Nonlinear Optimization

Jianing Yao
Department of MSIS-RUTCOR
Rutgers University, the State University of New Jersey
Piscataway, NJ 08854 USA

March 18, 2015

Abstract

This is the lecture notes when I took class with Prof. Andrezj Ruszczyński. The textbook we used is "nonlinear optimization" by himself, which is a quite comprehensive book. It integrates the theory and algorithm so that readers can really understand how those knowledge interacts. Although only finite-dimensional space is concerned, most of the proof and result can be extended to infinite-dimensional space easily. We covered the analysis of the smooth optimization problem for the restriction of time. The notes is used for preparing the qualify exam.

Contents

1 Introduction **5**

2 Elements of Convex Analysis **6**

2.1 Convex Sets 6

2.2 Projection 8

2.3 Separation 10

2.4 Cones and Separation of Cones 12

2.4.1 Cones and examples 12

2.4.2 Separation of cones 17

2.5 Extreme Point 21

2.6 Convex Function 22

2.7 Smooth convex functions 25

3 Optimality Conditions **28**

3.1 Unconstrained Optimization 28

3.2 Constrained Optimization 29

3.2.1 Tangent Cones 30

3.2.2 Metric Regularity 33

3.2.3 Algebraic Form 36
3.3 Optimality condition 39

4 Duality **42**

4.1 Intro to Lagrangian Duality 42
4.2 Simple Properties 44
4.3 Duality Relation 45
4.4 Application to Decomposition Problem 48
4.5 Augmented Lagrangian 49
4.5.1 Simplified System 49
4.5.2 Complete System 52

5 Unconstrained Optimization **54**

5.1 Line Search 55
5.2 Steepest Descent 57
5.2.1 Illustration of the method 57
5.2.2 Rate of Convergence 59
5.3 Conjugate Gradient Method 63
5.3.1 Illustration of the method 63
5.3.2 Rate of convergence 68
5.3.3 Pre-conditioning 68

6 Constrained Optimization **71**

6.1 The reduced Gradient Method 71

6.2 Penalty Method 75
6.2.1 general idea 75
6.2.2 Quadratic Penalty 77
6.3 Dual Method 78
6.4 Augmented Lagrangian Method 80

Chapter 1

Introduction

All optimization problems can be characterized in the following simple form: $\mathcal{X}$ is a set and $f : X \mapsto \mathbb{R}$ is a real function (in general, we can have $\mathbb{C}$ instead of $\mathbb{R}$), the problem is to find $\hat{x} \in \mathcal{X}$ such that

$$f(\hat{x}) \leq f(x), \text{ for all } x \in \mathcal{X}$$

Usually, $\mathcal{X}$ is called *feasible set* and f is called *objective function*. In this level of generality, very little can be said about optimization problems. We are interested in problems in which $\mathcal{X}$ is a subset of *Euclidean space* $\mathbb{R}^n$, and function f is sufficiently regular, e.g, convex, differentiable. Often times, $\mathcal{X}$ can be realized as a finite system of equalities and inequalities, which is always called *constrains*. As one of the most popular optimization problem, linear programming has objective function linear and the feasible set $\mathcal{X}$ is defined by a finite system of linear inequalities and equalities.

If the objective function f or some of the equations or inequalities defining the feasible set $\mathcal{X}$ is of non-linear form, then the optimization problem is called *non-linear optimization problem*. In this case, those specific tools for linear programming, simplex, dual simplex, e.t.c., can not be applied, we need novel technique to address such problems. This is the goal of this class. Before we get into the main business, let's first look at another characterization of optimization problems, the one that we will use,

$$\begin{aligned} \min . \ & f(x) \\ \text{s.t. } & g_i(x) \leq 0, \ i = 1, ..., m; \\ & h_i(x) = 0, \ i = 1, ..., p; \\ & x \in \mathcal{X}_0 \end{aligned}$$

where $g_i, h_j : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., m$, $j = 1, ..., p$ and $\mathcal{X}_0$ are some special set that is easy to deal with, e.g., $\mathcal{X}_0 = \{x : x \geq 0\}$.

Chapter 2

Elements of Convex Analysis

2.1 Convex Sets

The notion of *convexity of a set* is significant, especially for optimization problems. It basically says that if two points are in a set, then its line segment joining them lies entirely in the set as well. Let's give the formal definition:

Definition 2.1.1 A set $\mathcal{X} \subset \mathbb{R}^n$ is called *convex* if for all $x^1, x^2 \in \mathcal{X}$ it contains all points

$$\alpha x^1 + (1-\alpha)x^2,\ 0 < \alpha < 1$$

As one can imagine, convexity is preserved by the operation of intersection, namely,

Lemma 2.1.1 Let I be an arbitrary index set. If the set $\mathcal{X}_i \subset \mathbb{R}^n$, $i \in I$, are convex, then the set

$$\mathcal{X} = \cap_{i \in I} \mathcal{X}_i$$

is convex.

If we define the set operations as follows:

$$c\mathcal{X} := \{y \in \mathbb{R}^n : y = cx, x \in \mathcal{X}\},\ c \text{ is a scalar;}$$
$$\mathcal{X} + \mathcal{Y} := \{z \in \mathbb{R}^n : z = x + y, x \in \mathcal{X}, y \in \mathcal{Y}\}$$

the linear combination (in the same sense as for $\mathbb{R}^n$) preserves the convexity:

Lemma 2.1.2 Let $\mathcal{X}$ and $\mathcal{Y}$ be convex sets in $\mathbb{R}^n$ and let $c, d \in \mathbb{R}$. Then the set $\mathcal{Z} = c\mathcal{X} + d\mathcal{Y}$ is convex.

As we notice that the convex combination of $x^1, x^2 \in \mathcal{X}$,i.e., $\alpha x^1 + (1-\alpha)x^2$, for $\alpha \in (0,1)$ falls between the line segment joining x^1 and x^2, if we join more points in a 'similar' way, by *convex combination*,

Definition 2.1.2 A point x is a convex combination of points $x^1, ..., x^m$, if there exists $\alpha_1 \geq 0$,..., $\alpha_m \geq 0$ such that

$$x = \alpha_1 x^1 + \alpha_2 x^2 + \cdots + \alpha_m x^m$$

where

$$\sum_{i=1}^{m} \alpha_i = 1$$

then we construct so called *convex hull*. The usual characterization of convex hull is the following:

Definition 2.1.3 The convex hull of the set $\mathcal{X}$ denoted by conv$\mathcal{X}$ is the intersection of all convex sets containing $\mathcal{X}$

The relationship between them is provided by the following lemma

Lemma 2.1.3 The set conv$\mathcal{X}$ is the set of all convex combinations of points of $\mathcal{X}$.

Actually, we can represent arbitrary point in conv$\mathcal{X}$ by a convex combination of certain number of points.

Theorem 2.1.4 (*Caratheodory's lemma*) If $\mathcal{X} \subset \mathbb{R}^n$, then every element of conv$\mathcal{X}$ is a convex combination of at most $n+1$ points of $\mathcal{X}$.

Proof. Let x be a convex combination of $m > n+1$ points of $\mathcal{X}$. We shall show that m can be reduced by 1. If $\alpha_j = 0$ for some j, then we can delete the jth point and we are done. So, let all $\alpha_i > 0$. Since $m > n+1$, one can find $\gamma_1, \gamma_2, ..., \gamma_m$, not all equal 0 so that

$$\gamma_1 \begin{bmatrix} x^1 \\ 1 \end{bmatrix} + \gamma_2 \begin{bmatrix} x^2 \\ 1 \end{bmatrix} + \cdots + \gamma_m \begin{bmatrix} x^m \\ 1 \end{bmatrix} = 0 \tag{2.1}$$

Because number of equations is less than number of variables. Let $\tau = \min\{\frac{\alpha_i}{\gamma_i} : \gamma_i > 0\}$. Note that τ is well-defined, because some γ_j must be bigger than 0, if their sum is 0. Let $\bar{\alpha}_i = \alpha_i - \tau\gamma_i$, $i = 1, 2, ..., m$. By (2.1) we still have $\sum_{i=1}^m \bar{\alpha}_i = 1$ and $\sum_{i=1}^m \bar{\alpha}_i x^i = x$. By the definition of τ, at least one $\bar{\alpha}_j = 0$ and we can delete the jth point. Continuing in this way, we can reduce the number of points to $n+1$. □

Let's end this section with some basic relations of convexity and topological properties of a set.

Lemma 2.1.5 If $\mathcal{X}$ is convex, then its interior int$\mathcal{X}$ and its closure $\bar{\mathcal{X}}$ are convex.

Proof. Let B denote the unit ball. If $x^1, x^2 \in \text{int}\mathcal{X}$, then one can find $\epsilon > 0$ such that $x^1 + \epsilon B, x^2 + \epsilon B \subset \mathcal{X}$. Thus $\alpha x^1 + (1-\alpha)x^2 + \epsilon B \subset \mathcal{X}$ for $0 < \alpha < 1$. Therefore, $\alpha x^1 + (1-\alpha)x^2 \subset \text{int}\mathcal{X}$. To prove the second part of the lemma, let $x^k \to x$ and $y^k \to y$ with $x^k, y^k \in \mathcal{X}$ for all k. Then the sequence of points $\alpha x^k + (1-\alpha)y^k$ is contained in $\mathcal{X}$ is contained in $\mathcal{X}$ and converges to $\alpha x + (1-\alpha)y \in \mathcal{X}$. □

Lemma 2.1.6 * Assume that the set $\mathcal{X} \subset \mathbb{R}^n$ is convex. Then $\text{int}\mathcal{X} = \emptyset$ if and only if $\mathcal{X}$ is contained in a linear manifold of dimension smaller than n.

Proof. The proof is in the book, not that important ! □

2.2 Projection

For a closed and convex set $V \subset \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$, we call the point in V that is closest to x the *projection* of x on V and we denote it by $\Pi_V(x)$. The following theorem says the projection is always well-defined.

Theorem 2.2.1 If the set $V \subset \mathbb{R}^n$ is non-empty, convex and closed, then for every $x \in \mathbb{R}^n$ there exists exactly one point $z \in V$ that is closest to x.

Proof. Let $\mu = \inf\{||z - x|| : z \in V\}$. Since V is non-empty, μ is finite. Let us consider a sequence of points $\{z^k\}_{k\geq 1} \in V$ such that $||z^k - x|| \to \mu$, as $k \to \infty$. The sequence is bounded, it must have a convergent subsequence (*Bolzano Weierstrass theorem*), $\{z^k\}$, $k \in \mathcal{K}$. Denote the limit of this subsequence by z. We have

$$||z - x|| = \lim_{k\to\infty; k\in\mathcal{K}} ||z^k - x|| = \mu$$

Since V is closed, $z \in V$. This proves the existence.

Suppose two different points $z^1, z^2 \in V$ have distance μ from x. Consider the point $z = \frac{z^1+z^2}{2}$, it belongs to V, by convexity. Its distance to x can be calculated by *Pythagorean theorem*:

$$||z - x||^2 = \mu^2 - \frac{1}{4}||z^1 - z^2|| < \mu^2$$

a contradiction. Thus the projection is unique. □

Another characterization of projection is given by following lemma:

Lemma 2.2.2 Assume that $V \subset \mathbb{R}^n$ is a closed convex set and let $x \in \mathbb{R}^n$. Then $z = \Pi_V(x)$ if and only if $z \in V$ and

$$\langle v - z, x - z\rangle \leq 0, \text{ for all } v \in V \tag{2.2}$$

Proof. Let $z = \Pi_V(x)$ and $v \in V$, consider points of the form (in *Figure* 2.1)

$$\omega(\alpha) = \alpha v + (1-\alpha)z,\ 0 \leq \alpha \leq 1$$

Suppose $\langle x - z, v - z\rangle > 0$ for some $v \in V$, then

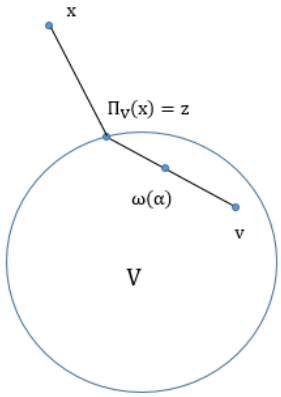

Figure 2.1: Projection

$$\begin{aligned} ||\omega(\alpha) - x||^2 &= \langle z + \alpha(v-z) - x, z + \alpha(v-z) - x\rangle \\ &= ||z-x||^2 + 2\alpha\langle z - x, v - z\rangle + \alpha^2||v-z||^2 \end{aligned}$$

The linear term is non-positive. However, as $\alpha \to 0$, the whole expression will be less than $||z-x||^2$ which can not happen, because, by convexity, the distance between $\omega(\alpha)$ and x cannot be smaller than $||z-x||$. Contradiction! Thus, we have $\langle v - z, x - z\rangle \leq 0$ for all $v \in V$.

For the other way around, pick an arbitrary $v \in V$,

$$\begin{aligned} ||v-x||^2 &= ||z + v - z - x||^2 \\ &= ||z-x||^2 + ||v-z||^2 + 2\langle z - x, v - z\rangle \end{aligned}$$

This implies $||v-x||^2 \geq ||z-x||^2$. Thus, z is the minimizer, i.e., $z = \Pi_V(x)$. □

Remark 2.2.3 In particular, if the set V is a linear manifold (vector subspace), for every $v \in V$ we have

$$\omega = 2\Pi_V(x) - v \in V$$

as well. Therefore,

$$\begin{aligned} \langle v - \Pi_V(x), x - \Pi_V(x)\rangle \leq 0 \\ \langle \omega - \Pi_V(x), x - \Pi_V(x)\rangle \leq 0 \end{aligned}$$

Consequently, $x - \Pi_V(x) \perp V$.

