject to} \quad & Ax = b, \\ & x \succcurlyeq 0. \end{aligned} \tag{48}$$

Here our logarithmic barrier function is given by $\varphi(x) = -\sum_{i=1}^n \log(x_i)$ and thus

$$\nabla\varphi(x) = \left(\frac{1}{x_1}, \frac{1}{x_2}, \dots, \frac{1}{x_n}\right),$$

and

$$\nabla^2\varphi(x) = \mathbf{diag}\left(\frac{1}{x_1^2}, \frac{1}{x_2^2}, \dots, \frac{1}{x_n^2}\right),$$

and our related minimization problem is given by

$$\begin{aligned} \text{minimize} \quad & c^T x - (1/t)\sum_{i=1}^n \log(x_i) \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{49}$$

for some $t > 0$. Multiplying the objective function by t yields an equivalent problem,

$$\begin{aligned} \text{minimize} \quad & tc^T x - \sum_{i=1}^n \log(x_i) \\ \text{subject to} \quad & Ax = b, \end{aligned} \tag{50}$$

which is 'better' behaved for large values of t. Using Newton's method for an equality constrained optimization problem, we solve for $x^*(t)$, the optimal solution to (50). Our solution $x^*(t)$ must be strictly feasible, i.e. $Ax^*(t) = b$ and $f_i(x^*(t)) \leq 0$ for all i, and must satisfy

$$\begin{aligned} 0 &= t\nabla f_0(x^*(t)) + \nabla\varphi(x^*(t)) + A^T\hat{\nu} \\ &= tc^T + \sum_{i=1}^m \frac{1}{x_i^*(t)} + A^T\hat{\nu} \end{aligned}$$

for some $\hat{\nu} \in \mathbb{R}^p$. To do this, given a strictly feasible x, we proceed with Newton's method with equality constraints. Here our KKT system can be written as

$$\begin{bmatrix} \nabla^2\varphi(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ w \end{bmatrix} = \begin{bmatrix} c - \nabla\varphi(x) \\ 0 \end{bmatrix}.$$

Using block elimination we can efficiently solve the system

$$\begin{aligned} A\nabla^2\varphi(x)^{-1}A^T w &= -A\nabla^2\varphi(x)^{-1}(c - \nabla\varphi(x)), \\ \Delta x_{nt} &= -\nabla^2\varphi(x)^{-1}(A^T w + (c - \nabla\varphi(x))), \end{aligned}$$

or more compactly

$$\begin{aligned} AH^{-1}A^T w &= -AH^{-1}(g), \\ \Delta x_{nt} &= -H^{-1}(A^T w + g), \end{aligned}$$

where H is the Hessian of the objective function and g is its gradient. This process is repeated until we have minimized (50). We then increase t and solve the above problem again, using the optimal $x^*(t)$ just found as our feasible starting point for the next minimization problem.

From this last example we see that to solve the minimization problem for a given $t > 0$, we find a strictly feasible $x^*(t)$ so that $Ax^*(t) = b$ and $f_i(x^*(t)) \leq 0$ for all i, and that satisfies

$$\begin{aligned} 0 &= t\nabla f_0(x^*(t)) + \nabla\varphi(x^*(t)) + A^T\hat{\nu} \\ &= t\nabla f_0(x^*(t)) + \sum_{i=1}^{m} \frac{\nabla f_i(x^*(t))}{-f_i(x^*(t)} + A^T\hat{\nu} \end{aligned}$$

for some $\hat{\nu} \in \mathbb{R}^p$. This is the general approach and for each t we solve

$$\begin{array}{ll} \text{minimize} & tf_0(x) - \sum_{i=1}^{m} \log(-f_i(x)) \\ \text{subject to} & Ax = b, \end{array} \tag{51}$$

using Newton's Method for equality constraints. Of course, the solution to (45) for some $t > 0$ will not be the same as the solution to (42), it will only be an approximation. So, since we are interested in the inequality constrained problem, we should be eager to know how good of an approximation the solution to (45) is for (42). The solutions $x^*(t)$ for increasing values of t are called the *central points,* and a collection of them make up the *central path.* From the characterization of these points above, we remark that each central point yields a dual feasible point $(\lambda * (t), \nu^*(t))$ which gives a lower bound for p^*, the optimal value of the original inequality constrained minimization problem. Observe that if we define $(\lambda * (t), \nu^*(t))$ as follows,

$$\lambda_i^*(t) = \frac{-1}{f_i(x^*(t)}, \ \nu^*(t) = \hat{\nu}/t,$$

then $(\lambda * (t), \nu^*(t))$ is a dual feasible point. To show this, we first notice that $\lambda^*(t) \succ 0$ since $f_i(x^*(t)) < 0$ for all i. We next realize that our optimality condition can be expressed as

$$\nabla f_0(x^*(t)) + \sum_{i=1}^{m} \lambda_i^*(t)\nabla f_i(x^*(t)) + A^T\nu^*(t) = 0.$$

This shows that $x^*(t)$ minimizes the Lagrangian

$$L(x, \lambda^*(t), \nu^*(t)) = f_0(x) + \sum_{i=1}^{m} \lambda^*(t) f_i(x) + A^T(\nu^*(t) - b),$$

thus $\lambda^*(t), \nu^*(t)$ is a dual feasible pair and

$$g(\lambda^*(t), \nu^*(t)) = f_0(x^*(t)) + \sum_{i=1}^{m} \lambda^*(t) f_i(x^*(t)) + \nu^*T(Ax^*(t) - b) = f_0(x^*(t)) - m/t.$$

So we have the lower bound on p^* given by

$$f_0(x^*(t)) - m/t \leq p^* \implies f_0(x^*(t)) - p^* \leq m/t.$$

In other words, $x^*(t)$ is no more than m/t-suboptimal, and thus as $t \to \infty$ we see that $x^*(t)$ converges to the optimal point x^* of the original inequality constrained problem given in (42).

7 Part I Conclusion

7.1 Recap

Thus far, we have developed an understanding of convex optimization, and have seen many examples of its use for problems that are both convex and those that are not.

Because of the robust understanding of convex optimization, it is a standard approach in applied mathematics to cast a well-known problem as a convex problem, as it enables us to immediately say many things about the problem, as well as allowing us to use the *numerous* techniques known to solve the problem efficiently.

This should not be confused with the process of formulating the dual problem as discussed in Chapter 5. Though the dual can often be very useful and informative to a general optimization problem, there are often more creative ways of reformulating the primal problem itself to have a convex structure. As discussed in the introduction, the process of casting any given problem as a convex problem in this way is extremely non-trivial and has not been a part of the discussion of this thesis. However, this work is very interesting, and if a convex problem can be formed it can often provide a new perspective and useful insight into the original problem, which should be celebrated and welcomed with open arms.

7.2 Let's Get Creative

The final chapter of this document contains a convex formulation of the *image registration problem* within a specific context. It is a re-implementation of a paper published by Taylor and Bhusnurmath [**Taylor08**], and is of particular interest to me because of some previous summer research I participated in that focused on mathematical image processing.

For me, this provided a very interesting and valuable take on the material presented in the previous 6 chapters. My first read through their paper had me excited to implement the procedure and see what it could do. I assumed I would be finished with the implementation within a week, but we all know what assuming does to a person, and I am no exception. Yet my struggle with the following work was healthy in the sense that I was able to understand how difficult this creativity can be, and how hard it can be to bring all of this beautiful mathematics to life.