Another theorem that maybe less important is that the projection operator is non-expansive:

Theorem 2.2.4 Assume that $V \subset \mathbb{R}^n$ is a closed convex set. Then for all $x, y \in \mathbb{R}^n$ we have

$$||\Pi_V(x) - \Pi_V(y)|| \leq ||x - y||.$$

2.3 Separation

Separation theorem is quite intuitive but has unbelievable significance to the proof of many theorems. We will learn several versions of separation theorem. Here comes the first one:

Theorem 2.3.1 Let $\mathcal{X} \subset \mathbb{R}^n$ be a closed convex set and let $x \notin \mathcal{X}$. Then there exists a non-zero $y \in \mathbb{R}^n$ and $\epsilon > 0$ such that

$$\langle y, v \rangle \leq \langle y, x \rangle - \epsilon, \ \forall x \in \mathcal{X}$$

Proof. Let z be a point of $\mathcal{X}$ which is the closest to x. It exists since $\mathcal{X}$ is closed. By *lemma* 2.2.2 above,

$$\langle x - z, v - z \rangle \leq 0 \text{ for all } v \in \mathcal{X}$$

Define $y = x - z$. Note that y is non-zero by assumption. For each $v \in \mathcal{X}$ the last inequality implies

$$\begin{aligned} \langle y, v \rangle &= \langle y, z \rangle + \langle y, v - z \rangle \\ &\leq \langle y, z \rangle \\ &= \langle y, z - x \rangle + \langle y, z \rangle \leq \langle y, z \rangle \end{aligned}$$

and the residual is $\epsilon = ||y||^2$. □

If the set $\mathcal{X}$ is not closed, the point $x \notin \mathcal{X}$ maybe a boundary point, we have following weak result:

Theorem 2.3.2 Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set and let $x \notin \mathcal{X}$. Then there exists a non-zero $y \in \mathbb{R}^n$ such that

$$\langle y, v \rangle \leq \langle y, x \rangle, \text{ for all } v \in \mathcal{X}$$

The most interesting case will be x on the boundary of $\mathcal{X}$, as in *figure* 2.2

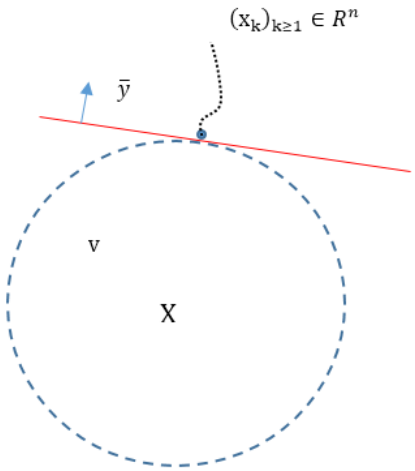

Figure 2.2: Separation

Proof. Consider a sequence $\{x_k\}_{k\geq 1}$ converge to x, where $x_k \in \bar{\mathcal{X}}$ for all k. By above theorem, we can always find $y_k \neq 0$ such that

$$\langle y_k, v\rangle \leq \langle y_k, x_k\rangle, \ \forall v \in \mathcal{X}$$

Normalize both sides by $\frac{1}{||y_k||}$,

$$\langle \frac{y_k}{||y_k||}, v\rangle \leq \langle \frac{y_k}{||y_k||}, x_k\rangle, \ \forall v \in \mathcal{X} \tag{2.3}$$

If we set $\bar{y}_k = \frac{y_k}{||y_k||}$, then $||\bar{y}_k|| = 1$. So we have a bounded sequence $\{\bar{y}_k\}_{k\geq 1}$, it must have an accumulation point thus a subsequence converging to $\bar{y}$. Passing limit on both sides of (2.3), we have

$$\langle \bar{y}, x\rangle \geq \langle \bar{y}, v\rangle, \ \forall v \in \mathcal{X}$$

□

The separation is possible also when both objects being separated are convex sets.

Theorem 2.3.3 Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be convex sets in $\mathbb{R}^n$. If $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$, then there exists a non-zero $y \in \mathbb{R}^n$ such that

$$\langle y, x_1\rangle \leq \langle y, x_2\rangle, \ \forall x_1, x_2 \in \mathcal{X}$$

Proof. Define $\mathcal{X} = \mathcal{X}_1 - \mathcal{X}_2$. Since $x = 0 \notin \mathcal{X}$. We can use *theorem* 2.3.2 to find $y \neq 0$ separating 0 from $\mathcal{X}$. This yields,

$$\langle y, 0\rangle \geq \langle y, v\rangle, \ \forall v \in \mathcal{X}$$

Since any $v \in \mathcal{X}$ can be decompose to x_1 and x_2 by definition, then

$$\langle y, x_2 \rangle \leq \langle y, x_1 \rangle$$

□

Above separation follows from *theorem* 2.3.2, actually, if one of the set is bounded, then strict separation is available.

Theorem 2.3.4 Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be closed convex sets in $\mathbb{R}^n$ and let $\mathcal{X}_1$ be bounded. If $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$, then $\exists y \in \mathbb{R}^n$ (not zero) and $\epsilon > 0$ such that

$$\langle y, x_1 \rangle \leq \langle y, x_2 \rangle - \epsilon$$

for all $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$.

The proof follows immediately from *theorem* 2.3.1 by the same argument in *theorem* 2.3.3

Remark 2.3.5 The boundedness assumption is essential. The closed sets $\mathcal{X}_1 = \{x \in \mathbb{R}^2 : x_2 \leq 0\}$ and $\mathcal{X}_2 = \{x \in \mathbb{R}^2 : x_2 \geq e^{-x_1}\}$ can not be separated strictly.

2.4 Cones and Separation of Cones

2.4.1 Cones and examples

A set $K \subset \mathbb{R}^n$ is called a *cone* if for every $x \in K$ and all $\alpha > 0$, one has $\alpha x \in K$. A *convex cone* is cone that is also convex. In this case, the convex combination can be relaxed to any combination with positive weights:

Lemma 2.4.1 Let K be a convex cone. If $x_1, x_2, ..., x_m \in K$ and $\alpha_1, \alpha_2, ..., \alpha_m > 0$, then

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m \in K$$

Proof. By convexity,

$$\frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m}{\alpha_1 + \alpha_2 + \cdots + \alpha_m} \in K$$

As K is a cone, we can multiply the point on the left hand side by $\alpha_1 + \alpha_2 + \cdots + \alpha_m$ and stay in K. □

Let's consider some important examples of convex cones.

Lemma 2.4.2 Assume that $\mathcal{X}$ is a convex set, then the set

$$\text{cone}(\mathcal{X}) = \{\gamma x : x \in \mathcal{X}, \gamma > 0\}$$

is a convex cone. $\mathcal{X}$.

The proof is uninteresting, we can visualize what happens in *figure* 2.3: The set cone($\mathcal{X}$)

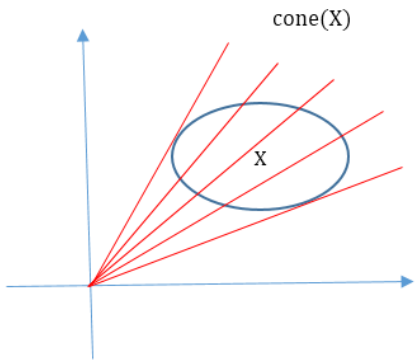

Figure 2.3: Cone generated by $\mathcal{X}$

is called the *cone generated by the set* $\mathcal{X}$. For a convex set $\mathcal{X}$ and a point $x \in \mathcal{X}$, the set

$$K_{\mathcal{X}}(x) := \text{cone}(\mathcal{X} - x) \tag{2.4}$$

is called the *cone of feasible directions of* $\mathcal{X}$ *at* x, which is of course a convex cone (see in *figure* 2.4). Here comes a very important example:

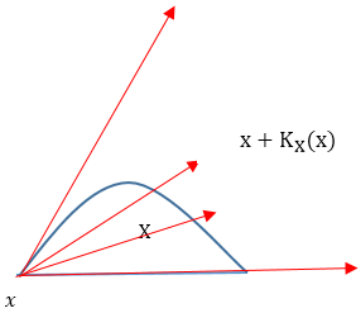

Figure 2.4: The cone of feasible dierctions (translated to the point x)

Example 2.4.3 Assume that set $\mathcal{X} \subset \mathbb{R}^n$ is a closed convex cone itself and $x \in \mathcal{X}$. Let us calculate the cone of feasible directions for $\mathcal{X}$ at x,

$$\begin{aligned} K_{\mathcal{X}}(x) &= \{d \in \mathbb{R}^n : d = \tau(y - x), y \in \mathcal{X}, \tau \geq 0\} \\ &= \{d \in \mathbb{R}^n : d = h - \tau x, h \in \mathcal{X}, \tau \geq 0\} \\ &= \mathcal{X} - \{\tau x : \tau \geq 0\} \\ &= \mathcal{X} + \{\tau x : \tau \in \mathbb{R}\} \end{aligned}$$

Let's also give the definition of *recession cone*:

Definition 2.4.1 Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex set and $x \in \mathcal{X}$ if there is a vector (representing a direction) d such that $x + \alpha d \in \mathcal{X}$ for all $\alpha \geq 0$, then d is called a *recession vector* of $\mathcal{X}$.

We can visualize it in *figure* 2.5 And it is obvious that if d is a recession vector from a

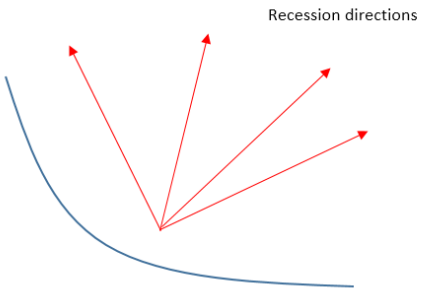

Figure 2.5: Recession vector of convex set $\mathcal{X}$

point $x \in \mathcal{X}$, it is a recession vector from any point $\bar{x} \in \mathcal{X}$.

Theorem 2.4.4 The set of all recession vectors of a convex set $\mathcal{X}$ forms a convex cone (we call it *recession cone* of $\mathcal{X}$, denote as $\text{rec}(\mathcal{X})$ or X_∞).

Proof. If $d \in \text{rec}(\mathcal{X})$, then for any $x \in \mathcal{X}$, $x + \alpha d$, $\forall \alpha \geq 0$. This implies $x + \lambda\alpha d \in \mathcal{X}$ for some $\lambda > 0$. Thus, $\text{rec}(\mathcal{X})$ is a cone. To show its convexity, let d_1, d_2 be recession directions, then

$$x + (\lambda d_1 + (1 - \lambda)d_2) = \lambda(x + d_1) + (1 - \alpha)(x + d_2) \in \mathcal{X}$$

for any $x \in \mathcal{X}$ and $\alpha \in [0, 1]$. □

A consequence of this is if $d \in \text{rec}(\mathcal{X})$ then $\mathcal{X} + d \subset \mathcal{X}$.

Next, we want to introduce *polar cone*:

Definition 2.4.2 Let K be a cone in $\mathbb{R}^n$. The set

$$K^\circ := \{y \in \mathbb{R}^n : \langle y, x\rangle \leq 0, \text{ for all } x \in K\}$$

is called the *polar cone of* K.

The picture is *figure* 2.6 There is an algebraic result follows from the definition. Let

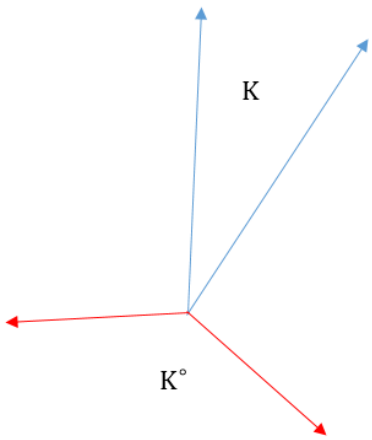

Figure 2.6: Polar Cone of K

$K_1, ..., K_m$ be cones in $\mathbb{R}^n$ and let $K = K_1 + K_2 + \cdots + K_m$. Clearly, K is a cone. We shall calculate its polar cone. If $z \in K^\circ$, then for every $x_1 \in K_1$, ..., $x_m \in K_m$, we have

$$\langle z, x_1\rangle + \cdots + \langle z, x_m\rangle \leq 0. \tag{2.5}$$

Let us choose $j \in \{1, ..., n\}$. Setting all $x_i = 0$ in (2.5), except for $i = j$, we conclude that

$$\langle z, x_j\rangle \leq 0 \text{ for all } x_j \in K_j$$

Consequently, $z \in K_j^\circ$. As j was arbitrary,

$$K^\circ \subset K_1^\circ \cup \cdots \cup K_m^\circ$$

On the other hand, for every element of z of $K_1^\circ \cup \cdots \cup K_m^\circ$ inequality (2.5) is satisfied, and thus $z \in K^\circ$. Therefore,

$$(K_1 + \cdots + K_m)^\circ = K_1^\circ \cup \cdots \cup K_m^\circ.$$

Following properties also follow from the definition immediately: 1) K° is a cone; 2)K° is convex; 3)K° is closed. The last property is true since K° is the intersection of closed half-space. We also have the useful property below:

Lemma 2.4.5 Let K be a cone in $\mathbb{R}^n$ and let $y \in \mathbb{R}^n$ be such that the scalar product $\langle y, x\rangle$ is bounded from above for all $x \in K$, then $y \in K^\circ$.

Proof. (Let's prove the contrapositive) Suppose $y \notin K^\circ$, then there exists $x \in K$

$$\langle y, x\rangle > 0$$

Take $\alpha > 0$, we know that $\langle y, \alpha x\rangle = \alpha\langle y, x\rangle$. As $\alpha \to \infty$, $\langle y, \alpha x\rangle \to \infty$. So it is not bounded above, which contradicts our assumption. □

The notation $K^{\circ\circ}$ stands for the polar cone of the polar cone of K. We have *Bipolar Theorem*:

Theorem 2.4.6 If $K \subset \mathbb{R}^n$ is closed convex cone, then

$$K^{\circ\circ} = K.$$

The proof is an application of separation theorem, we will skip.

In linear programming, *Farkas Lemma* is a famous axiom of alternative choice. We can be prove it via the notion of polar cone. But first, let's work on a more general result and then show *Farkas lemma* is a special case.