Part II

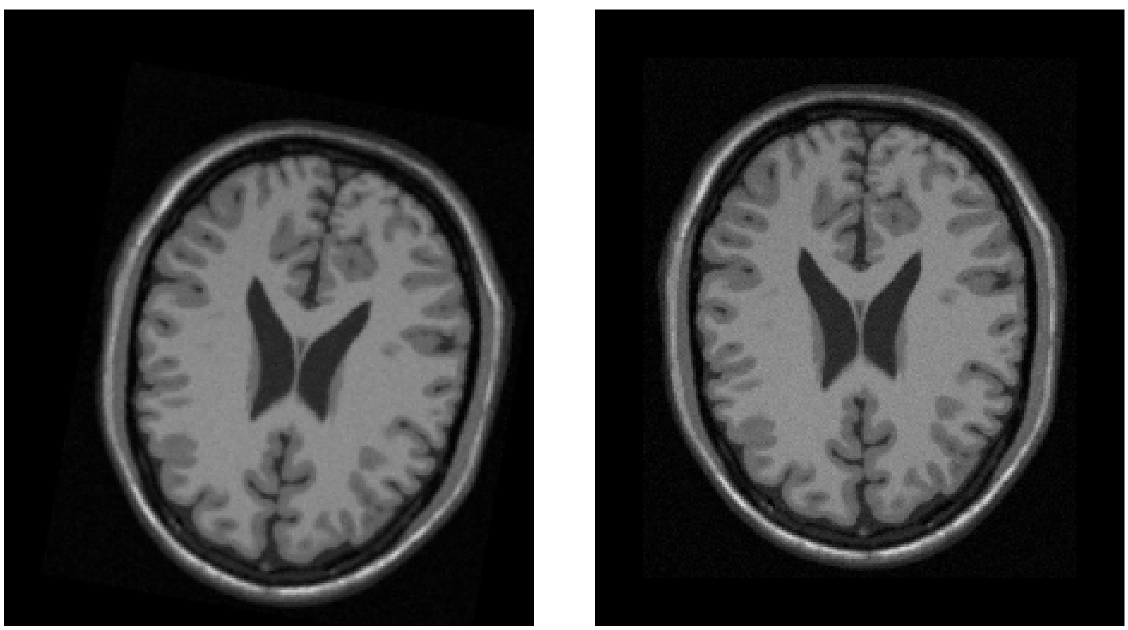
The Image Registration Problem

8 A Convex Image Registration Algorithm

The image registration problem is an optimization problem concerned with finding a map or transformation between two sufficiently 'similar images' (5). At the core of this problem is the idea that two images of the same object should be able to be considered 'the same' in some context. This leads naturally into the study of image metrics, which seek to answer what it means, mathematically, for two images to be the same. Is there a way to treat images like points in a space, for which we may define a metric? The image registration problem goes a step further, and asks how similar the two images can be to one another after a transformation from a specific set is applied to one of the images.

Continuing our analogy for points in a space here fails. We can certainly always find a function which maps one point to another, in which case the distance between these two points would be 0, but this misses the heart of the image registration problem. Indeed, if we have one image of a cat in a field, and another of a steaming cup of coffee, we should have no interest in finding the optimal transformation of one to the other because they have little business being compared.

Keeping this in mind, we formulate the standard image registration problem below. Let's consider two images A and B, as show below. To our eyes, it is no stretch to call these images 'similar.'



(a) Target image A

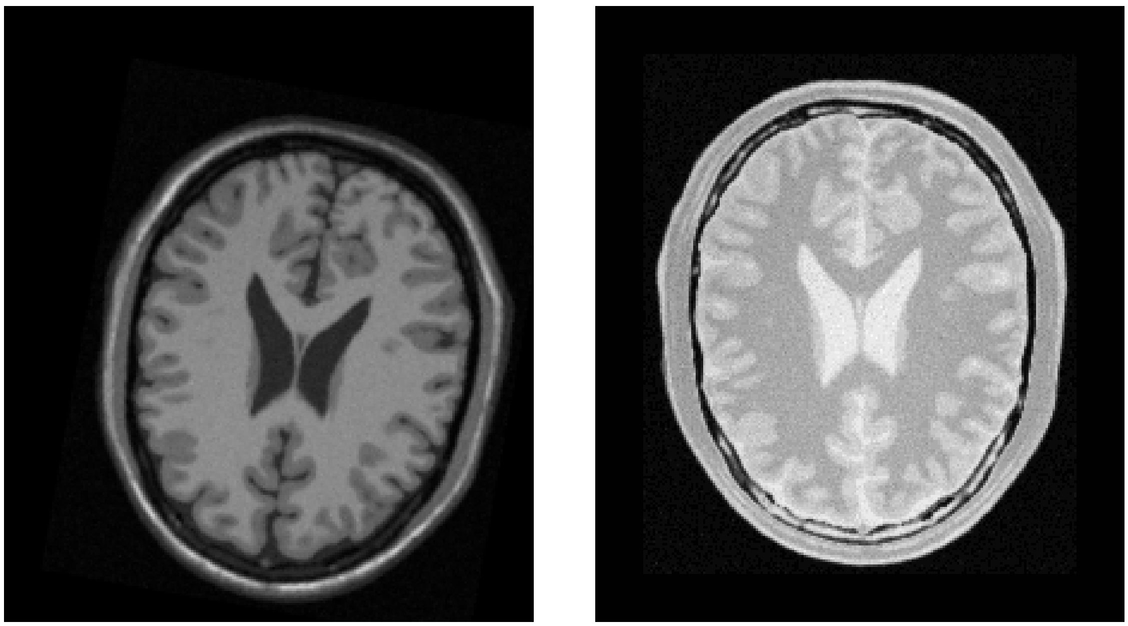
(b) Base image B

Figure 5: 'Similar' Images

We will refer to the *target image*, A, as the image on the left, and the *base image* B as that on the right of Figure (5). Our goal will be to find a transformation that maps our base image onto our target image. These two images clearly depict the same object, but with different orientations. The idea of image registration is to find a transformation from a base image onto a target image. So we seek to find the 'best' transformation of the base image, B, namely the transformation that makes

B indistinguishable from A. In order to do this we need some concept of the distance between two images. This will be the formulated as a function D that takes two images and outputs some real number.

Of course, this function D may need to have different properties to adjust to how we want to measure similarity. In contrast to the cat and coffee, consider the images in Figure (6).



(a) Target image A

(b) Base image B

Figure 6: 'Similar' Images of different modality

As humans we would say with confidence that these are two images in Figure (6) are of the same object, namely a slice of some brain. However, to consider these images to be the same, or 'similar' mathematically can be quite challenging. The mutual information metric discussed attempts to deal with this particular situation. For more details see [**miRegistrationSurvey**]. In general, similarity must be defined based upon what images the user wishes to consider similar, and this must be reflected in the mathematical construction of D.

8.1 The Image Registration Problem

In image processing the registration problem is formulated mathematically as follows:

Given two images A, B, a distance function D, and a transform space τ, find the transformation Φ such that

$$\Phi = \operatorname{argmin}_{\varphi \in \tau} D(A, \varphi(B)). \tag{52}$$

For two fixed images, A and B, this is an unconstrained optimization problem

$$\underset{\varphi \in \tau}{\text{minimize}} \quad D(A, \varphi(B)), \tag{53}$$

the domain of which is the transformation space τ. In general there is no reason to assume that the function space τ is a convex, much less that D is a convex function on τ.

8.2 Applications of Image Registration

The image registration problem (52) is of significant interest in many medical imaging contexts. This is because it is very common to have multiple images of the same patient taken at different times, with different modalities (i.e. MRI vs. CT Scan), and at different orientations. Even more explicitly, any 3-D medical image is constructed via a sequence of 2-D slices, which form a sequence of 'similar' images.

The ability to mathematically describe the map between any two images allows for processes such as segmentation to be extrapolated from one image to another 'similar' image, and thus potentially provide many segmentations of a set of images from a single segmentation. In many cases these segmentations are done by hand, so this process could be extremely useful in this context.

Furthermore, the ability to register two 'similar' images of a particular patient allows for a more meaningful measurement of how the structure of the patients brain, lung, etc., has changed from one time to another. For more information see [**Maintz98asurvey**].

Outside of medical applications, image registration is heavily studied in computer vision for many different purposes such as target recognition and autonomous vehicle navigation. It is also vital for data analysis of sets of images collected in a time series such images from as satellites. These applications are discussed in detail in [**Brown92asurvey**].

8.3 Image Metrics

We begin with a discussion of the distance function D described in the optimization problems in (52) and (53). There are several common 'metrics' or distance functions used in image processing to form a notion of distance or difference between two images, each aimed at capturing some form of similarity.

8.3.1 Mean Squares Metric

The *Mean Squares Metric* calculates the sum of the squared differences between two images A, B: $\mathbb{R}^{m\times n} \to \mathbb{R}$,

Definition 40 (*Mean Squares Metric*)**.** Given two images A and B, the *Mean Squares Metric* is the function

$$MS(A,B) = \frac{1}{mn}\sum_{x=1}^{n}\sum_{y=1}^{m}(A_{xy} - B_{xy})^2, \tag{54}$$

where A_{xy} and B_{xy} represent the intensity values at pixel (x, y) of the images A and B respectively.

Note: for the remainder of this discussion, the pixel at location (x, y) refers to the pixel in row y, column x. This is done so that we may consider our image to lie in the first quadrant of the usual xy-plane so that our discussion of image transformations is less cumbersome.

We take the standard approach that grayscale images are 8-bit and thus have intensity values in the set $\{0, 1, \ldots 255\}$, where 0 represents black, and 255 represents white. If they are color images (RGB), then each pixel is assigned 3 values, each in the set $\{0, 1, \ldots 255\}$, and interpreted similarly. If A and B are colored images, i.e. if A, B: $\mathbb{R}^{m\times n} \to \mathbb{R}^3$, then A_{xy} and B_{xy} are vectors in $\mathbb{R}^3$, and we interpret $(A_{xy} - B_{xy})^2$ as the 2-norm $\|A_{xy} - B_{xy}\|_2^2$.