Theorem 2.4.7 Assume that C is a closed convex cone in $\mathbb{R}^m$ and A is an $m \times n$ matrix. Let

$$K = \{x \in \mathbb{R}^n : Ax \in C\}$$

Then K is closed convex cone and

$$K^\circ = \mathrm{cl}(\{A^\top \lambda : \lambda \in C^\circ\})$$

Proof. By construction K is convex and closed. If $z = A^\top\lambda$ with $\lambda \in C^\circ$, then for every $x \in K$ we have

$$\langle x, z\rangle = \langle x, A^\top\lambda\rangle = \langle Ax, \lambda\rangle \leq 0$$

Because $Ax \in C$. Hence

$$\mathrm{cl}(\{A^\top \lambda : \lambda \in C^\circ\}) \subset K^\circ$$

The set on the left hand side is convex and closed (easy to verify). Suppose $h \in K^\circ$. Then we can strictly separate h and set in question: there exists $y \neq 0$ and $\epsilon > 0$ such that

$$\langle y, h\rangle \geq \langle y, A^\top\lambda\rangle + \epsilon, \text{ for all } \lambda \in C^\circ \tag{2.6}$$

Thus, the product $\langle y, A^\top\lambda\rangle = \langle Ay, \lambda\rangle$ is bounded above for all $\lambda \in C^\circ$. Thus, $Ay \in C^{\circ\circ}$. Since C is closed, $Ay \in C$ and therefore $y \in K$. On the other hand, $h \in K^\circ$, we have $\langle y, h\rangle \leq 0$. But then (2.6) with $\lambda = 0$ yields a contradiction: $0 \geq \epsilon$. □

Setting

$$C = \{y \in \mathbb{R}^m : y_i \leq 0, i = 1, ..., m\}$$

We obtain the *Farkas Lemma*:

Corollary 2.4.8 Let A be an $m \times n$ matrix and let

$$K = \{x \in \mathbb{R}^n : Ax \leq 0\}.$$

Then

$$K^\circ = \{y \in \mathbb{R}^n : y = A^\top \lambda, \lambda \in \mathbb{R}^m, \lambda \geq 0\}.$$

The above fact is frequently formulate as an alternative: exactly one of the following two systems has a solution, *either* (i) $Ax \leq 0$ and $\langle c, x\rangle > 0$; *or* (ii) $c = A^\top \lambda, \lambda \geq 0$.

Now the normal cone.

Definition 2.4.3 Consider a convex closed set $\mathcal{X} \in \mathbb{R}^n$ and a point $x \in \mathcal{X}$. The set

$$N_{\mathcal{X}}(x) := [\text{cone}(\mathcal{X} - x)]^\circ$$

is called the *normal cone to* $\mathcal{X}$ *at* x.

Obviously, the normal cone is a polar thus closed and convex. Also, it follows from the definition: $v \in N_{\mathcal{X}}(x)$ if and only if

$$\langle v, y - x\rangle \leq 0 \text{ for all } y \in \mathcal{X}$$

Therefore, we can have another characterization of normal cone:

Lemma 2.4.9 Let X be a closed convex set and $x \in \mathcal{X}$. Then

$$N_{\mathcal{X}}(x) = \{v \in \mathbb{R}^n : \Pi_{\mathcal{X}}(x + v) = x\}$$

2.4.2 Separation of cones

We have learnt separation theorem, in this section, we will see how this can be applied to cones, especially convex cones, since they arise so frequently in optimization theory.

Let K_1 and K_2 be convex cones such that $K_1 \cap K_2 = \emptyset$. By separation theorem, we have $y \neq 0$,

$$\langle y, x^1\rangle \leq \langle y, x^2\rangle$$

for all $x^i \in K_i$, $i = 1, 2$. By *lemma* 2.4.5, we have $y \in K_1^\circ$ and $-y \in K_2^\circ$. Thus, if convex cone $K_1 \cap K_2 = \emptyset$, there exists $y^1 \in K_1^\circ$ and $y^2 \in K_2^\circ$ such that

$$y^1 + y^2 = 0$$

As a generalization, we have the following theorem:

Theorem 2.4.10 Let $K_1, K_2, ..., K_m$ be convex cones in $\mathbb{R}^n$. If $K_1 \cap K_2 \cap ... \cap K_m = \emptyset$, then there exists $y^i \in K_i^\circ$, $i = 1, 2, ..., m$, not all equal 0, such that

$$y^1 + y^2 + \cdots + y^m = 0$$

Proof. Let us define two cones in $\mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n = \mathbb{R}^{mn}$:

$$C_1 = \{z = (z^1, ..., z^m) : z^i \in K_i, i = 1, ..., m\},$$
$$C_2 = \{\omega = (x, ..., x), x \in \mathbb{R}^n\}$$

Since $\cap_{i=1}^m K_i = \emptyset$, we have $C_1 \cap C_2 = \emptyset$. Then by separation theorem, we can find a non-zero vector $y \in \mathbb{R}^{mn}$ such that

$$\langle y, \omega \rangle \geq \langle y, z \rangle$$

for all $\omega = (x, x, ..., x)$ and all $z \in C_1$. Writing $y = (y^1, y^2, ..., y^m)$, we obtain:

$$\langle y^1 + y^2 + \cdots + y^m, x \rangle \geq \langle y^1, z^1 \rangle + \langle y^2, z^2 \rangle + \cdots + \langle y^m, z^m \rangle$$

for all $x \in \mathbb{R}^n$ and all $z^i \in K_i$, $i = 1, 2, ..., m$. Setting $x = 0$, we see that the right hand side is bounded from above for all $z^i \in K_i$, $i = 1, 2, ..., m$, which implies that each $\langle y^i, z^i \rangle$ is bounded from above for all $z^i \in K_i$. By *lemma* 2.4.5, we have $y^i \in K_i^\circ$. The left hand side is bounded from below for all $x \in \mathbb{R}^n$, which is possible only when $y^1 + y^2 + \cdots + y^m = 0$. □

To use above theorem, we need to be capable to derive the polar cones. The following theorem will be very useful, but the prerequisite is the lemma below:

Lemma 2.4.11 If $x \in \text{int}K$, then $\langle y, x \rangle < 0$ for all non-zero $y \in K^\circ$.

The visualization is easier than a proof (in *figure* 2.7)

Theorem 2.4.12 Let $K_1, .., K_m$ be convex cones in $\mathbb{R}^n$ and let $K = \cap_{i=1}^m K_i$. If $K_1 \cap \text{int}K_2 \cap \cdots \cap \text{int } K_m \neq \emptyset$, then

$$K^\circ = K_1^\circ + K_2^\circ + \cdots + K_m^\circ$$

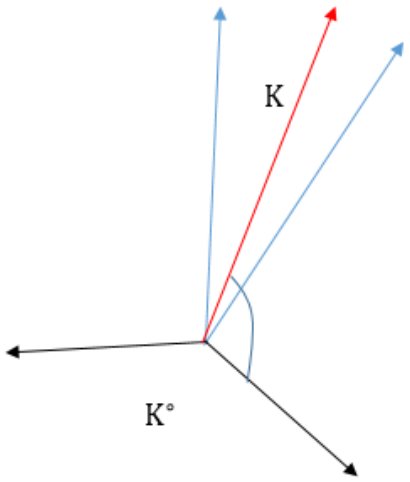

Figure 2.7: Visualization Proof

Proof. (" $\supseteq$ ") Take $x \in K$, $y = y^1 + \cdots + y^m$, where $y^i \in K_i^\circ$, for all i, then

$$\langle y, x\rangle = \langle y^1, x\rangle + \cdots + \langle y^m, x\rangle \leq 0$$

since $y^i \in K_i^\circ$.

(" $\subseteq$ ") Choose $y \in K^\circ$ and define the cone:

$$C = \{x : \langle x, y\rangle > 0\}$$

Clearly, $C \cap K_1 \cap K_2 \cap \cdots \cap K_m = \emptyset$. By separation theorem, we can find $d \in C^\circ$, $y^i \in K_i^\circ$, not all zero such that

$$d + y^1 + y^2 + \cdots + y^m = 0$$

Directly from the definition of C we see that $d = -\alpha y$ for some $\alpha > 0$. Thus, we can rewrite as:

$$-\alpha y + y^1 + \cdots + y^m = 0$$

If $\alpha \neq 0$, then

$$y = \frac{y^1}{\alpha} + \cdots + \frac{y_m}{\alpha}$$

where every $\frac{y^i}{\alpha} \in K_1^\circ$, $\forall i$.

If $\alpha = 0$, then

$$y^1 + \cdots + y^m = 0$$

Choose $\bar{x} \in K_1 \cap \text{int}K_2 \cap \cdots \cap \text{int}K_m$,

$$\langle y^1 + \cdots + y^m, \bar{x} \rangle = \sum_{i=1}^{m} \langle x, y^i \rangle = 0$$

All components in the summation are non-positive, thus $\langle x, y^i \rangle = 0$, $i = 1, ..., m$. Since $x \in \text{int}K_i$, $i = 2, ..., m$, *lemma* 2.4.11 implies $y^i = 0$, $i = 2, ..., m$. Then $y^1 = 0$, a contradiction. Thus $\alpha > 0$, which returns to the previous case. □

Remark 2.4.13 If the cones $K_1, .., K_m$ are polyhedral, the regularity assumption $K_1 \cap \text{int}K_2 \cap \cdots \cap \text{int } K_m \neq \emptyset$ is not needed; the result will follow directly from the *corollary* 2.4.8.

Another application of above theorem is the calculation of the polar cone of :

$$K = \{x \in K_1 : Ax \in K_2\}$$

which arises in a natura way in the analysis of optimality conditions.

Theorem 2.4.14 Assume that K_1 and K_2 are closed convex cones, and K is defined above. If

$$0 \in \text{int } \{Ax - y : x \in K_1, y \in K_2\}$$

then,

$$K^\circ = K_1^\circ + \text{cl}\left(\{A^\top \lambda : \lambda \in K_2^\circ\}\right)$$

Remark 2.4.15 Again, when all cones are polyhedral, the regularity assumption is not needed.

If we recall the notion – normal cones, we have another important theorem:

Theorem 2.4.16 Assume that $X = X_1 \cap \cdots \cap X_m$, where X_i are closed convex sets, $i = 1, ..., m$, and let $x \in X$. If $X_1 \cap \text{int } X_2 \cap \cdots \cap \text{int } X_m \neq \emptyset$, then

$$N_X(x) = N_{X_1}(x) + \cdots + N_{X_m}(x)$$

Proof. We have

$$K_X(x) = K_{X_1}(x) \cap \cdots \cap K_{X_m}(x)$$

By assumption,

$$K_{X_1}(x) \cap \text{int } K_{X_2}(x) \cdots \cap \text{int } K_{X_m}(x) \neq \emptyset$$

Applying the *theorem* 2.4.12, we obtain the desired result. □

2.5 Extreme Point

Let's first introduce the notion of extreme point

Definition 2.5.1 A point x of a convex set X is called an *extreme point* of X if no other points $x^1, x^2 \in X$ exists such that

$$x = \frac{1}{2}x^1 + \frac{1}{2}x^2$$

As we may remember that a convex hull of X is the set of all convex combination of points of X, but if the set X is closed, it can be characterized by just its extreme points.

Theorem 2.5.1 A convex and compact set in $\mathbb{R}^n$ is equal to the convex hull of the set of its extreme points.

The proof is a little bit long, we will skip. The above result can be extended to unbounded convex sets with the use of the concept of the *recession cone*, we will not give those proofs, but refer you to the book.

Lemma 2.5.2 A closed convex set X is bounded if and only if $X_\infty = \{0\}$.

If we denote by $E(X)$ the set of all extreme points of a set X, then

Theorem 2.5.3 A close convex set X, which has at least one extreme point, can be represented as follows:

$$X = \text{conv}E(X) + X_\infty$$

These results above are essential for *linear programming*:

Theorem 2.5.4 Let $A \in \mathbb{R}^{m\times n}$ and let $X \subset \mathbb{R}^n$ be defined as:

$$X = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$$

A point x is an extreme point of X if and only if the columns of A that correspond to positive component of x are linearly independent.

Those extreme points are called *basic feasible solution*, their role is evident by *theorem* 2.5.1:

Theorem 2.5.5 If the set $X = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ is bounded, it is the convex hull of the set of basic feasible solutions.

2.6 Convex Function

For every $f : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ (extended real), we can associate it with two sets:

$$\text{dom } f := \{x : f(x) < +\infty\}$$

and the *epigraph*

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

It is displayed in *figure* 2.8 Since always ∞ function are not of interests, we shall restrict

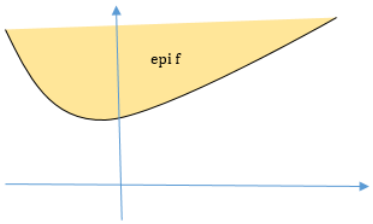

Figure 2.8: Epigraph of function f

our attention only to so called *proper function*: it satisfies that $f(x) > -\infty$ for all x and $f(x) < +\infty$ for at least one x. Then, we can define *convexity* of function via the epigraph:

Definition 2.6.1 A function f is called convex if epi f is a convex set. It is called concave if $-f$ is convex.

And there exists an equivalent characterization:

Lemma 2.6.1 A function f is convex if and only if for all x^1, x^2 and $0 \leq \alpha \leq 1$, we have

$$f(\alpha x^1 + (1-\alpha)x^2) \leq \alpha f(x^1) + (1-\alpha) f(x^2) \tag{2.7}$$

Proof. If $x^1 \notin$ dom f or $x^2 \notin$ dom f, the above inequality is satisfied trivially. It remains to consider the case when $x^1, x^2 \in$ dom f. Then the points $(x^1, f(x^1))$ and $(x^2, f(x^2))$ belong to epi f. If f is convex, then their convex combination is in the epigraph, i.e.,

$$(\alpha x^1 + (1-\alpha)x^2, \alpha f(x^1) + (1-\alpha) f(x^2)) \in \text{epi } f \tag{2.8}$$

which implies (2.7). On the other hand, (2.7) entails (2.8), which in turn yields the convexity of the epigraph. □

Definition 2.6.2 A function f is called *strictly convex* if inequality (2.7) is strict for all $x^1 \neq x^1$ and all $0 < \alpha < 1$.

Several simple observations about convex functions given below as lemmas.

Lemma 2.6.2 If f is convex then dom$f$ is a convex set.