It is clear that

$$MS(A,B) = 0 \Leftrightarrow A = B,$$

which is a property we would like any of these distance functions to have. It is also clear that for small amounts of corruption to the intensity values of an image B will cause small changes in the output of the MS function. What is not so clear is how this function behaves for simple transformations such as scaling or rotation. Depending on the structure of an image A, even a small rotation or translation of A, call it A', can 'fool' the MS metric into thinking A and A' are very different images, as intensity values can vary largely from pixel to pixel.

8.3.2 Normalized Correlation Metric

Another standard image metric is the *Normalized Correlation Metric* computes pixel-wise cross-correlation and normalizes the output by the square root of the autocorrelation function.

Definition 41. (*Normalized Correlation Metric*)
Let A, B be $m \times n$ images. Then the normalized correlation metric between A and B is

$$NC(A,B) := -\frac{\sum_{i=1}^{n\cdot m}(A_i \cdot B_i)}{\sqrt{\sum_{i=1}^{m\cdot n} A_i^2 \cdot \sum_{i=1}^{m\cdot n} B_i^2}}$$

where $A_i \cdot B_i$ is the element-wise multiplication of the images, and the negative value causes the optimized return to be a minima.

The optimal value of the normalized correlation metric is -1, and thus can be made to have optimal value 0 by a shift of 1. The metric requires the compared images to be of the same modality.

We may derive a few more results about this metric by realizing it as an inner product[1]. We can write the normalized correlation metric as

$$NC(A,B) = -\langle \frac{A}{||A||}, \frac{B}{||B||} \rangle.$$

In this regard it is clear that the optimal value is -1, and that any scalar multiples of the image A gives

$$\begin{aligned} NC(A, cA) &= -\langle \frac{A}{||A||}, \frac{cA}{c||A||} \rangle \\ &= -\langle \frac{A}{||A||}, \frac{A}{||A||} \rangle \\ &= -1. \end{aligned}$$

So we see that the Normalized Correlation Metric is able to disregard scalar transformations of intensity. In this way it is more general or robust than the Mean Squares Metric. However, it should be noted that one may not always want to ignore these intensity scalings, so the utility of the metric should be gauged based upon the context of the problem.

8.3.3 Mutual Information Metric

A third commonly used image metric is the *Mutual Information Metric.* As the name implies, it is an information theoretic metric that defines the distance between two images based upon probability distributions of pixel values. This metric is the primary metric used in medical imaging contexts as it may be applied for images of different modalities. However, it will not be used in our convex formulation of the image registration problem, so all further discussion of this metric is omitted. A good discussion of this metric and a survey of its usage in registration can be found here [**miRegistrationSurvey**].

[1] In this case we use the vector inner product, that is $\langle A, B \rangle = \sum_{i,j} a_{i,j} b_{i,j}$.

8.3.4 Image Metrics and Image Registration

In the registration problem we are concerned precisely with minimizing the distance function within a given transformation space τ. These transformation spaces are usually divided into rigid or non-rigid. An example of a rigid transformation would be a linear transformation, whereas a non-rigid is generally modeled by radial basis functions or b-splines. See [**Crum04non-rigidimage**], [**principalwarps:**] and [**deformableTemplateKinematics**]. As is the case with b-splines, it is often helpful to parameterize In a practical sense we will have to search the given transformation space for a transformation φ which makes $D(A, \varphi(B))$ minimal. The discussion above illustrates there are many local minimums in this space, which makes finding the optimal transformation difficult in practice. These local minimums are completely dependent upon the distance function itself. To alleviate this issue we want to construct some sort of convex approximation of the distance function, which in theory will allow us to effectively search the transform space and converge to a global minimum distance between A and $\varphi(B)$, thus solving the registration problem with $\Phi = \varphi$. It should be noted that this is a well-studied problem, and that there are many different approaches and current solutions to it that do not demand the problem to be a convex optimization problem. [**Brown92asurvey**]. For our purposes, however, we are exploring ways in which this problem can be tackled using the techniques of convex optimization.

8.4 Non-convexity of Mean Squares

Supposing we are working with the Mean Squares Metric and looking at affine transformations, the image registration problem here is to find the affine transformation Φ that minimizes $MS(A, \varphi(B))$. The first way to approach this problem would be to search the transformation space until a minimum of $MS(A, \varphi(B))$ is found. This could be framed as the problem

$$\underset{\varphi \in \tau}{\text{minimize}} \quad MS(A, \varphi(B)) = \frac{1}{mn} \sum_{x,y} e_{xy}(\varphi(B)_{xy}), \tag{55}$$

where

$$e_{xy}(\varphi(B)_{xy}) = \|A_{xy} - \varphi(B)_{xy}\|_2^2$$

is the function that finds the l^2 difference between the intensity values in row y column x of the target image and the transformed base image.

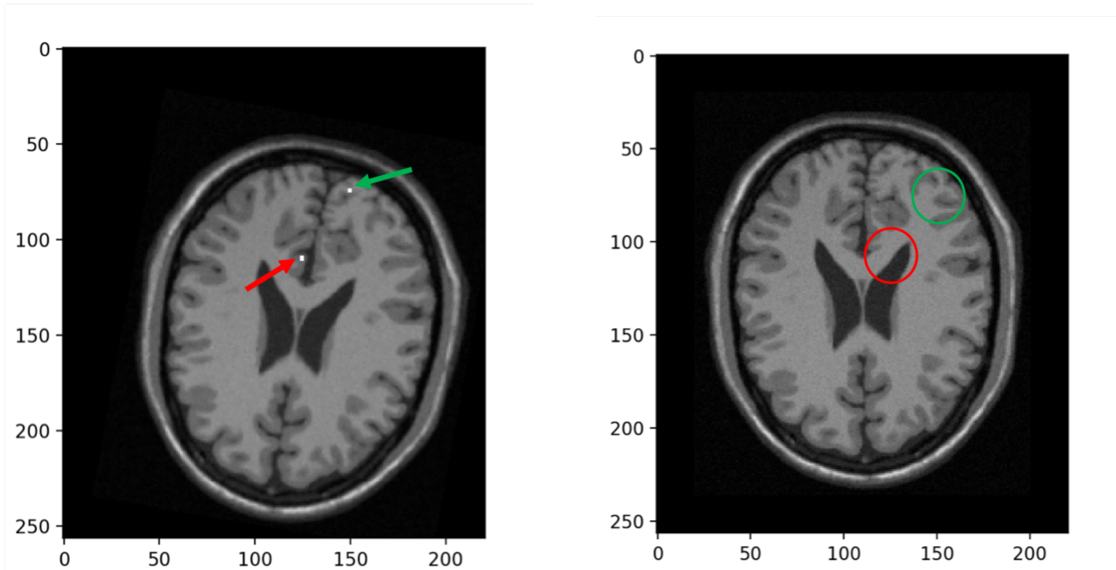
To do this we begin by looking at every individual pixel in the target image. For the pixel at location (x, y) we look at the distance (l^2 distance on intensity values) between the target image and a corresponding window (of predefined size) in the base image. To exemplify what this process looks like, consider two pixels in the target image, $(125, 110)$ and $(150, 75)$. We search equal windows around them in the base image to see how the pixel intensities vary.

In Figure (7) we isolate two pixels of the target image, seen in white, and a corresponding window in the base image. The region circled in red is qualitatively more 'distinct' than that in green, as the corresponding region in the base image contains a sharp edge or feature of the brain. In this sort of region our l^2 difference will vary substantially in a given window. In contrast, the region circled in green has less features and so our l^2 difference is more uniform. For our standard optimization problem we search for a transformation φ so that the sum of these l^2 distances between each pair of pixels is minimal, or

$$\sum_{ij} e_{ij}(\varphi(b)_{ij})$$

is minimized, and ideally zero.

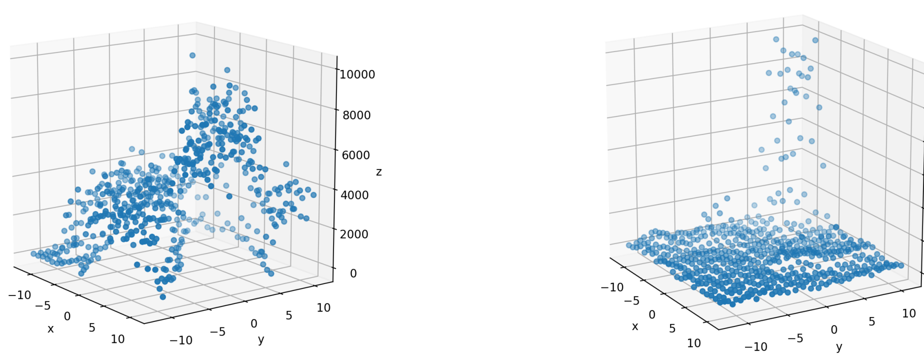
From the scatter plots in Figure (8), most notably in (a), around a distinct feature of the image, this difference will have many local minima, as discussed previously.



(a) Target image, red arrow pointing to pixel $(125, 110)$ with more distinct features, green arrow pointing to pixel $(150, 75)$ in a more homogeneous region.

(b) Base image, red circle around pixel $(125, 110)$ with more distinct features, green circle around pixel $(150, 75)$ in a more homogeneous region.

Figure 7: ‘Distinct’ and ‘homogeneous’ regions



(a) l^2 difference between Target image at pixel $(125, 110)$ and Base image at $(125+i, 110+j)$ for i, j varying from -12 to 12 independently.

(b) l^2 difference between Target image at pixel (150, 75) and Base image at $(150+i, 75+j)$ for i, j varying from -12 to 12 independently.

Figure 8: l^2 Differences

Our objective is to construct a global convex approximation of the MS metric, MS′, and thus to state this problem as a convex optimization problem,

$$\underset{\varphi}{\text{minimize}} \quad \text{MS}'(A, \varphi(B)). \tag{56}$$

To construct MS' appropriately, we must first determine how our transformations will be defined.

8.5 Deformation Models

For the purpose of creating a convex function space in which to search, we model the image deformations as a displacement at each pixel (x, y) in the following way: construct two scalar functions, $D_x(x, y, \mathbf{p_X}), D_y(x, y, \mathbf{p_Y})$ which give the displacement from the point (x, y) in terms of some deformation parameters $\mathbf{p_X}$, and $\mathbf{p_Y}$. We assume that D_x and D_y are linear functions of $\mathbf{p_X}$, and $\mathbf{p_Y}$ respectively. *i.e.* we could model an affine transformation in the following way:

$$D_x(x, y, \mathbf{p_X}) = \mathbf{p_X}^T(1, x, y)$$
$$D_y(x, y, \mathbf{p_Y}) = \mathbf{p_Y}^T(1, x, y),$$

for $\mathbf{p_X}, \mathbf{p_Y} \in \mathbb{R}^3$. The parameters

$$p_X = (0, 0, 0), \quad p_Y = (0, 0, 0)$$

result in no displacement and thus represent the identity transformation. The parameters

$$p_X = (13, 0, 0), \quad p_Y = (17, 0, 0)$$

correspond to the displacements

$$D_x(x, y, \mathbf{p_X}) = 13, \quad D_y(x, y, \mathbf{p_X}) = 17$$

and thus represent a horizontal translation 13 units to the right, and a vertical translation 17 units 'down'.

In the general case for an affine function, we have basis vectors of the displacements in terms of x, y and some constant, which we will take to be 1. With this structure, we could represent all the horizontal displacements by the vector $D_X = C\mathbf{p_X}$, where C is the concatenation of all the basis vectors of the deformation,

$$C^T = \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & 1 & \cdots & x & \cdots & n \\ 0 & 0 & \cdots & y & \cdots & m \end{bmatrix}.$$

Similarly, we have $D_Y = C\mathbf{p_Y}$. Again considering the translation example with $p_X = (13, 0, 0)$ and $p_Y = (17, 0, 0)$, this results in the vectors

$$D_X = (13, 13, \ldots, 13), \quad D_Y = (17, 17, \ldots, 17).$$

With this approach our function space is simply $\mathbb{R}^k$, where k is the number of parameters necessary to describe the displacements. If we wish to restrict the parameters so that we have maximum and minimum displacement values, then our function space could be some convex subset of $\mathbb{R}^k$. In the case of the affine transform our function space is $\mathbb{R}^6$.

In this work we have implemented the above deformation as well as deformations of higher order polynomials in x and y, along with radial functions of x and y.

8.5.1 Second, Third and n^{th} Order Models

Similarly to our affine, or first order, deformation model listed above, we can model our deformations using polynomials of higher order. The Second Order model is constructed using the basis vector $(1, x, y, xy, x^2, y^2)$, with parameters $\mathbf{p_X}, \mathbf{p_Y} \in \mathbb{R}^6$, and the Third Order is constructed using the basis vectors $(1, x, y, xy, x^2, y^2, x^2y, xy^2, x^3, y^3)$, with parameters with parameters $\mathbf{p_X}, \mathbf{p_Y} \in \mathbb{R}^{10}$. We can construct an n^{th} order model in the same way, with $\mathbf{p_X}, \mathbf{p_Y} \in \mathbb{R}^{\binom{n+2}{2}}$.

8.5.2 Gaussian Deformation Model

The Gaussian Deformation model described by Taylor and Bhusnurmath is given by the basis vectors $(1, x, y, e^{-(r_1/\sigma)^2}, \ldots, e^{-(r_i/\sigma)^2}), \ldots e^{-(r_k/\sigma)^2})$, where $r_i = \|(x, y) - (x_i, y_i)\|$ for a given set of symmetrically placed kernels (x_i, y_i) throughout the image. To reproduce the work in this paper, we used 16 kernels arranged in a four by four grid across the image. This gives us parameters $\mathbf{p_X}, \mathbf{p_Y} \in \mathbb{R}^{19}$.

8.6 A Convex Approximation of Mean Squares

At this point we have developed a deformation model that describes transformations by vectors in $\mathbb{R}^N$, so that we have a convex transformation space. We now formalize the convex approximation of the Mean Squares Metric proposed by Taylor and Bhursnurmath [**Taylor08**], and state the image registration problem as a linear program.

8.6.1 Setting up the deformation model

Given a target image A and base image B, we model the deformation by two scalar functions, $D_x(x, y, \mathbf{p_X})$, and $D_y(x, y, \mathbf{p_Y})$ as described previously, which define the displacement from the point (x, y) in terms of some deformation parameters $\mathbf{p_X}$, and $\mathbf{p_Y} \in \mathbb{R}^N$. In this way our transformation φ is defined at any particular pixel (x, y) by

$$\varphi(B)_{x,y} = (x + D_x(x, y, \mathbf{p_X}),\;\; y + D_y(x, y, \mathbf{p_Y})).$$

To form a convex optimization problem we new need a convex objective function to minimize subject to convex constraints.

8.6.2 Constructing Convex the Lower Bound

With our deformations as linear functions of the parameters $\mathbf{p_X}$, and $\mathbf{p_Y}$, we now revisit our idea of the convex Mean Squares Metric. The proposed method by Taylor and Bhusnurmath is to construct a convex lower bound e'_{xy} for each individual objective function at a given target pixel (x, y) by replacing the error surface with its lower convex hull.

In Figure (9), we construct the lower convex hull for the l^2 differences around the pixels $(125, 110)$ and $(150, 75)$ respectively. These are seen by the green facets. We will see that the coefficients defining these facets will provide the constraints for our new minimization problem.

Letting E be the convex lower hull, then, for a proposed transformation defined by the pair $\mathbf{p_X}, \mathbf{p_y}$, each pixel has the constraints that the value of the error function at that point must be at least the value of the error surface at that point, *i.e.* $e'_{xy}(\varphi(B)_{xy}) \geq E(x, y)$.

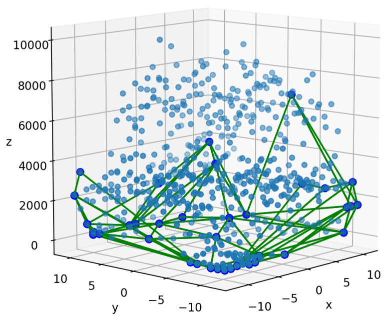
Thus we now seek a transformation φ to solve the new minimization problem

$$\underset{\varphi}{\text{minimize}} \quad \sum_{x,y} e'_{xy}(\varphi(B)_{xy}). \tag{57}$$

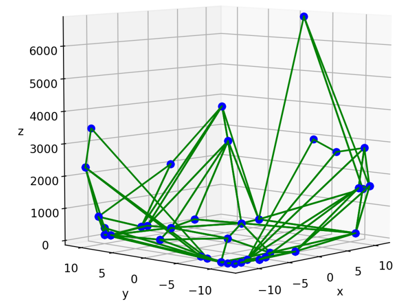
Remarks:

(i) It is important to note that this is a convex approximation of the *MS metric relative to to A and B*.