Proof. If $x^1 \in$ dom f and $x^2 \in$ dom f, then, $f(\alpha x^1 + (1-\alpha)x^2) \leq \alpha f(x^1) + (1-\alpha)f(x^2) < +\infty$. □

Lemma 2.6.3 If f_i, $i \in I$, is a family of convex functions, then

$$f(x) = \sup_{i \in I} f_i(x)$$

is convex.

Proof. Note that epi $f = \cap_{i \in I}$epi f_i, we obtain the result from the fact that the intersection of convex sets are convex. □

Lemma 2.6.4 If f is a convex function, then for all $x^1, x^2, ..., x^m$ and all $\alpha_1 \geq 0$, $\alpha_2 \geq 0$, ..., $\alpha_m \geq 0$ such that $\alpha_1 + \alpha_2 + \cdots + \alpha_m = 1$, one has

$$f(\alpha_1 x^1 + \alpha_2 x^2 + \cdots + \alpha_m x^m) \leq \alpha_1 f(x^1) + \alpha_2 f(x^2) + \cdots + \alpha_m f(x^m).$$

Proof. The result follows from the convexity of epi f: te points $(x^i, f(x^i))$, $i = 1, ..., m$ belong to the epigraph, so their convex combination is in epi f, too. □

Lemma 2.6.5 If the functions f_i, $i = 1, ..., m$, are convex, then for all $c_i \geq 0$, $i = 1, ..., m$ the function:

$$f(x) = c_1 f_1(x) + \cdots + c_m f_m(x)$$

is convex.

Proof. Since (2.7) holds true for each f_i, we can multiply these inequalities by $c_i \geq 0$ and sum up to get (2.7) for f. □

We now define the *level set*:

$$M_\beta = \{x \in \mathbb{R}^n : f(x) \leq \beta\}$$

Lemma 2.6.6 If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex, then for each $\beta \in \mathbb{R}$, M_β is convex.

Proof. If $x, y \in M_\beta$, then

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \leq \beta$$

So $\alpha x + (1-\alpha)y \in M_\beta$. □

Let's see how convex function affect the optimization problem:

Lemma 2.6.7 Let $X \subset \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ be a convex function. Then the set $\tilde{X}$ of solutions of the optimization problem:

$$\min_{x \in X} f(x) \tag{2.9}$$

is convex.

Proof. If no solution exists, then $\tilde{X}$ is trivially convex. Suppose $\tilde{X}$ is non-empty, so we can choose $\tilde{x} \in \tilde{X}$, define $\beta = f(\tilde{x})$. Then

$$\tilde{X} = X \cap M_\beta$$

with M_β defined as level set. Thus, $\tilde{X}$ is convex. □

The maxima of convex functions in convex sets can be characterized as well:

Theorem 2.6.8 Let f be a convex function and let $X \subset \text{dom } f$ be a convex, closed and bounded set. Then the set of solutions of the problem:

$$\max_{x \in X} f(x) \tag{2.10}$$

contains at least one extreme point of X. If, in addition, the function $f(\cdot)$ is *affine*, then the set of solutions of (2.10) is the convex hull of the set of extreme points of X that are solutions of (2.10).

Proof. Let $\tilde{x}$ be a solution, since X is compact, we can find extreme points $x^1, x^2, ..., x^m$ of X such that

$$\tilde{x} = \alpha_1 x^1 + \cdots + \alpha_m x^m$$

where $\alpha_i > 0$, $\sum_{i=1}^m \alpha_i = 1$. Since f is convex,

$$f(\tilde{x}) \leq \sum_{i=1}^m \alpha_i f(x^i)$$

This can be rewritten as:

$$\sum_{i=1}^m \alpha_i [f(\tilde{x}) - f(x^i)] = 0 \tag{2.11}$$

Because $\tilde{x}$ is optimal, $f(\tilde{x}) \geq f(x^i)$, $i = 1, ..., m$. Then it follows from (2.11) that $f(\tilde{x}) = f(x^i)$, $i = 1, ..., m$. All points x^i are optimal as well. We see that the set of solutions of (2.10) is included in the convex hull of the extreme points that are solutions of (2.10).

If the function $f(\cdot)$ is affine, then $-f(\cdot)$ is convex. The set of the solution of it is the same as the set of minima of $-f(x)$ over $x \in X$, it is convex. Therefore, the convex hull of the extreme points that are soultion of it is included in the set of all solutions of it. □

Returning to linear programing problem,

Theorem 2.6.9 If the feasible set of the linear programming problem

$$\begin{aligned} &\min . \ \langle c, x\rangle \\ \text{subject to: } &Ax = b, \\ &x \geq 0 \end{aligned}$$

is bounded, then the set of optimal solutions is the convex hull of the set of optimal basic feasible solutions.

Remark 2.6.10 The above fact is used by the simplex method for solving linear programming problems. It moves from one basic feasible solution to a better one, as long as progress is possible. The best basic feasible solution is guaranteed to be optimal. It can be found after finitely many steps if a solution exists. If the set is unbounded, we may discover a ray from the recession cone, along which the objective can be decreased without limits. In this case no optimal solution exists.

2.7 Smooth convex functions

Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is *differentiable*, then we denote the *gradient* of f at x,

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} (x)$$

If f is twice continuously differentiable, $\nabla^2 f(x)$ denotes the *Hessian* of f at x,

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} (x)$$

Theorem 2.7.1 Assume that a function f is continuously differentiable, then

- f is convex if and only if for all x and y,

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

- f is strictly convex if and only if for all $x \neq y$,

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

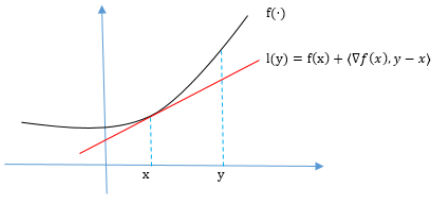

Figure 2.9: Convex function gradient

We can easily visualize it (in *figure* 2.9)

Proof. (i) Suppose f is convex, but for some x and y,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle - \epsilon \tag{2.12}$$

with $\epsilon > 0$. Let us consider $z = \alpha y + (1 - \alpha)x$, for $\alpha \in (0, 1)$. Then,

$$f(x) \leq \alpha f(y) + (1 - \alpha) f(x) \leq f(x) + \alpha \langle \nabla f(x), y - x \rangle - \alpha\epsilon$$

Rearrange the term and divide both sides by α, we have

$$\frac{f(z) - f(x)}{\alpha} \leq \langle \nabla f(x), y - x \rangle - \epsilon$$

Let $\alpha \downarrow 0$. Since $z = x + \alpha d$ with $d = y - x$, the left hand side converges to the directional derivative at x in the direction d,

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

We obtain a contradiction.

For the converse, we assume that (2.12) is true. Let y and z be arbitrary points, $y \neq z$, and let $x = \alpha y + (1 - \alpha)x$ with $\alpha \in (0, 1)$. Then,

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

Simple maniuplation yields,

$$\alpha f(y) + (1 - \alpha) f(z) \geq f(x)$$

which was what we set out to prove. (ii) follows similar reasoning, we will skip. □

The second derivative, hessian, also gives a characterization of convexity:

Theorem 2.7.2 Assume that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is twice continuously differentiable. Then

- f is convex if and ony if its hessian $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbb{R}^n$;

- if the Hessian $\nabla^2 f(x)$ is positive definite for all $x \in \mathbb{R}^n$ then f is strictly convex.

Proof. If f is twice differentiable , then for all x and y,

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\langle y - x, \nabla^2 f(x_\theta)(y - x) \rangle$$

where $x_\theta = x + \theta(y - x)$ with $\theta \in [0, 1]$. If the hessian $\nabla^2 f(x)$ is positive semi-definite for all x, then the quadratic term is non-negative, and we obtain what we want.

Suppose that hessian is not positive semi-definite for some x, then there exists d such that

$$\langle d, \nabla^2 f(x), d \rangle < 0$$

Let $y = x + \epsilon d$ for some $\epsilon > 0$. If ϵ is small enough, then y and x_θ are so close to x that

$$\langle d, \nabla^2 f(x_\theta) d \rangle < 0$$

Because hessian is continuous. But quadratic term is negative, we get contradiction. □

Remark 2.7.3 Note the second statement does not have the only if part, which appeared in the quadratic case. For example, the function $f(x) = x^4$ is stricly convex, but its second derivative vanishes at 0.

In the proof, we used the notion of directional derivative. The definition is the following:

Definition 2.7.1 Let $f : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ be a convex function and let $x \in \text{dom } f$. Then for every $d \in \mathbb{R}^n$, the quantity:

$$f'(x; d) = \lim_{\tau \downarrow 0} \frac{f(x + \tau d) - f(x)}{\tau}$$

is called the directional derivative of f at x in the direction d.

If the function f is differentiable at x, then the directional derivative exists along any vector d, and one has

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

Intuitively, the directional derivative of f at a point x represents teh rate of change of f with respect to time when it is moving at a speed and direction given by d.

Chapter 3

Optimality Conditions

3.1 Unconstrained Optimization

The unconstrained optimization problem can be described as below:

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

where we assume $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable. Then,

Theorem 3.1.1 If f is differentiable at $\tilde{x}$, then

- If $f(\cdot)$ attain its local minimum at $\tilde{x}$, then

$$\nabla f(\tilde{x}) = 0 \tag{3.1}$$

- If $f(\cdot)$ is convex and (3.1) is satisfied, then $\tilde{x}$ is the global minimum of $f(\cdot)$.

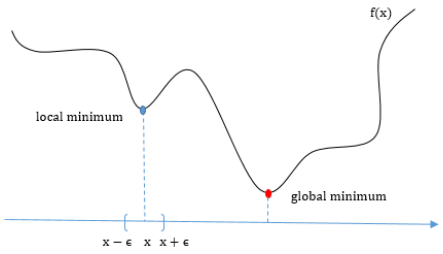

Figure 3.1: Local minimum and global minimum

Proof. Suppose $\nabla f(\tilde{x}) \neq 0$, we consider the points:

$$x(\tau) = \tilde{x} - \tau \nabla f(\tilde{x}) \tag{3.2}$$

From *taylor expansion*, we obtain

$$f(x(\tau)) = f(\tilde{x}) + \langle \nabla f(\tilde{x}), x(\tau) - \tilde{x} \rangle + o(x(\tau), \tilde{x})$$

where $\lim_{x(\tau)\to\tilde{x}} \frac{o(x(\tau);\tilde{x})}{||x(\tau)-\tilde{x}||} = 0$. Thus,

$$\begin{aligned} f(x(\tau)) :&= f(\tilde{x}) - \langle \nabla f(x), \tau \nabla f(\tilde{x}) \rangle + o(x(\tau); \tilde{x}) \\ &= f(\tilde{x}) - \tau \langle \nabla f(\tilde{x}), \nabla f(\tilde{x}) \rangle + o(x(\tau); \tilde{x}) \end{aligned}$$

Since $||x(\tau) - \tilde{x}|| = \tau ||\nabla f(\tilde{x})||$,

$$\begin{aligned} f(x(\tau)) &= f(\tilde{x}) - \tau ||\nabla f(\tilde{x})||^2 + \frac{o(x(\tau); \tilde{x})}{\tau ||\nabla f(\tilde{x})||} \tau ||\nabla f(\tilde{x})|| \\ &= f(\tilde{x}) + \tau (-||\nabla f(\tilde{x})||^2 + \frac{o(x(\tau); \tilde{x})}{\tau ||\nabla f(\tilde{x})||} ||\nabla f(\tilde{x})||) \\ &\neq f(\tilde{x}) \end{aligned}$$

for all sufficiently small τ. A better point $x(\tau)$ exists in any neighbourhood of $\tilde{x}$, therefore $\tilde{x}$ can not be a local minimum.

If $f(\cdot)$ is convex, then for all $y \in \mathbb{R}^n$, we have

$$f(y) \geq f(\tilde{x}) + \langle \nabla f(\tilde{x}), y - x \rangle = f(\tilde{x})$$

Therefore, $\tilde{x}$ is global minimum. □

Remark 3.1.2 Points satisfying condition (3.1) are called stationary.

3.2 Constrained Optimization

The constrained optimization is often written in the following format:

$$\min_{x \in X} \ f(x)$$

where we again assume $f(\cdot)$ is differentiable. In this case, $\tilde{x} \in X$ is a *global minimum* if $f(\tilde{x}) \leq f(x)$, for all $x \in X$; $x \in X$ is a *local minimum* if there exists an ϵ such that $f(\tilde{x}) \leq f(x)$ for all $x \in \mathbb{R}^n$ satisfying $||x - \tilde{x}|| \leq \epsilon$. To analyse such optimization problem, the notion of tangent cone is essential. It will in the end lead to the optimality condition. Actually, all optimality conditions, in one way or another, decipher the inclusion

$$-\nabla f(\tilde{x}) \in [T_X(\tilde{x})]^\circ$$

which we will see shortly. This is the main motivation for the consideration of this section.

3.2.1 Tangent Cones

Let's recall the notion of cone of feasible direction at x_0,

$$K_X(x_0) = \{d \in \mathbb{R}^n : \exists \tilde{\tau} > 0, \forall \tau \in (0, \tilde{\tau}], x_0 + \tau d \in X\}$$

or, equivalently,

$$K_X(x_0) = \{d \in \mathbb{R}^n : d = \beta(y - x), y \in X, \beta \geq 0\}$$

However,it is impossible to find feasible set of directions. For example, if the set X is a sphere, at any point on the surface of it, one can not perturb it even a bit along the tangent direction. Therefore, we introduce the concept of *tangent direction* at $x_0 \in X$.

Definition 3.2.1 A direction d is called *tangent* to the set $X \subset \mathbb{R}^n$ at the pint $x \in X$ if there exists a sequence of points $x^k \in X$ and scalars $\tau_k > 0$, $k = 1, 2, ...$, such that $\tau_k \downarrow 0$ and

$$d = \lim_{k \to \infty} \frac{x^k - x}{\tau_k}$$

Lemma 3.2.1 Let $x \subset \mathbb{R}^n$ and let $x_0 \in X$. The set $T_X(x_0)$ of all tangent directions for X at x_0 is a closed cone.