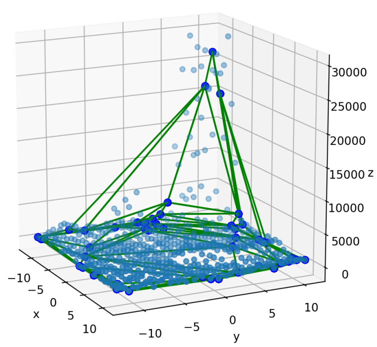
(ii) In both of the cases above it is evident this approximation can significantly underestimate the cost associated with choosing a particular pixel in the base image. This is seen by points on the error surface that lie far above the lower convex hull. However, because our objective function is a sum over the convex lower bound for each pixel given some deformation, these individual mismatches may be ameliorated.



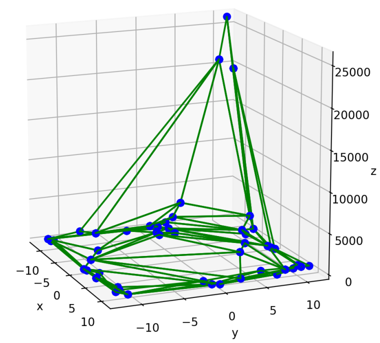
(a) Convex lower bound near 'distinct feature' at $(x, y) = (125, 110)$ of image A with respect to image B



(b) Convex Lower Hull used to approximate MS metric for any given displacement from (125, 110)



(c) Convex lower bound of homogeneous region at $(x, y) = (150, 75)$ of image A with respect to image B



(d) Convex Lower Hull used to approximate MS metric for any given displacement from (150, 75)

Figure 9: Convex Lower Hull of a 'Distinct' and a 'Homogeneous' region

(iii) We also notice that in a homogeneous region of the base image B our approximation provides little constraint to the deformation parameters, whereas a detailed part of the base image can create steep facets that have more control over the optimal deformation.

8.7 Formulating the Linear Program

We now seek to form this problem as a linear program.

8.7.1 Recasting Convex Hull Constraints as Linear Constraints

To capture the convex approximation of the Mean Squares metric as linear constraints, we introduce an auxiliary variable $z(x, y)$ for each pixel in the target image which must be larger than the value of the convex error surface $E(x, y)$. Practically, this is accomplished by ensuring that the point $(x, y, z(x, y))$ lies above all of the planar facets which comprise the corresponding convex lower hull, or convex error surface, E. So if $a_x^i(x, y), a_y^i(x, y), b_i(x, y)$ are the coefficients corresponding to a planar facets in the lower convex hull around (x, y), there is a corresponding equation defining the plane in $\tilde{x}, \tilde{y}, \tilde{z}$, given by:

$$a_x^i(x, y)\tilde{x} + a_y^i(x, y)\tilde{y} + a_z^i(x, y)\tilde{z} = b_i(x, y),$$

with $\left\|(a_x^i, a_y^i, a_z^i)\right\|_2 = 1$. Here $\tilde{x}$ and $\tilde{y}$ represent the displacements $D_x(x, y, \mathbf{p_X})$ and $D_y(x, y, \mathbf{p_Y})$ from (x, y). The lower facets are those for which $a_z^i < 0$, so for each of those facets we replace $a_z^i\tilde{z}$ with $-z(x, y)$, and ensure that $z(x, y)$ is such that our point $(D_x, D_y, z(x, y))$ lies above the lower facet defined by $a_x^i(x, y), a_y^i(x, y)$ and $b_i(x, y)$, or that

$$a_x^i(x, y)D_x(x, y, \mathbf{p_X}) + a_y^i(x, y)D_y(x, y, \mathbf{p_Y}) - z(x, y) \geq b_i(x, y).$$

Each lower convex hull will have a variable number of facets, F_j, which define it, and thus each auxiliary variable $z(x, y)$ has F_j constraints associated with it, where j varies from 1 to mn, the total number of pixels in the images. Thus, by ensuring $z(x, y)$ satisfies

$$a_x^i(x, y)D_x(x, y, \mathbf{p_X}) + a_y^i(x, y)D_y(x, y, \mathbf{p_Y}) - z(x, y) \geq b_i(x, y),$$

for all i, we ensure the point $(D_x, D_y, z(x, y))$ lies above the lower convex hull, and we have a cost associated with the choice of displacements D_x and D_y. Doing this for each pixel in the target image, we can rephrase the minimization problem (57) as

$$\begin{aligned} &\text{minimize}_{z, \mathbf{p_X}, \mathbf{p_Y}} \sum_{x,y} z(x, y) \\ &\text{subject to} \\ &z(x, y) \geq a_x^i(x, y)D_x(x, y, \mathbf{p_X}) + a_y^i(x, y)D_y(x, y, \mathbf{p_Y}) - b_i(x, y) \text{ for all } i, \text{ for all } x, y. \end{aligned} \tag{58}$$

8.8 Solving the Linear Program

Before attempting to solve the LP given in (58), we do some heavy notational lifting for a cleaner implementation.

8.8.1 Planar Facet Coefficient matrices

As stated in the previous section, each pixel (x, y) in the target image has associated with it a convex hull that comes with some number planar facets which define it. We called this number F_j. Each of these facets can in turn be described by three coefficients in $\mathbb{R}$, which we denoted as a_x^i, a_y^i and b_i for

$i = 1, \ldots, F_j$. By defining the matrix

$$A = \begin{bmatrix} a_1^1 & 0 & \cdots & 0 \\ a_2^1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 \\ a_{F_1}^1 & 0 & \cdots & 0 \\ 0 & a_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & a_1^{mn} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & a_{F_{mn}}^{mn} \end{bmatrix},$$

we may collect the facet coefficients in the matrices A_X, A_Y, and I_z. This forms an $F \times mn$ matrix where

$$F = \sum_{j=1}^{mn} F_j,$$

is the total number of planar facets used to construct the convex hulls. Each column of the matrix A contains the coefficients associated with a particular pixel (x, y) and has been listed in row major order. The matrix I_z contains only 1's as entries for any non-zero entry in A_X or A_Y, and corresponds to the coefficient of 1 on each of our auxiliary variables $z(x, y)$. We also let b be the vector of length F constructed by concatenating the scalars b_i for $i = 1, \ldots, F$. So

$$b = (b_1, b_2, \ldots, b_F).$$

Now, making use of the concatenated displacements

$$D_X = C\mathbf{p_X} \text{ and } D_Y = C\mathbf{p_Y},$$

we may rewrite (58) as

$$\begin{aligned} &\text{minimize }_{\mathbf{p_X},\mathbf{p_Y},z} \mathbb{1}^T z \\ &\text{subject to} \\ &A_X C\mathbf{p_X} + A_Y C\mathbf{p_Y} - I_z z \preccurlyeq b. \end{aligned} \tag{59}$$

The inequality

$$A_X C\mathbf{p_X} + A_Y C\mathbf{p_Y} - I_z z \preccurlyeq b$$

economically captures the constraints imposed by each lower convex hull that defines the convex approximation.

8.8.2 Displacement Bounds

The barrier method discussed in Chapter 7 can readily be applied to solve (59), but we may also benefit from adding additional constraints on the displacements we allow. Letting $b_{X_u}, b_{X_l}, b_{Y_u}, b_{Y_l} \in \mathbb{R}^{mn}$, we have the related problem

$$\begin{aligned} &\text{minimize }_{\mathbf{p_X},\mathbf{p_Y},z} \mathbb{1}^T z \\ &\text{subject to} \\ &A_X C\mathbf{p_X} + A_Y C\mathbf{p_Y} - I_z z \preccurlyeq b, \\ &b_{X_l} \preccurlyeq C\mathbf{p_X} \preccurlyeq b_{X_u} \\ &b_{Y_l} \preccurlyeq C\mathbf{p_Y} \preccurlyeq b_{Y_u}. \end{aligned} \tag{60}$$

8.8.3 Implementing the Barrier Method

For the sake of minimizing notation and better illustrating the technique, we focus on the program given in (59) with no displacement bounds. The following can be readily extended to include these constraints. The Barrier Method produces a sequence of minimization problems of the form

$$\text{minimize}_{\,\mathbf{p_X},\mathbf{p_Y},z} \quad f(\mathbf{p_X},\mathbf{p_Y},z) = t\mathbb{1}^T z - \sum_{i=1}^{F} \log\left(-(b_i + \mathbb{1}^T z - a_X^i c_i^T \mathbf{p_X} - a_Y^i c_i^T \mathbf{p_Y})\right), \tag{61}$$

for increasing values of t.