Proof. Suppose $d \in T_X(x_0)$. For every $\beta > 0$,

$$\beta d = \lim_{k \to \infty} \frac{x^k - x_0}{\tau_k / \beta}$$

So the sequence $\{x^k\}$ and $\{\tau_k/\beta\}$ satisfying the definition of tangent direction βd. Hence $T_X(x_0)$ is a cone.

Let direction d^j be tangent to X at x_0 with corresponding sequence $\{x_0^{j,k}\}$ and $\{\tau_{j,k}\}$, $k = 1, 2, ...$, satisfying the definition of tangent direction and let

$$\lim_{j \to \infty} d^j = d$$

Since direction d^j are tangent, for every j we can find $k(j)$ such that

$$||\frac{x_0^{j,k} - x_0}{\tau_{j,k(j)}} - d^j|| \leq ||d^j - d||$$

Therefore,

$$||\frac{x_0^{j,k} - x_0}{\tau_{j,k(j)}} - d|| \leq 2||d^j - d||$$

It follows that the sequence $\{x^{j,k(j)}\}$ and $\{\tau_{j,k(j)}\}$, $j = 1, 2, ...$ satisfying definition for direction d. As a consequence, the cone $T_X(x_0)$ is closed. □

Lemma 3.2.2 All $d \in K_X(x_0)$ belongs to the tangent cone at x_0, $\forall x_0 \in X$.

Proof. Suppose $d \in K_X(x_0)$, then $\forall x$ including x^k has the form: $x^k = x_0 + \tau_k d \in X$. As $\tau_k \downarrow 0$,

$$\lim_{k\to\infty} \frac{x^k - x_0}{\tau_k} = d$$

□

From this, we saw the close relationship between tangent cones and cone of feasible direction at particular point. And if the set X is convex, we can obtain more (the proof is skipped).

Lemma 3.2.3 Let $X \subset \mathbb{R}^n$ be a convex set and let $x_0 \in X$, then

$$T_X(x_0) = \text{cl } (K_X(x_0))$$

Now, recall the constrained optimization problem:

$$\min_{x\in X} \ f(x) \tag{3.3}$$

with a differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ and a set $X \subset \mathbb{R}^n$. If its solution $\hat{x}$ is a boundary point of the feasible set X, the necessary condition of optimality formulated in unconstrained problem does not have to be satisfied. The main reason is that the perturbations of the point $\hat{x}$ which take it out of the feasible set X are not allowed, and therefore they may correspond to a decrease of the objective function. In order to obtain necessary conditions of optimality, we restrict the set of possible perturbations to tangent direction at $\hat{x}$.

Theorem 3.2.4 Assume that $\hat{x}$ is a *local minimum* of problem (3.3) and that $f(\cdot)$ is differentiable at $\hat{x}$. Let $T_X(\hat{x})$ be the *tangent cone* to the set X at $\hat{x}$. Then,

$$-\nabla f(\hat{x}) \in [T_X(\hat{x})]^\circ \tag{3.4}$$

Proof. Suppose assertion is false:

$$-\nabla f(\hat{x}) \notin [T_X(\hat{x})]^\circ$$

This means there exists a direction $d \in T_X(\hat{x})$ such that

$$\langle \nabla f(\hat{x}), d\rangle < 0 \tag{3.5}$$

As d is tangent cone, there exists a sequence of points $x^k \in X$ converging to $\hat{x}$ and a sequence of scalar $\tau_k \downarrow 0$ such that

$$\lim_{k\to\infty} \frac{x^k - \hat{x}}{\tau_k} = d \tag{3.6}$$

Since $f(\cdot)$ is differentiable at $\hat{x}$,

$$f(x^k) - f(\hat{x}) = \langle \nabla f(\hat{x}), x^k - x \rangle + o(x^k; \hat{x})$$

where $\lim_{k\to\infty} \frac{o(x^k;\hat{x})}{||x^k - \hat{x}||} = 0$. Dividing both sides by τ_k, we obtain

$$\frac{f(x^k) - f(\hat{x})}{\tau_k} = \langle \nabla f(\hat{x}), d \rangle + \langle \nabla f(\hat{x}), \frac{x^k - \hat{x}}{\tau_k} - d \rangle + \frac{o(x^k; \hat{x})}{\tau_k}$$

By (3.5), we know $||d|| \neq 0$ and thus,

$$\lim_{k\to\infty} \frac{o(x^k; \hat{x})}{\tau_k} = \lim_{k\to\infty} \frac{o(x^k; \hat{x})||d||}{||x^k - \hat{x}||} = 0$$

By virtue of (3.5) and (3.6), we conclude that

$$\lim_{k\to\infty} \frac{f(x^k) - f(\hat{x})}{\tau_k} = \langle \nabla f(\hat{x}), d \rangle < 0$$

On the other hand, all points x^k are feasible and they approach $\hat{x}$. Since $\hat{x}$ is a local minimum, $f(x^k) \geq f(\hat{x})$ for all sufficiently large k. Hence

$$\liminf_{k\to\infty} \frac{f(x^k) - f(\hat{x})}{\tau_k} \geq 0$$

a contradiction ! Therefore, relation (3.4) is valid. □

Above is the necessary condition for both local minimum and global minimum. If more condition is imposed, then we can have sufficient condition for global minimum.

Theorem 3.2.5 If the function $f(\cdot)$ is convex, the set X is convex, and a point $\hat{x} \in X$ satisfies:

$$-\nabla f(\hat{x}) \in [T_X(\hat{x})]^\circ \tag{3.7}$$

then $\hat{x}$ is *global minimum.*

Proof. Since the set X is convex, for every $y \in X$, the direction

$$d = y - \hat{x}$$

is a tangent direction for X at $\hat{x}$. Thus, condition (3.7) implies

$$\langle \nabla f(\hat{x}), y - \hat{x} \rangle \geq 0$$

Also, since the function $f(\cdot)$ is convex,

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

Therefore,

$$f(y) \geq f(\hat{x})$$

for all $y \in X$, as required. □

As we observe, the tangent cone of X at a particular point $x \in X$ is of great interests. Usually, we encounter set X defined by an intersection:

$$X = X_1 \cap X_2 \cap \cdots \cap X_m.$$

For a point $x \in X$, we always have

$$T_X(x) \subset T_{X_1}(x) \cap T_{X_2}(x) \cap \cdots \cap T_m(x)$$

But an equality is not guaranteed. We will in the following see when the equality can be established.

3.2.2 Metric Regularity

Feasible sets of non-linear optimization problems are usually defined by systems of inequalities and equalities:

$$\begin{aligned} &g_i(x) \leq 0, \ i = 1, ..., m; \\ &h_i(x) = 0, \ i = 1, ..., p. \end{aligned}$$

with $g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., m$, $h_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., p$. Additionally, we may have abstract set constraints of the form $x \in X_0$.

Now, let's consider the following problem first:

$$\begin{aligned} &g(x) \in Y_0; \\ &x \in X_0 \end{aligned} \tag{3.8}$$

Here $g : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable, Y_0 is a closed convex set in $\mathbb{R}^m$ and X_0 is a closed convex set in $\mathbb{R}^n$. The *Jacobian* of g is defined as:

$$g'(x) := \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{pmatrix} (x)$$

We denote by X the set defined by (3.8),

$$X = \{x \in X_0 : g(x) \in Y_0\} \tag{3.9}$$

and consider a point $x_0 \in X$. For a direction d to be tangent to X at x_0, it is necessary $d \in T_{X_0}(x_0)$. Furthermore, when x_0 is perturbed in the direction d, then $g(x_0)$ is perturbed in the direction $g'(x_0)d$.Thus, it is also necessary that $g'(x_0)d \in T_{Y_0}(y_0)$. Consequently,

$$T_X(x_0) \subset \{d \in \mathbb{R}^n : d \in T_{X_0}(x_0), g'(x_0)d \in T_{Y_0}(g(x_0))\} \tag{3.10}$$

It becomes an equation, if the set X_0 and Y_0 are convex closed polyhedra, and if the function $g(\cdot)$ is affine. In general, we need metric regularity to ensure the equality relation. But let's first prove the inclusion (3.9) above:

Proof. Since $X \subset X_0$, $\forall d \in T_X(x_0)$, it will be an element of $T_X(x_0)$. On the other hand, by definition, there exists $x^k \in X$ and $\tau_k \downarrow 0$ such that

$$d = \lim_{k\to\infty} \frac{x^k - x_0}{\tau_k}$$

Set $y^k = g(x^k)$, $y_0 = g(x_0)$, we have

$$y^k = y_0 + g'(x_0)(x^k - x_0) + o(x^k; x_0)$$

this is equivalent to

$$\lim_{k\to\infty} \frac{y^k - y_0}{\tau_k} = g'(x_0)d$$

As $y^k \in Y_0$, we have $g'(x_0)d \in T_{Y_0}(y_0)$. □

Next, we shall talk about *metric regularity*. If x_0 is a solution of the system (3.8). The system is *metric-regular* if we can find $\epsilon > 0$ such that for every point $\tilde{x}$ with $||\tilde{x} - x_0|| \leq \epsilon$ and every $\tilde{u}$, $||\tilde{u}|| \leq \epsilon$, a point x_R exists such that

$$g(x_R) - \tilde{u} \in Y_0,\ x_R \in X_0$$

and

$$||x_R - \tilde{x}|| \leq C(d(\tilde{x}, X_0) + d\left(g(\tilde{x}) - \tilde{u}, Y_0\right))$$

where the left hand side is the distance between $\tilde{x}$ and corrected point x_R, the first term in the bracket of left hand side is the error in $x \in X_0$ and the second term is the error in $g(\tilde{x}) - \tilde{u} \in Y_0$. This says: **the system allows compensation of small perturbations in (3.8) with small adjustment of** x. If x_0 satisfies the condition: $x_0 \in X_0, g(x_0) \in Y_0$, a point that is not too far from x_0, call it $\tilde{x}$, exists and should not violate the perturbed system too much in the following sense, if x_R satisfies the perturbed system $x_R \in X_0, g(X_R) - u \in Y_0$, the distance between $\tilde{x}$ and x_R can not be too much but bounded by some constant times the violation of the perturbed system by $\tilde{x}$.

The concept of metric regularity of a system $g(x) = 0, x \in X_0$ is illustrated in *figure* 3.2 Let's give two example of not being metric regular, in *figure* 3.3 and 3.4: In the first example, there exists no such corrected point x_R satisfying the perturbed system. While in the second example, x_R can be very far from $\tilde{x}$ (not proportional to the violation of the two constraints of the perturbed system).

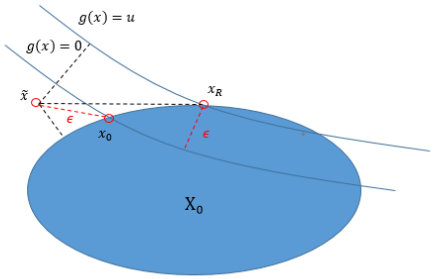

Figure 3.2: Illustration of metric regularity

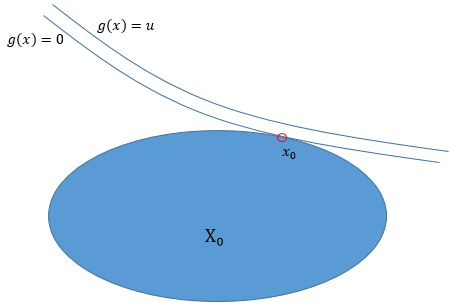

Figure 3.3: Counter Example 1

It can be proved that above description is equivalent to *Robinson condition*, that is,

$$0 \in \text{int}\{g'(x_0)d - v : d \in K_{X_0}(x_0), v \in K_{Y_0}(g(x_0))\}$$

Now, we can complete the reverse inclusion:

Theorem 3.2.6 If the system (3.8) is *metric regular*, then

$$T_X(x_0) = \{d \in \mathbb{R}^n : d \in T_{X_0}(x_0), g'(x_0)d \in T_{Y_0}(g(x_0))\} \tag{3.11}$$

Proof. We only need to proof the other way around inclusion. Let d be a direction in the right hand side of (3.11). Consider point of the form:

$$x(\tau) = x_0 + \tau d, \ \tau > 0$$

As $d \in T_{X_0}(x_0)$,

$$\text{dist } (x(\tau), X_0) = o_1(\tau) \tag{3.12}$$

with $\frac{o_1(\tau)}{\tau} \to 0$, when $\tau \downarrow 0$. Also,

$$g(x(\tau)) = g(x_0) + \tau g'(x_0)d + o_2(\tau)$$

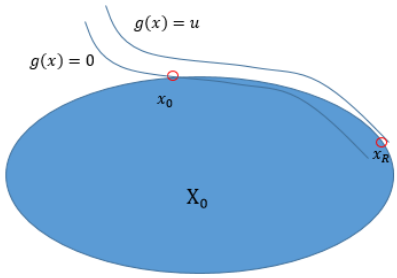

Figure 3.4: Counter Example 2

with $||o_2(\tau)||/\tau \to 0$, when $\tau \downarrow 0$. Since $g'(x_0)d \in T_{Y_0}(g(x_0))$, it follows that

$$\text{dist}(g(x(\tau)), Y_0) \leq \text{dist}(g(x(\tau)), g(x_0) + \tau g'(x_0)d) + \text{dist}(g(x_0) + \tau g'(x_0)d, Y_0) = o_3(\tau) \tag{3.13}$$

with $\frac{o_3(\tau)}{\tau} \to 0$, with $\tau \downarrow 0$. Consequently, the point $x(\tau)$ "almost" belongs to X_0 and almost satisfies the constraints $g(x_0) \in Y_0$. The error is negligible with respect to τ. We now use the property of metric regularity. Set $\tilde{x} = x(\tau)$ and $\tilde{u} = 0$, we deduce for sufficiently small τ, we can find $x_R(\tau) \in X$ such that

$$||x_R(\tau) - x(\tau)|| \leq C(\text{dist } (x(\tau), X_0) + \text{dist } (g(x(\tau)), Y_0))$$

Using (3.12)-(3.13), we conclude

$$\lim_{\tau \downarrow 0} \frac{x_R(\tau) - x(\tau)}{\tau} = 0$$

Equivalently,

$$\lim_{\tau \downarrow 0} \frac{x_R(\tau) - x_0}{\tau} - d = 0$$

Thus,

$$\lim_{\tau \downarrow 0} \frac{x_R(\tau) - x_0}{\tau} = d$$

Hence, d is a tangent direction to X at x_0. □

3.2.3 Algebraic Form

Now, let's consider the follow system:

$$\begin{aligned} &g_i(x) \leq 0, \ i = 1, ..., m; \\ &h_i(x) = 0, \ i = 1, ..., p \\ &x \in X_0 \end{aligned} \tag{3.14}$$

with continuously differentiable $g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $h_i : \mathbb{R}^n \mapsto \mathbb{R}$ and X_0 a convex closed set.