We denote

$$s = (s_1, s_2, \ldots, s_F)$$
$$\text{for } s_i = -(b_i + \mathbb{1}^T z - a_X^i c_i^T \mathbf{p_X} - a_Y^i c_i^T \mathbf{p_Y}).$$

From (61) we have gradients given by

$$\begin{aligned} g_{\mathbf{p_X}} &= C^T A_X^T s^{-1}, \\ g_{\mathbf{p_Y}} &= C^T A_Y^T s^{-1}, \\ g_z &= t\mathbb{1} - I_z^T s^{-1}. \end{aligned}$$

Further we have the Hessian

$$H = \begin{bmatrix} \frac{\partial^2}{\partial \mathbf{p_X}^2} f & \frac{\partial^2}{\partial \mathbf{p_X} \partial \mathbf{p_Y}} f & \frac{\partial^2}{\partial \mathbf{p_X} \partial z} f \\ \frac{\partial^2}{\partial \mathbf{p_X} \partial \mathbf{p_Y}} f & \frac{\partial^2}{\partial \mathbf{p_Y}^2} f & \frac{\partial^2}{\partial \mathbf{p_Y} \partial z} \\ \frac{\partial^2}{\partial z \partial \mathbf{p_X}} f & \frac{\partial^2}{\partial z \partial \mathbf{p_Y}} f & \frac{\partial^2}{\partial z^2} f \end{bmatrix},$$

given by

$$H = \begin{bmatrix} C^T D_1 C & C^T D_2 C & -D_3 C \\ C^T D_2 C & C^T D_4 C & -D_5 C \\ -D_3 C & -D_5 C & D_6 \end{bmatrix},$$

where D_1, D_2, D_3, D_4, D_5 and D_6 are the diagonal matrices

$$\begin{aligned} D_1 &= A_X^T \mathbf{diag}(s^{-2}) A_X, \\ D_2 &= A_X^T \mathbf{diag}(s^{-2}) A_Y, \\ D_3 &= A_X^T \mathbf{diag}(s^{-2}) I_z \\ D_4 &= A_Y^T \mathbf{diag}(s^{-2}) A_Y, \\ D_5 &= A_Y^T \mathbf{diag}(s^{-2}) I_z, \\ D_6 &= I_z^T \mathbf{diag}(s^{-2}) I_z. \end{aligned}$$

Thus the KKT system defining our Newton Step can be written as

$$H \begin{pmatrix} \delta_{\mathbf{p_X}} \\ \delta_{\mathbf{p_Y}} \\ \delta_z \end{pmatrix} = - \begin{pmatrix} g_{\mathbf{p_X}} \\ g_{\mathbf{p_Y}} \\ g_z \end{pmatrix}. \tag{62}$$

We may write this more compactly by letting

$$H_p = \begin{bmatrix} C^T D_1 C & C^T D_2 C \\ C^T D_2 C & C^T D_4 C \end{bmatrix}$$
$$H_z = \begin{bmatrix} -D_3 C & -D_5 C, \end{bmatrix}$$

So that the system in (62) becomes

$$\begin{bmatrix} H_p & H_z \\ H_z^T & D_6 \end{bmatrix} \begin{pmatrix} \delta_{\mathbf{p}} \\ \delta_z \end{pmatrix} = - \begin{pmatrix} g_{\mathbf{p}} \\ g_z \end{pmatrix}$$

By utilizing the fact that D_6 is diagonal and of full rank we can solve for δ_z efficiently in (62) to get

$$\delta_z = D_6^{-1}(-g_z - H_z^T \delta_{\mathbf{p}}),$$

which can be substituted into

$$H_{\mathbf{p}}\delta_{\mathbf{p}} + H_z \delta_z = -g_{\mathbf{p}}$$

to yield the equation

$$(H_{\mathbf{p}} - H_z D_6^{-1} H_z^T)\delta_{\mathbf{p}} = g_p - H_z D_6^{-1} g_z.$$

If $\mathbf{p_X}, \mathbf{p_Y} \in \mathbb{R}^N$, then the above is a system of $2N$ linear equations. So our linear system defining the Newton step is only as large as the parameter space chosen to define the transformations. Thus the number of facets defining any convex hull used in our approximation does not add complexity to the system.

8.9 The Image Registration Algorithm

Following Taylor and Bhursnurmath [**Taylor08**] we introduce multiple image scales in order to downsize the image initially for the sake of computational efficiency.

To do this for some scale n, we downsize the target A and base B by a factor of 2^n via bilinear interpolation to produce A' and B'. We then redefine the basis vectors collected in the matrix C (8.5) in the down scaled images in order to remain in same physical space as the target image A. i.e.

$$x \mapsto 2^n x, \quad y \mapsto 2^n y.$$

In this way the pixel at (x, y) in the full size image remains at (x, y) in the down sized image. The registration the defines the function φ at each pixel by

$$\varphi(B)_{x,y} = (x + D_x(x, y, \mathbf{p_X}), y + D_y(x, y, \mathbf{p_Y})).$$

We can then define Φ on the original base image B by simply upscaling the parameters

$$\mathbf{p_X} \mapsto 2^n \mathbf{p_X}, \quad \mathbf{p_Y} \mapsto 2^n \mathbf{p_Y},$$

and so we have an interpolation of

$$\Phi(B)_{x,y} = (x + D_x(x, y, \mathbf{p_X}), y + D_y(x, y, \mathbf{p_Y})).$$

In this way we find an initial guess for the optimal deformation on the base image. By scaling up one factor at a time we are allowed to increase our window size while significantly decreasing the computational cost of doing so. Furthermore, as the window used to define the convex approximation of the mean squares metric (59) decreases, the approximate error surface becomes a better approximation of the actual error surface.

This multi-scale algorithm is outlined recursively below.

Algorithm 4: Unconstrained Convex Image Registration (Recursive)

given target A, base B, basis vectors C, window size w, number of scales n, $\alpha \in (0, 1)$, $\mathbf{p_X}, \mathbf{p_Y}$;
returns $\mathbf{p_X}, \mathbf{p_Y}$;
for: $j = n-1, \dots, 1, 0$;
(0) get current base $\varphi(B)$ defined by $\mathbf{p_X}, \mathbf{p_Y}$;
(1) downsize target, base by a factor of 2^j. $A \mapsto A', \varphi(B) \mapsto B'$;
(2) find relative window size $w' := w/2^j$;
(3) upscale basis vectors by $x \mapsto 2^j x, y \mapsto 2^j y$;
(4) register A' to B' by finding optimal $\mathbf{p_X}', \mathbf{p_Y}'$ defining φ' with window w';
(5) decrease window $w := \alpha w$;
(6) update parameters, scale back to physical space of target $\mathbf{p_X}$ += $2^j \mathbf{p_X}'$, $\mathbf{p_X}$ += $2^j \mathbf{p_X}'$;
(7) **if:** j = 0 exit;
(8) **else:** register A, B, with basis vectors C, window w, $j-1$ scales, $\mathbf{p_X}, \mathbf{p_Y}$.

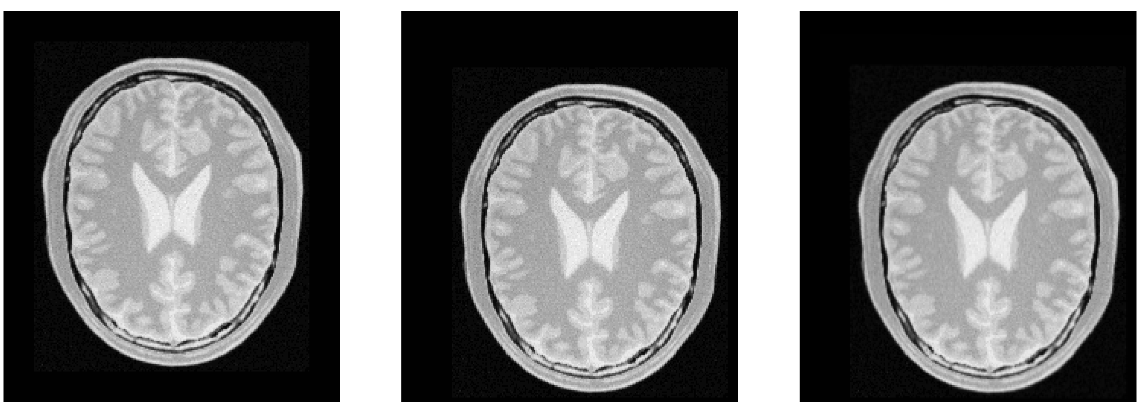
8.10 Examples

8.10.1 Affine Transformation Mode

The following registrations were found using the affine or first order deformation model given by

$$D_x(x, y, \mathbf{p_X}) = \mathbf{p_X}^T(1, x, y)$$
$$D_y(x, y, \mathbf{p_Y}) = \mathbf{p_Y}^T(1, x, y),$$

for $\mathbf{p_X}, \mathbf{p_X} \in \mathbb{R}^3$.