Suppose $\hat{x}$ is local minimum, we visualize

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \\ h_1(x) \\ \vdots \\ h_p(x) \end{bmatrix}_{(m+p)\times 1}$$

and

$$Y_0 = \{ \begin{bmatrix} y_1 \\ \vdots \\ y_m \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(m+p)\times 1} : y_i \leq 0, i = 1, ..., m\}$$

The system (3.11) can be rewritten as

$$g(x) \in Y_0, \ x \in X_0 \tag{3.15}$$

Then, for any $s = [s_1 \ \cdots \ s_m \ s_{m+1} \ \cdots \ s_{m+p}]^\top \in T_{Y_0}(g(\hat{x}))$, it must have the form that $s_{m+1} = \cdots = s_{m+p} = 0$. But what can we say about s_1 to s_m ? Let's define *active constraints*:

$$I^0(\hat{x}) = \{1 \leq i \leq m : g_i(\hat{x}) = 0\}$$

Then,

$$T_{X_0}(g(\hat{x})) = \{s \in \mathbb{R}^{m+p}; s_i \leq 0, i \in I^0(\hat{x}), s_i \in \mathbb{R}, i \notin I^0, \text{ for } i = 1, ..., m; s_i = 0, i = m+1, ..., m+p\}$$

Example 3.2.7 In $\mathbb{R}^3$, we consider

$$Y_0 = \{ \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} : x_1 \leq 0, x_2 \leq 0\}$$

As we observed in *figure* 3.5, $T_{Y_0}(g(\tilde{x}))$ must have x_3 direction 0. Now we only look at x_1, x_2 coordinates. If $g(\hat{x}) = [0, \ < 0]$, then obviously the tangent cone of Y_0 at $g(\hat{x})$ will have $x_1 \leq 0$, $x_2 \in \mathbb{R}$ (illustrated in *Figure* 3.6).

As a result, by

$$T_X(\hat{x}) = \{d \in \mathbb{R}^n; d \in T_{X_0}(\hat{x}), g'(\hat{x})d \in T_{Y_0}(g(\hat{x}))\}$$

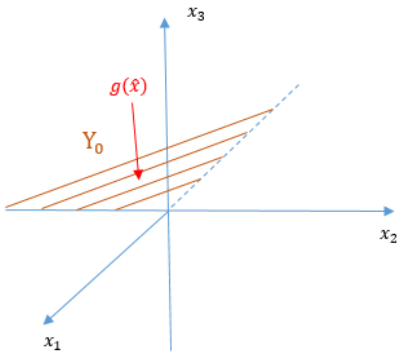

Figure 3.5: Visualization 1

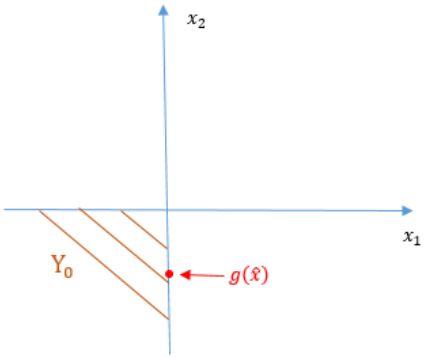

Figure 3.6: Visualization 2

We have

$$\begin{gathered}\langle \nabla g_i(\hat{x}), d\rangle \leq 0,\ \ i \in I^0(\hat{x}),\\ \langle \nabla h_i(\hat{x}), d\rangle = 0,\ \ i = 1, ..., p,\\ d \in T_{X_0}(\hat{x}).\end{gathered}$$

Or,

$$T_X(\hat{x}) = T_{X_0}(\hat{x}) \cap \{d : \langle \nabla g_i(\hat{x}), d\rangle \leq 0, i \in I^0(\hat{x}); \langle \nabla h_i(\hat{x}), d\rangle = 0, i = 1, ..., p\} \tag{3.16}$$

Define

$$A = \begin{bmatrix} (\nabla g_i(\hat{x}))^\top,\ i \in I^0(\hat{x}) \\ (\nabla h_i(\hat{x}))^\top,\ i = 1, ..., p \end{bmatrix}$$

then the second term on the right hand side of (refeq:inter2) can be expressed as:

$$\{d : Ad \in K\}$$

where K is the set $K = \{k_1 \leq 0, \text{ if } i \in I^0(\hat{x}); k_1 = 0, \text{otherwise}\}$. Since *Robinson condition* is satisfied, the assumption of *theorem* 2.4.14 is valid, thus,

$$[T_X(\hat{x})]^\circ = [T_{X_0}(\hat{x})]^\circ + \{A^\top \lambda : \lambda_i \geq 0,\ i \in I^0(\hat{x});\ \lambda_k \in \mathbb{R},\ \text{otherwise}\}$$

Set

- λ_i, $i \in I^0(\hat{x})$: coefficients corresponding to $\nabla g_i(\hat{x})$, $i \in I^0(\hat{x})$, those λ_i's are bigger than 0;
- $\mu_i, i = 1, ..., p$: coefficients corresponding to $\nabla h_i(\hat{x})$, $i = 1, ..., p$, those μ_i's are unrestricted, i.e., $\mu_i \in \mathbb{R}$, $i = 1, ..., p$.

Then

$$[T_X(\hat{x})]^\circ = N_{X_0}(\hat{x}) + \{\sum_{i \in I^0(\hat{x})} \lambda_i \nabla g_i(\hat{x}) + \sum_{i=1}^{p} \mu_i \nabla h_i(\hat{x}),\ \lambda_i \geq 0, i \in I^0(\hat{x}),\ \mu_i \in \mathbb{R}, i = 1, ..., p\}$$

3.3 Optimality condition

After all those preparations, we can now formally describe the optimality conditions for problem:

$$\begin{aligned} \min .\ & f(x) \\ \text{subject to: } & g_i(x) \leq 0,\ i = 1, ..., m; \\ & h_i(x) = 0,\ i = 1, ..., p; \\ & x \in X_0 \end{aligned} \tag{3.17}$$

where f, g_i, h_i are continuously differentiable and that the set X_0 is closed and convex, we denote the feasible set of this problem as X.

Theorem 3.3.1 Let $\hat{x}$ be a local minimum of problem (3.17). Assume that at $\hat{x}$ the constraint qualification condition is satisfied. Then, there exists multipliers $\hat{\lambda}_i \geq 0$, $i = 1, .., m$ and $\mu_i \in \mathbb{R}$, $i = 1, ..., p$ such that

$$0 \in \nabla f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) + N_{X_0}(\hat{x}) \tag{3.18}$$

and

$$\hat{\lambda}_i g_i(\hat{x}) = 0,\ i = 1, ..., m. \tag{3.19}$$

Remark 3.3.2 Actually, we shall have

$$0 \in \nabla f(\hat{x}) + \sum_{i \in I^0(\hat{x})} \hat{\lambda}_i \nabla g_i(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) + N_{X_0}(\hat{x})$$

But, we can assign $\lambda_i = 0$, $i \in I^0(\hat{x})$, thus obtaining (3.18) and (3.19). This set of conditions are called *Karush–Kuhn–Tucker* condition.

Example 3.3.3 Given $f(x) = x^\top Ax = \langle x, Ax \rangle$, A is $n \times n$ symmetric, we have constraints $h(x) = x^\top x = 1 = 0$. Then

$$\nabla h(x) = 2x$$
$$\nabla f(x) = 2Ax$$

By KKT condition, we shall have

$$0 = 2Ax + 2\mu x$$

That is equivalent to: $Ax = -\mu x$. Given $\langle x, Ax \rangle = -\mu ||x||^2 = -\mu$, to minimize $\langle x, Ax \rangle$ is to choose the smallest eigenvalue of A.

Let's introduce the *Lagrangian* associated with constrained problem:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$$

Then, (3.18) says

$$-\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) \in N_{X_0}(\hat{x}) \tag{3.20}$$

We have in *theorem* 3.3.1 the necessary condition for optimality, now we give the sufficient condition but under some regularity conditions of f, g, h:

Theorem 3.3.4 Assume that the function $f(\cdot)$ and $g_i(\cdot)$, $i = 1, ..., m$ are convex and the function $h_i(\cdot)$, $i = 1, .., p$ is affine. If the point $\hat{x} \in X$ and multiplier $\hat{\lambda}_i \geq 0$, $i = 1, ..., m$, and $\mu_i \in \mathbb{R}$, $i = 1, ..., p$, satisfies condition (3.18) and (3.19), then $\hat{x}$ is *global minimum.*

Proof. By assumption $L(x, \hat{\lambda}, \hat{\mu})$ is convex with respect to x. By (3.20) and *theorem* **??**, we have

$$L(\hat{x}, \hat{\lambda}, \hat{\mu}) \leq L(x, \hat{\lambda}, \hat{\mu}), \ \forall x \in X_0$$

At feasible point, we have

$$L(x, \hat{\lambda}, \hat{\mu}) \leq f(x)$$

and at the point $\hat{x}$,

$$f(\hat{x}) = L(\hat{x}, \hat{\lambda}, \hat{\mu})$$

Hence,

$$f(\hat{x}) \leq f(x) \text{ for all } x \in X$$

□

Chapter 4

Duality

4.1 Intro to Lagrangian Duality

Recall the general non-linear optimization problem:

$$\begin{aligned} \min . \ & f(x) \\ \text{subject to: } & g_i(x) \leq 0, \ i = 1, ..., m; \\ & h_i(x) = 0, \ i = 1, ..., p; \\ & x \in X_0 \end{aligned} \tag{4.1}$$

Suppose we make no assumption on the function $g(\cdot)$, $h(\cdot)$, $f(\cdot)$ and also the structure of the set X_0.

Let $\hat{x}$ be a local minimum of (4.1) and assume at $\hat{x}$ the constraint qualification condition is satisfied, then there exists $\hat{\lambda}_i \geq 0$, $i = 1, ..., m$ and $\mu_i \in \mathbb{R}$, $i = 1, ..., p$, such that

$$0 \in \nabla f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) + N_{X_0}(\hat{x}) \tag{4.2}$$

and complementarity condition

$$\hat{\lambda}_i g_i(\hat{x}) = 0, \ i = 1, ..., m \tag{4.3}$$

should be satisfied.

Define the *Lagrangian function* as before:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x) \tag{4.4}$$

It is a function of both $x \in X_0$ and $(\lambda, \mu) \in \Lambda_0$, where

$$\Lambda_0 = \mathbb{R}^m_+ + \mathbb{R}^p$$

then, its quite obvious that (4.2) can be rewritten in a compact form:

$$-\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) \in N_{X_0}(\hat{x}) \tag{4.5}$$

This is a necessary condition of optimality.

Remark 4.1.1 This also the necessary condition of optimality for the following optimization problem:

$$\min_{x \in X_0} L(x, \hat{\lambda}, \hat{\mu})$$

Let's define the so called *primal function*:

$$L_P(x) = \max_{(\lambda,\mu)\in\Lambda_0} L(x, \lambda, \mu) \tag{4.6}$$

then,

$$L_P(x) = \begin{cases} f(x), & \text{if } g_i(x) \leq 0,\ i = 1, ..., m,\ h_i(x) = 0,\ i = 1, ..., p; \\ +\infty, & \text{otherwise.} \end{cases}$$

Because if one of the constraint is violated, we can make $L_P(\cdot)$ arbitrary large. Let's also define the *dual function*:

$$L_D(\lambda, \mu) = \min_{x \in X_0} L(x, \lambda, \mu) \tag{4.7}$$

The *primal problem* is then

$$\min_{x \in X_0} L_P(x) \tag{4.8}$$

which is exactly equivalent to (4.1) and can be explicitly written as:

$$\min_{x \in X_0} \max_{(\lambda,\mu)\in\Lambda_0} L(x, \lambda, \mu) \tag{4.9}$$

On the other hand, we define the *dual problem*:

$$\max_{(\lambda,\mu)\in\Lambda_0} L_D(\lambda, \mu) \tag{4.10}$$

which is equivalent to:

$$\max_{(\lambda,\mu)\in\Lambda_0} \min_{x \in X_0} L(x, \lambda, \mu) \tag{4.11}$$

Now, the natural question to ask is whether they are equivalent, or, are there any conditions under which the equivalence relationship can be established, i.e.,

$$\min_{x \in X_0} \max_{(\lambda,\mu)\in\Lambda_0} L(x, \lambda, \mu) \stackrel{?}{=} \max_{(\lambda,\mu)\in\Lambda_0} \min_{x \in X_0} L(x, \lambda, \mu)$$

4.2 Simple Properties

Let's observe some simple properties of dual or primal functions:

Lemma 4.2.1 The dual function $L_D(\lambda, \mu)$ is concave.

Proof. Since the *Lagrangian* is affine in (λ, μ) for every $x \in X_0$, dual function is an infinimum of affine functions. Thus, $-L_D(\cdot)$ is a supremum of a family of affine function and is convex. □

Lemma 4.2.2 Assume that for (λ^0, μ^0) we can find $x^0 \in X_0$ such that

$$L_D(\lambda^0, \mu^0) = L(x^0, \lambda^0, \mu^0)$$

Then for all (λ, μ) we have

$$L_D(\lambda, \mu) \leq L_D(\lambda^0, \mu^0) + \langle g(x^0), \lambda - \lambda^0 \rangle + \langle h(x^0), \mu - \mu^0 \rangle \tag{4.12}$$

Proof. By the definition of dual function, $\forall (\lambda, \mu) \in \Lambda_0$,

$$\begin{aligned} L_D(\lambda, \mu) &= \min_{x \in X_0} L(x, \lambda, \mu) \\ &\leq L(x^0, \lambda, \mu) \\ &= L(x^0, \lambda^0, \mu^0) + \langle g(x^0), \lambda - \lambda^0 \rangle + \langle h(x^0), \mu - \mu^0 \rangle \end{aligned}$$

which is what we set out to prove. □

Lemma 4.2.3 If $L_D(\cdot, \cdot)$ is differentiable, i.e.,

$$\begin{aligned} \nabla_\lambda L_D(\lambda^0, \mu^0) &= g(x^0); \\ \nabla_\mu L_D(\lambda^0, \mu^0) &= h(x^0), \end{aligned}$$

then

$$L_P(\bar{x}) \geq L_D(\bar{\lambda}, \bar{\mu})$$

for all $\bar{x} \in X_0$ and $(\bar{\lambda}, \bar{\mu}) \in \Lambda_0$.

Proof.

$$L_P(\bar{x}) = \max_{(\lambda, \mu) \in \Lambda_0} L(\bar{x}, \lambda, \mu) \geq L(\bar{x}, \bar{\lambda}, \bar{\mu}) \geq \min_{x \in X_0} L(x, \bar{\lambda}, \bar{\mu}) = L_D(\bar{\lambda}, \bar{\mu})$$

□

4.3 Duality Relation

Definition 4.3.1 A point $(\tilde{x}, (\tilde{\lambda}, \tilde{\mu})) \in X_0 \times \Lambda_0$ is a *saddle point* of $L(x, \lambda, \mu)$ if

$$L(\tilde{x}, \lambda, \mu) \leq L(\tilde{x}, \tilde{\lambda}) \leq L(x, \tilde{\lambda}, \tilde{\mu}) \tag{4.13}$$

for all $x \in X_0$ and $(\lambda, \mu) \in \Lambda_0$.

Remark 4.3.1 A saddle point is such a point at which the maximum of *Lagrangian* w.r.t $(\lambda, \mu) \in \Lambda_0$ and minimum w.r.t $x \in X_0$ are attained:

$$\max_{(\lambda,\mu)\in\Lambda_0} L(\tilde{x}, \lambda, \mu) = L(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = \min_{x\in X_0} L(x, \tilde{\lambda}, \tilde{\mu})$$

We can answer the question raised in the first section by the following theorem:

Theorem 4.3.2 Assume that the function $f(\cdot)$ and $g_i(\cdot)$, $i = 1, ..., m$ in problem (4.1) are convex, the function $h_i(\cdot)$, $i = 1, ..., p$, are affine, and the set X_0 is convex. Then a point $\hat{x}$ satisfies the first order optimality condition with *Lagrangian multipliers* $(\hat{\lambda}, \hat{\mu})$ if and only if $(\hat{x}, (\hat{\lambda}, \hat{\mu}))$ is a *saddle point* of *Lagrangian.*

Proof. (*We only give a proof when the all functions are differentiable, but the non-smooth case can be easily adapted*)1. Optimality condition $\Rightarrow$ saddle point: assume that $\hat{x}$ is an optimal solution satisfying the optimality condition with multipliers $(\hat{\lambda}, \hat{\mu})$. For fixed values of $(\hat{\lambda}, \hat{\mu}) \in \Lambda_0$, the *Lagrangian* is a convex function of x. If we differentiate both sides with respect to x and evaluated at $\hat{x}$,

$$\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) = \nabla f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x})$$

According to the first order optimality condition, we have

$$0 \in \nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) + N_{X_0}(\hat{x})$$

It follows this is also the optimality condition for the *Lagrangian function* $L(x, \hat{\lambda}, \hat{\mu})$ for $x \in X_0$, i.e.,

$$L(\hat{x}, \hat{\lambda}, \hat{\mu}) = \min_{x\in X_0} L(x, \hat{\lambda}, \hat{\mu})$$

Thus, the right inequality of (4.13) holds true for $x \in X_0$. We shall prove the left inequality. By the complementarity condition, we have

$$L(\hat{x}, \hat{\lambda}, \hat{\mu}) = f(\hat{x})$$

Since $g_i(\hat{x}) \leq 0$, $i = 1, ..., m$, and $h_i(\hat{x}) = 0$, $i = 1, ..., p$, for every $(\lambda, \mu) \in \Lambda_0$ we have

$$L(\hat{x}, \lambda, \mu) = f(\hat{x}) + \sum_{i=1}^{m} \lambda g_i(\hat{x}) + \sum_{i=1}^{p} \mu_i h_i(\hat{x}) \leq f(\hat{x})$$

Combining the last two relations we obtain the left inequality of (4.13) for all $(\lambda, \mu) \in \Lambda_0$. Therefore, $(\hat{x}, (\hat{\lambda}, \hat{\mu}))$ is a saddle point of the *Lagrangian*.

2. Optimality condition $\Leftarrow$ saddle point: suppose $(\hat{x}, (\hat{\lambda}, \hat{\mu}))$ is a saddle point, since:

$$L(\hat{x}, \hat{\lambda}, \hat{\mu}) \leq L(x, \hat{\lambda}, \hat{\mu})$$

this implies:

$$L(\hat{x}, \hat{\lambda}, \hat{\mu}) = \min_{x \in X_0} L(x, \hat{\lambda}, \hat{\mu})$$

Then,

$$-\nabla_x L(\hat{x}, \hat{\lambda}, \hat{\mu}) \in N_{X_0}(\hat{x})$$

This is equivalent to one of the first order optimality condition. On the other hand,

$$L(\hat{x}, \lambda, \mu) \leq L(\hat{x}, \hat{\lambda}, \hat{\mu})$$

implies $\sum_{i=1}^{m} \lambda_i g_i(\hat{x})$ is bounded from above for all $\lambda \geq 0$, hence $g_i(\hat{x}) \leq 0$, $i = 1, ..., m$. Similarly, $h_i(\hat{x}) = 0$, $i = 1, ..., p$. Consequently, the point $\hat{x}$ is feasible. As the maximum of $\sum_{i=1}^{m} \lambda_i g_i(\hat{x})$ is attained at $\hat{\lambda}$, we have

$$\hat{\lambda}_i g_i(\hat{x}) = 0, \ i = 1, ..., m$$

Thus, the assertion follows. □

The relation between saddle points and the primal and dual problem is straightforward:

Theorem 4.3.3 If $(\tilde{x}, (\tilde{\lambda}, \tilde{\mu}))$ is a saddle point, then

1. $\tilde{x}$ is a solution of the primal problem;
2. $(\tilde{\lambda}, \tilde{\mu})$ is a solution of the dual problem;
3. Duality holds, i.e., $L_P(\tilde{x}) = L_D(\tilde{\lambda}, \tilde{\mu})$

Proof. Take $(\lambda, \mu) \in \Lambda_0$,

$$\begin{aligned} L_D(\lambda, \mu) &= \min_{x \in X_0} L(x, \lambda, \mu) \leq L(\tilde{x}, \lambda, \mu) \\ &\leq L(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = \min_{x \in X_0}(x, \tilde{\lambda}, \tilde{\mu}) \\ &= L_D(\tilde{\lambda}, \tilde{\mu}) \end{aligned}$$

This proves (2). Similarly, take $x \in X_0$,

$$\begin{aligned} L_P(x) &= \max_{(\lambda,\mu)\in\Lambda_0} L(x,\lambda,\mu) \geq L(x,\tilde{\lambda},\tilde{\mu}) \\ &\geq L(\tilde{x},\tilde{\lambda},\tilde{\mu}) = \max_{(\lambda,\mu)\in\Lambda_0} L(\tilde{x},\lambda,\mu) = L_P(\tilde{x}) \end{aligned}$$

This proves (1). (3) follows easily. □

Remark 4.3.4 Note that no explicit convexity assumptions are made in this theorem. The duality relation always holds.

The converse of above theorem is also true:

Theorem 4.3.5 Assume that the duality relation is satisfied with finite values of the primal and dual functions. Then for every solution $\hat{x}$ of the primal problem and for every solution $(\hat{\lambda},\hat{\mu})$ of the dual problem the point $(\hat{x},\hat{\lambda},\hat{\mu})$ is a saddle point of *Lagrangian*.

Proof. We have

$$L_P(x) \geq L_P(\hat{x}) = \max_{(\lambda,\mu)\in\Lambda_0} L(\hat{x},\lambda,\mu) \geq L(\hat{x},\hat{\lambda},\hat{\mu})$$

also,

$$L_D(\lambda,\mu) \geq L_D(\hat{\lambda},\hat{\mu}) = \min_{x\in X_0} L(x,\hat{\lambda},\hat{\mu}) \leq L(\hat{x},\hat{\lambda},\hat{\mu})$$

Since duality relation holds, the assertion follows. □

Based on these results, we can look for a solution to the primal problem by first solving the dual problem to get $(\hat{\lambda},\hat{\mu})$ and then determining the primal solution $\hat{x}$ from the saddle point conditions.

Theorem 4.3.6 Assume that the duality relation holds true. If $(\bar{x},\bar{\mu}) \in \Lambda_0$ is a feasible point of the dual problem, then every $\hat{x} \in X_0$ such that

1. $L(\hat{x},\bar{\lambda},\bar{\mu}) = \min_{x\in X_0} L(x,\bar{\lambda},\bar{\mu})$;
2. all constraints of (4.1) are satisfied at $\hat{x}$;
3. $\bar{\lambda}_i g_i(\hat{x}) = 0$, $i = 1,...,m$,

is a solution of (4.1).

Proof. (i) says $L(\hat{x}, \bar{\lambda}, \bar{\mu}) \leq L(x, \bar{\lambda}, \bar{\mu})$, while (ii) and (iii) says

$$L(\hat{x}, \lambda, \mu) = f(\hat{x}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{x}) + \sum_{i=1}^{p} \mu_i h_i(\hat{x}) \leq f(x) = L(\hat{x}, \bar{\lambda}, \bar{\mu}, \bar{\mu})$$

Thus, $(\hat{x}, (\bar{\lambda}, \bar{\mu}))$ is a saddle point. Saddle point is equivalent to optimality condition. □

To summarize, if the problem has convex structure (f, g_i are convex and h_i is affine, X_0 is convex), then satisfying optimality condition is equivalent to the existence of saddle point. The saddle point reveals the dual solution and primal solution simultaneously and makes the duality relation holds. If the duality holds then, we can solve the dual problem and get the primal solution through saddle point conditions.

4.4 Application to Decomposition Problem

Consider the non-linear optimization problem:

$$\begin{aligned} \min . \ & f(x) \\ \text{subject to: } & g_i(x) \leq b_i, \ , i = 1, ..., m; \\ & h_i(x) = c_i, \ i = 1, ..., p; \\ & x \in X_0 \end{aligned} \tag{4.14}$$

with $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^p$. Assume that we can partition vector x as

$$X = (x^1, x^2, ..., x^K), \ x^k \in \mathbb{R}^{n_K}$$

where $\sum_{k=1}^{K} n_k = n$, in such a way that the objective function and constraint functions can be represented for all x as sums:

$$\begin{aligned} f(x) &= \sum_{k=1}^{K} f^k(x^k), \\ g_i(x) &= \sum_{k=1}^{K} g_i^k(x^k), \ i = 1, ..., m, \\ h_i(x) &= \sum_{k=1}^{K} h_i^k(x^k), \ i = 1, ..., p. \end{aligned}$$

Moreover, we assume that

$$X_0 = X_0^1 \times \cdots \times X_0^K, \ X_0^k \subset \mathbb{R}^{n_k}, \ k = 1, ..., K.$$

The *Lagrangian* has the form:

$$L(x,\lambda,\mu)=\sum_{k=1}^{K}\left(f^k(x^k)+\sum_{i=1}^{m}\lambda_i g_i^k(x^k)+\sum_{i=1}^{p}\mu_i h_i^k(x^k)\right)-\langle\lambda,b\rangle-\langle\mu,c\rangle$$

Therefore, the dual function is:

$$\begin{aligned}L_D(\lambda,\mu)&=\min_{x\in X_0}L(x,\lambda,\mu)\\&=\sum_{k=1}^{K}\min_{x^k\in X_0^k}\left[f^k(x^k)+\sum_{i=1}^{m}\lambda_i g_i^k(x^k)+\sum_{i=1}^{p}\mu_i h_i^k(x^k)\right]-\langle\lambda,b\rangle-\langle\mu,c\rangle\end{aligned}$$

It follows that calculation of the dual function decomposes into K smaller problems, each for the sub-vector x^k, $k=1,...,K$:

$$L_D^k(\lambda,\mu)=\min_{x^k\in X_0^k}\left[f^k(x^k)+\sum_{i=1}^{m}\lambda_i g_i^k(x^k)+\sum_{i=1}^{p}\mu_i h_i^k(x^k)\right]$$

In some cases the problems can be solved in a closed form. In other cases, efficient numerical methods can be employed for their solution. In any case, solving the dual problem:

$$\max_{\lambda\geq 0,\mu\in\mathbb{R}^p}\sum_{k=1}^{K}L_D^k(\lambda,\mu)-\langle\lambda,b\rangle-\langle\mu,c\rangle$$

maybe much easier than solving the primal problem. If the duality relation is satisfied, the above decomposition approach provides the optimal solution of the primal problem. If duality does not hold true, a lower bound for the optimal value of the primal problem can be obtained.

4.5 Augmented Lagrangian

4.5.1 Simplified System

Consider the non-linear optimization problem with equality constraints:

$$\begin{aligned}&\min.\ f(x)\\&\text{subject to: } h_i(x)=0,\ i=1,...,p,\\&\qquad x\in X_0\end{aligned}\tag{4.15}$$

We assume that the functions $f:\mathbb{R}^n\mapsto\mathbb{R}$ and $h_i:\mathbb{R}^n\mapsto\mathbb{R}$ are twice continuously differentiable, and the set X_0 is convex and closed. In general, if the problem (4.15) does not have

convex structure, it will not have a saddle point. However, in some case, a local saddle point of an *augmented Lagrangian function* exists. We define this function as follows:

$$L_\rho(x,\mu) = f(x) + \sum_{i=1}^{p} \mu_i h_i(x) + \frac{\rho}{2} \sum_{i=1}^{p} [h_i(x)]^2 \tag{4.16}$$

Here, $\rho > 0$ is a fixed parameter of the function.