(a) Base (b) Target (c) Registered

Figure (10) shows several examples of the registration algorithm deployed to recover some basic rotations and shifts. The image on the far left is the base image, the middle is the target, and the far right is the result of the recovered registration. In general the algorithm does a good job of recovering these translations and rotations.

With these simple transformations it is easy to quantify the success of the registrations in Figure (10). We can represent a translation and rotation using the linear transformation

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) & h \\ \sin(\theta) & \cos(\theta) & v \\ 0 & 0 & 1, \end{bmatrix}$$

(a) Base (b) Target (c) Registered

Figure 10: Affine Transformation Examples

where θ is the counterclockwise rotation, h is the horizontal shift, and v is the vertical shift. The pure translation example is a shift 13 units right, and 17 units down, therefore we we have

$$\theta = 0, \quad h = 13, \quad v = 17,$$

and so applying the corresponding tranformation to a pixel (x, y) gives

$$\begin{bmatrix} 1 & 0 & 13 \\ 0 & 1 & 17 \\ 0 & 0 & 1, \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x+13 \\ y+17 \\ 1 \end{bmatrix}.$$

In our deformation model we describe the deformation in terms of displacement, so that

$$D_x = \mathbf{p_X}(1, x, y) = c_1 + c_2 x + c_3 y,$$
$$D_x = \mathbf{p_X}(1, x, y) = c_4 + c_5 x + c_6 y,$$

should represent the constant displacements of each pixel, i.e. an perfect registration would yield $D_x = 13$ and $D_y = 17$ for all (x, y), thus we should find the parameters

$$c_1 = 13, \quad c_4 = 17, \quad c_2 = c_3 = c + 5 = c_6 = 0.$$

In this case the registration algorithm returns the parameters

$$c_1 = 12.51, \quad c_4 = 15.12, \quad c_2, c_3, c_5, c_6 < .01$$

Similarly, for the translation and rotation, the ground truth transformation is the same translation, 13 left, 17 down, along with a rotation 10 degrees clockwise. So we have the matrix

$$\begin{bmatrix} \cos(-10°) & -\sin(-10°) & 13 \\ \sin(-10°) & \cos(-10°) & 17 \\ 0 & 0 & 1, \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(-10°)x - \sin(-10°)y + 13 \\ \sin(-10°)x + \cos(-10°)y - 17 \\ 1 \end{bmatrix},$$

which corresponds to the deformation parameters

$$D_x = c_1 + c_2 x + c_3 y = x(\cos(-10°) - 1) - y\sin(-10°) + 13$$
$$D_y = c_4 + c_5 x + c_6 y = x\sin(-10°) + y(\cos(-10°) - 1) + 17,$$

so that we have

$$
\begin{aligned}
c_1 &= 13, \\
c_2 &= \cos(-10^\circ) - 1 = -0.0151, \\
c_3 &= -\sin(-10^\circ) = 0.1736, \\
c_4 &= 17, \\
c_5 &= \sin(-10^\circ) = -0.1736, \\
c_6 &= (\cos(-10^\circ) - 1) = -0.0151.
\end{aligned}
$$

In this case the registration algorithm returns

$$
\begin{aligned}
c_1 &= 29.37, \\
c_2 &= 0.0086, \\
c_3 &= -0.1237 \\
c_4 &= 31.69, \\
c_5 &= 0.1020, \\
c_6 &= 0.0016.
\end{aligned}
$$

Here we see that we did not recover the ground truth transformation, and thus the global optimal solution determined by the convex approximation was not able to capture the ground truth transformation, though visually we observe the registration is qualitatively successful.

We may also measure the true Mean Squares Difference of the target image to the registered image to find the average difference in intensity value. In the pure translation we have

$$MS(A, \varphi(A)) = 0.0073,$$

and in the translation and rotation we have

$$MS(A, \varphi(A)) = 0.0063.$$

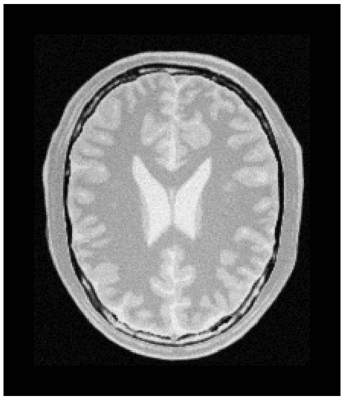
It is worth nothing that, because of it's dependence on the size of the window used to define the approximation, this algorithm struggles to find 'large' affine transformations with scaling or shearing. By construction Convex Approximation cannot detect both large shifts and local scaling, although some sort of iterated algorithm may be constructed to approach this problem, and we leave such an algorithm for future work.

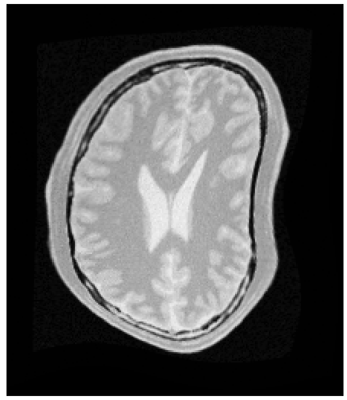
8.10.2 Gaussian Deformation Model

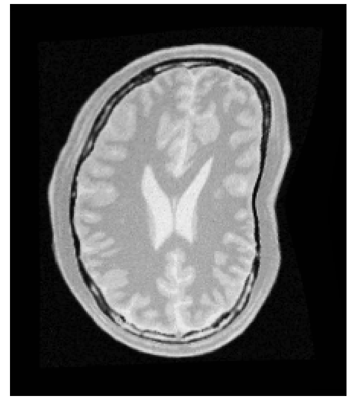
For local non-linear deformations the best results attained were those using the Gaussian Deformation model discussed in (8.5.2).

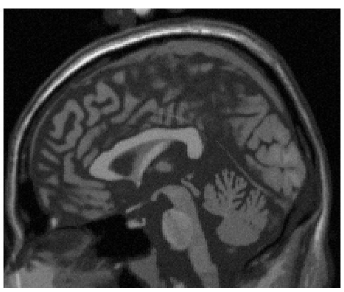
Figure (11) shows several examples of the registration algorithm deployed using the Gaussian deformation model. The first two examples are fairly standard deformations that could be seen in the context of medical imaging, whereas the latter two are more extreme deformations. In these more extreme cases it is clear that the algorithm tends to focus in on particular details and is unable to detect others well. In fact, in the first two examples there are still key details that the registration misses, though qualitatively the result is good.

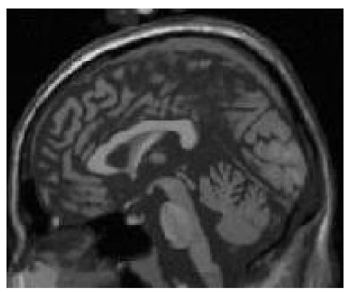
In each of the registrations show in this section the registration was performed using 3 scales with a starting window size between varied between 20 and 50 pixels, and $\alpha = 2/5$.

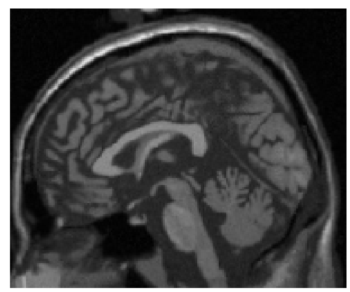

(a) Base


(b) Target

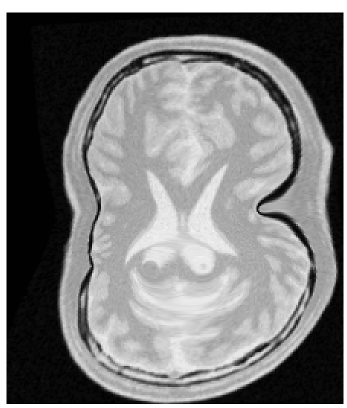

(c) Registered

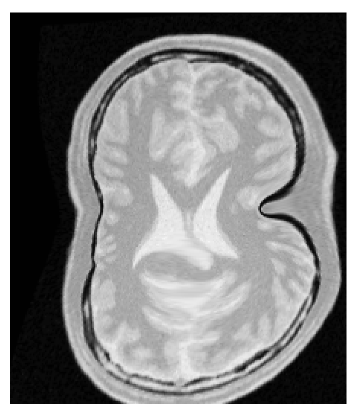

(a) Base


(b) Target


(c) Registered

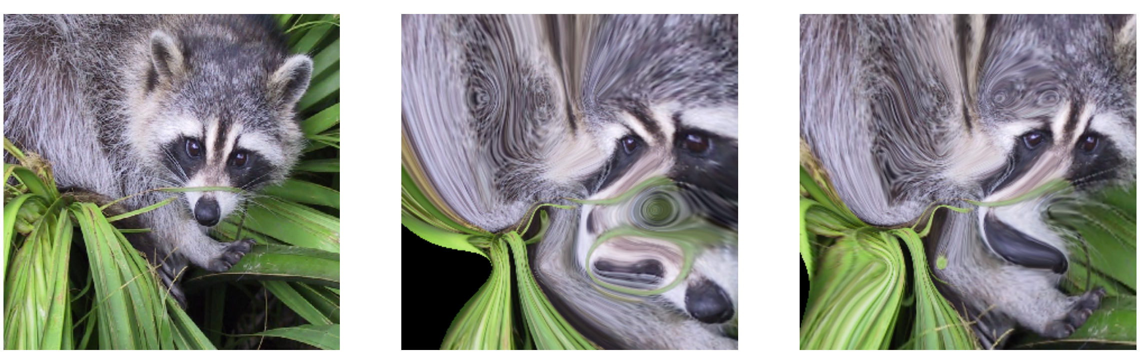

(a) Base


(b) Target


(c) Registered

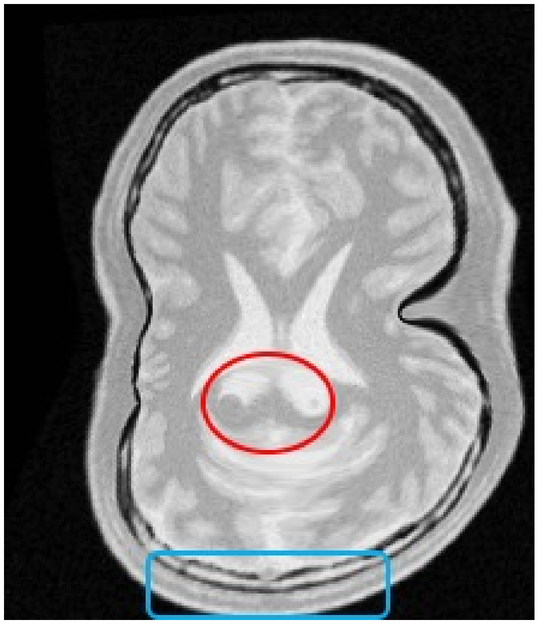
(a) Base (b) Target (c) Registered

(a) Base (b) Target (c) Registered

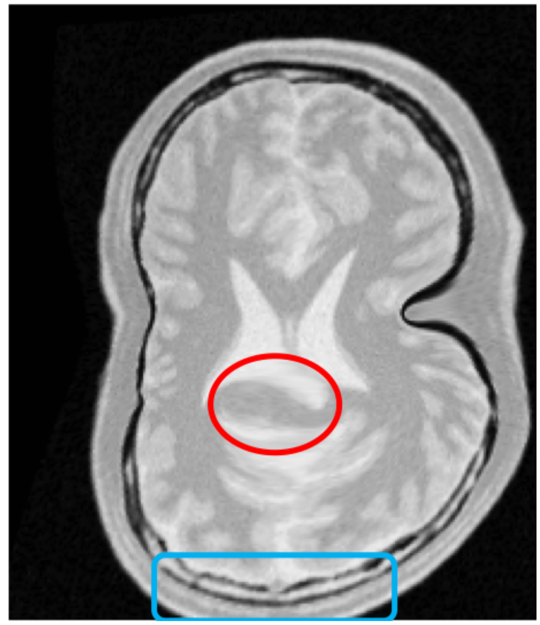


(a) Base (b) Target (c) Registered

Figure 11: Gaussian Deformation Examples

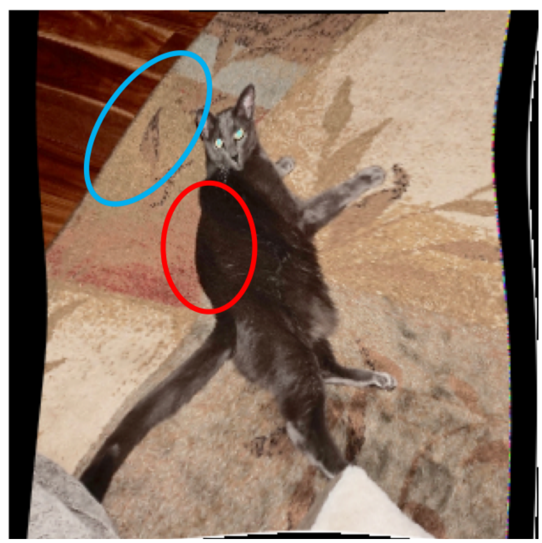


(a) Target



(b) Registered

(a) Target

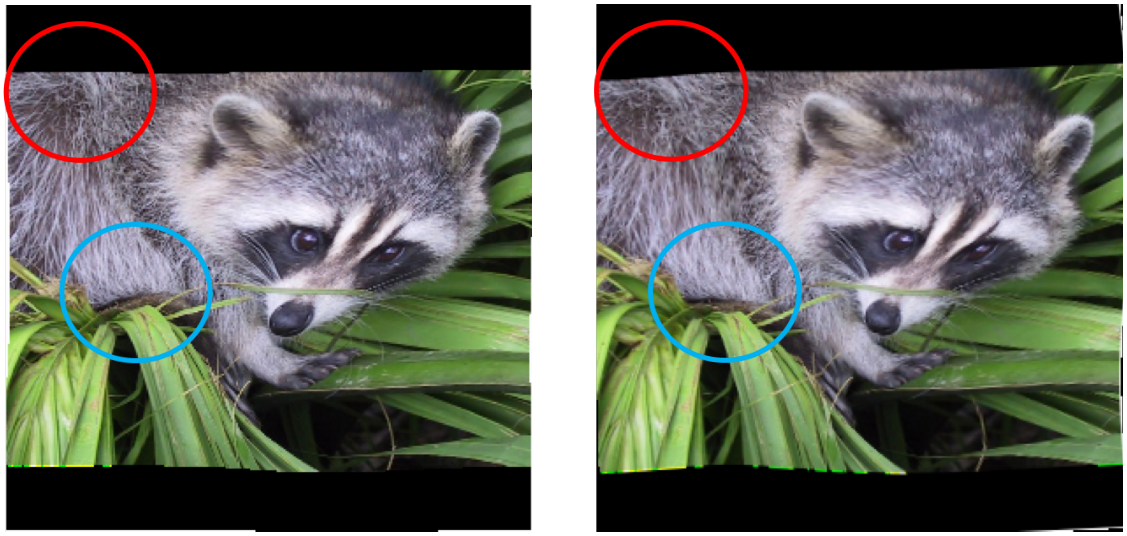


(b) Registered

8.10.3 Discussion of Registration, Future Work

To exemplify this, we look 3 examples from above and point out specific areas of mismatch between the target image and the registered image in Figure (12).

In solving the linear program, these areas of mismatch were considered to be of minimal cost associated relative to the convex approximation of the MS metric. The window size chosen to define each lower convex hull directly effects the shape of these hulls and therefore of the entire approximation. Thus, changing this starting window size, the α chosen to define how quickly this window shrinks, and the number of scales used in the registration algorithm will alter the approximation and emphasize



(a) Target (b) Registered

Figure 12: Registration Mismaches

different regions in the images. To quantify the impact of these parameters is left as a topic for further study. In particular, it would be interesting to quantify an optimal window size for successful registration of particular regions in the target image. To get better insight into this question, it may also be helpful to construct a quantitative analysis of registration success in subregions of the target image and the registered image.

In many applications of image registration there is noise introduced to both the target and the base image, which adds an additional layer of difficulty to the problem, in that what may be 'similar' images relative to the image metric could now be corrupted enough to disqualify these images from being registered. There are multi-scale algorithms designed for registration under the presence of noise in particular contexts [**paquin**]. The convex algorithm discussed here would be particularly sensitive to impulse, or salt and pepper noise, as large jumps in intensity value throughout the image will severely confound the convex approximation of the MS metric. More uniform noise, such as Gaussian noise, however, may be less destructive to the success of this algorithm.

We remark that the above construction can be extended readily to work with the Normalized Correlation Metric by first finding an appropriate intensity scaling between images and then performing the registration on the rescaled images. A more ambitious direction for further work would be to construct a analog of this work for the Mutual Information Metric [**miRegistrationSurvey**] in order to produce a convex multi-modal image registration scheme.

Appendix

All relevant (unoptimized) code used for the implementation of the convex image registration algorithm can be found in my personal Github repository. The core packages used in the this implementation were `scipy`'s ConvexHull module and `SimpleITK`'s image registration modules. `PIL` and `cvxpy` were also used throughout the implementation process.