We already knew that if *Robinson's condition* is satisfied, then there exists $\hat{\mu} \in \mathbb{R}^p$ such that

$$0 \in \nabla f(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) + N_{X_0}(\hat{x})$$

These are also the necessary conditions of the local optimality for the problem

$$\min_{x \in X_0} L_\rho(x, \hat{\mu})$$

evaluated at the point $\hat{x}$. To verify whether sufficient condition of a local minimum satisfied at the point $(\hat{x}, \hat{\mu})$, we calculate the *Hessian* of the augmented Lagrangian with respect to x:

$$\nabla^2_{xx} L_\rho(\hat{x}, \hat{\mu}) = \nabla^2 f(\hat{x}) + \rho \sum_{i=1}^{p} \hat{\mu}_i \nabla^2 h_i(\hat{x}) + \rho \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) [\nabla h_i(\hat{x})]^\top$$

Our intention is to show that for all sufficiently large ρ the hessian is a positive definite matrix and that the point $\hat{x}$ is therefore a local minimum of the augmented Lagrangian.

Lemma 4.5.1 Assume that a symmetric matrix Q of dimension n and a matrix B of dimension $m \times n$ are such that

$$\langle x, Qx \rangle > 0 \text{ for all } x \neq 0 \text{ such that } Bx = 0$$

Then there exists ρ_0 such that for all $\rho > \rho_0$ the matrix $Q + \rho B^\top B$ is positive definite.

Proof. Suppose the assertion is false, then for every ρ_k we can find $\bar{\rho}_k > \rho_k$ and some x_k such that

$$\langle x_k, (Q + \bar{\rho}_k B^\top B) x_k \rangle < 0. \tag{4.17}$$

We can always normalize x_k in such a way that $||x_k|| = 1$. Consider a sequence $\{\rho_k\}$ diverging to $+\infty$ and the corresponding $\{x_k\}$. By choosing a subsequence, if necessary, we can assume that the sequence $\{x_k\}$ is convergent. Let z be its limit. Dividing both sides of (4.17) by $\bar{\rho}_k$ we obtain

$$\frac{1}{\bar{\rho}_k} \langle x_k, Q x_k \rangle + ||Bx_k||^2 \leq 0$$

Passing to the limit, as $k \to \infty$, we conclude that $Bz = 0$. Skipping the first term, we also have

$$\langle x_k, Qx_k \rangle \leq 0$$

Passing to the limit,

$$\langle z, Qz \rangle \leq 0$$

which contradicts the assumption. □

Using this little lemma, we will be able to show the local convexity of augmented Lagrangian.

Theorem 4.5.2 Assume that a point $\hat{x}$ satisfying the second order sufficient conditions of optimality with *Lagrangian multiplier* $\hat{\mu}$. Then there exists $\rho_0 \geq 0$ such that for all $\rho > \rho_0$ the point $\hat{x}$ is a local minimum over X_0 of the *augmented Lagrangian* with $\mu = \hat{\mu}$.

Proof. According to the theorem regarding second optimality condition, if $\hat{x}$ satisfies:

$$0 \in \nabla f(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) + N_{X_0}(\hat{x})$$

and $\hat{\Lambda}(\hat{x})$ is the set of *Lagrangian multipliers* $\hat{\mu}$ for this problem. For every non-zero $s \in T_X(\hat{x})$, where

$$T_X(\hat{x}) = \{s \in T_{X_0}(\hat{x}) : \ \langle \nabla g_i(\hat{x}), s \rangle \leq 0, \ i \in I^0(\hat{x}); \ \langle \nabla h_i(\hat{x}), s \rangle = 0, \ i = 1, ..., p\}$$

We have the assumption of above theorem, with

$$Q = \nabla^2_{xx} L(\hat{x}, \hat{\mu}), \ B = h'(\hat{x}) = [(\nabla h_1(\hat{x}))^\top \ \cdots \ (\nabla h_p(\hat{x}))^\top]^\top$$

because, second optimality condition says

$$\langle s, \nabla^2_{xx} L(\hat{x}, \hat{\mu}) s \rangle > 0$$

for some $\hat{\mu}$. It follows immediately that

$$\langle s, \nabla^2_{xx} L_\rho(\hat{x}, \hat{\mu}) s \rangle > 0$$

□

It follows that the point $(\hat{x}, \hat{\mu})$ is a *local saddle point* of the augmented Lagrangian: there exists $\rho_0 > 0$ and a neighbourhood U of $\hat{x}$ such that for all $\rho > \rho_0$,

$$\max_{\mu \in \mathbb{R}^n} L_\rho(\hat{x}, \mu) = L_\rho(\hat{x}, \hat{\mu}) = \min_{x \in X_0 \cap U} L_\rho(x, \hat{\mu})$$

The converse is also true.

Theorem 4.5.3 Assume that a point $(\bar{x}, \bar{\mu})$ is a local saddle point of the augmented Lagrangian for some $\rho \geq 0$. Then $\bar{x}$ is a local minimum of problem (4.15) and $\bar{\mu}$ is the vector of Lagrangian multipliers associated with the equality constraints.

Proof. Skipped. □

4.5.2 Complete System

Let's now consider the full system:

$$\begin{aligned} &\min . \ f(x) \\ &\text{subject to: } g_i(x) \leq 0, \ i = 1, ..., m, \\ &\qquad\qquad h_i(x) = 0, \ i = 1, ..., p, \\ &\qquad\qquad x \in X_0. \end{aligned} \tag{4.18}$$

We assume that all functions involved in this problem are twice continuously differentiable and that the set X_0 is convex and closed. The *Lagrangian* associated with (4.18) has the form:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x) \tag{4.19}$$

Suppose $\hat{x}$ is a local minimum of problem (4.18) and that the set $\hat{\Lambda}(\hat{x})$ of the Lagrange multipliers associated with the inequality constraints, respectively, is non-empty. Consider the set of active constraints:

$$I^0(\hat{x}) = \{1 \leq i \leq m : g_i(\hat{x}) = 0\}.$$

Definition 4.5.1 Problem (4.18) satisfies the strong second order sufficient condition if there exists multipliers such that

1. $\hat{\lambda}_i > 0$, $i \in I^0(\hat{x})$;
2. $\langle s, \nabla^2_{xx} L(\hat{x}, \hat{\lambda}, \hat{\mu}) > 0$, for all non-zero s satisfying the equations:

$$\langle \nabla g_i(\hat{x}), s \rangle = 0, \ i = 1 \in I^0(\hat{x}) \tag{4.20}$$

$$\langle \nabla h_i(\hat{x}), s \rangle = 0, \ i = 1, ..., p. \tag{4.21}$$

In order to define the augmented Lagrangian for the inequality constrained problem, we transform (4.18) to the equality constrained problem:

$$\begin{aligned} &\min . \ f(x) \\ &\text{subject to: } g_i(x) + (z_i)^2 = 0, \ i = 1, ..., m, \\ &\qquad\qquad h_i(x) = 0, \ i = 1, ..., p, \\ &\qquad\qquad x \in X_0. \end{aligned} \tag{4.22}$$

It is apparent that $\hat{x}$ that $\hat{x}$ and $\hat{z}_i^2 = -g_i(\hat{x})$, $i = 1, ..., m$, constitute a local minimum of this problem. The set of Lagrange multipliers $\hat{\Lambda}(\hat{x})$ remains unchanged. The Lagrangian of above problem has the form:

$$\bar{L}(x, z, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i [g_i(x) + z_i^2] + \sum_{i=1}^{p} \mu_i h_i(x) \tag{4.23}$$

The augmented Lagrangian for problem can be written as follows:

$$\bar{L}_\rho(x,z,\lambda,\mu) = \bar{L}(x,z,\lambda,\mu) + \frac{\rho}{2}\sum_{i=1}^{m}[g_i(x)+z_i^2]^2 + \frac{\rho}{2}\sum_{i=1}^{p}[h_i(x)]^2 \tag{4.24}$$

Here $\rho \geq 0$ is a fixed parameter of the function.

Theorem 4.5.4 Assume that problem (4.18) satisfies the strong second order sufficient condition. Then there exists $\rho_0 \geq 0$ such that for all $\rho > \rho_0$ the point $(\hat{x},\hat{z})$ is a local minimum over X_0 of the augmented Lagrangian $\bar{L}_\rho(x,z,\hat{\lambda},\hat{\mu})$.

It follows that the point $(\hat{x},\hat{z},\hat{\lambda},\hat{\mu})$ is local saddle point of the augmented Lagrangian: there exists ρ_0 $geq0$ and a neighborhood U of $(\hat{x},\hat{z})$ such that for all $\rho > \rho_0$,

$$\max_{\lambda\geq 0,\mu\in\mathbb{R}^n} \bar{L}_\rho(\hat{x},\hat{z},\hat{\lambda},\hat{\mu}) = \bar{L}_\rho(\hat{x},\hat{z},\hat{\lambda},\hat{\mu}) = \min_{(x,z)\in(X_0\times\mathbb{R}^m)\cap U} \bar{L}_\rho(x,z,\hat{\lambda},\hat{\mu})$$

In a way identical to previous theorem in the reduced case,

Theorem 4.5.5 Assume that a point $(\bar{x},\bar{z},\bar{\lambda},\bar{\mu})$ is a local saddle point of the augmented Lagrangian for some $\rho \geq 0$. Then $(\bar{x},\bar{z})$ is a local minimum of (4.22) and $(\bar{\lambda},\bar{\mu})$ is the vector of Lagrange multipliers associated with its constraints.

In practice, we rarely deal with the Lagrangian in its full form. The dependence of this function of z is simple, and the minimum with respect to z can be calculated in a closed form:

$$\begin{aligned} L_\rho(x,\lambda,\mu) = f(x) &+ \sum_{i=1}^{m}\lambda_i \max(g_i(x),-\frac{\lambda_i}{\rho},-\frac{\lambda_i}{\rho}) + \sum_{i=1}^{p}\mu_i h_i(x) \\ &+ \frac{\rho}{2}\sum_{i=1}^{m}\left(\max(g_i(x),-\frac{\lambda_i}{\rho})\right)^2 + \frac{\rho}{2}\sum_{i=1}^{p}[h_i(x)]^2 \end{aligned}$$

By simple manipulations, it can be shown that we have the following equivalent characterization:

$$L_\rho(x,\lambda,\mu) = f(x) + \frac{\rho}{2}\sum_{i=1}^{m}[\max(0,g_i(x)+\frac{\lambda_i}{\rho})]^2 + \frac{\rho}{2}\sum_{i=1}^{p}[h_i(x)+\frac{\mu_i}{\rho}]^2 - \frac{1}{2\rho}\left(\sum_{i=1}^{m}\lambda_i^2 + \sum_{i=1}^{p}\mu_i^2\right) \tag{4.25}$$

If all functions involved are continuously differentiable, then the augmented Lagrangian L_ρ is continuously differentiable, because the function $t \mapsto [\max(0,t)]^2$ is smooth.

Chapter 5

Unconstrained Optimization

Consider the following *unconstrained problem*:

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

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a continuously differentiable function. The *local minimum* $\hat{x}$ has the property

$$\nabla f(\hat{x}) = 0 \tag{5.1}$$

But it is not always easy to get full derivatives of $f(x)$, actually, in some case, it will involve unexpected amount of efforts. So, for large scale system, we always rely on the *iterative method*.

Intuitively, we want to construct a *sequence*: $x^0, x^1, ..., x^k, ... \to \hat{x}$ (where $\hat{x}$ is a *local minimum*), for which we have the mechanism,

$$(x^0, x^1, x^2, ..., x^k) \to x^{k+1}$$

such that $f(x^{k+1}) < f(x^k)$. In particular, we would like to have the following construction (in *figure* 5.1),

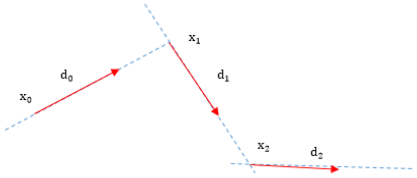

Figure 5.1: Iterative Method

$$x^{k+1} = x^k + \tau d^k,\ \tau \in \mathbb{R}^+$$

There are *two steps*:

1. choose direction d^k at point x^k;
2. find $x^{k+1} = x^k + \tau_k d^k$, where τ_k needs to be choose.

Let's first address the second step.

5.1 Line Search

Let's state the probelm again, given x^k and direction d^k, we want to

$$\min_{\tau \in \mathbb{R}} \ \psi(\tau) \tag{5.2}$$

where,

$$\psi(\tau) = f(x^k + \tau d^k)$$

This is relative simple task because we have function ψ depending on just on variable.

1. Golden Section

We firstly initialize $\alpha_0 < \beta_0 < \gamma_0 < \delta_0$ such that

$$\psi(\alpha_0) > \psi(\beta_0) \text{ and } \psi(\gamma_0) < \psi(\delta_0)$$

Since ψ is a continuous function, we are sure that a local minimum is contained in the interval $[\alpha_0, \delta_0]$. We shall maintain such relation for all iterations k. It is convenient to satisfy the proportions:

$$\frac{\gamma_k - \alpha_k}{\delta_k - \alpha_k} = \frac{\delta_k - \beta_k}{\delta_k - \alpha_k} = q$$

with *golden ratio*, $q \approx 0.618$. Then it is easy to calculate that also:

$$\frac{\beta_k - \alpha_k}{\gamma_k - \alpha_k} = \frac{\delta_k - \gamma_k}{\delta_k - \beta_k} = q$$

Now, we compare $\psi(\beta_k)$ and $\psi(\gamma_k)$, two cases can happen:

1. If $\psi(\beta_k) < \psi(\gamma_k)$, then the local minimum is in $[\alpha_k, \gamma_k]$, we discard the point δ_k to make:

$$\alpha_{k+1} := \alpha_k,\ \delta_{k+1} = \gamma_k,$$
$$\gamma_{k+1} = \beta_k,\ \beta_{k+1} = q\alpha_k + (1-q)\gamma_k.$$

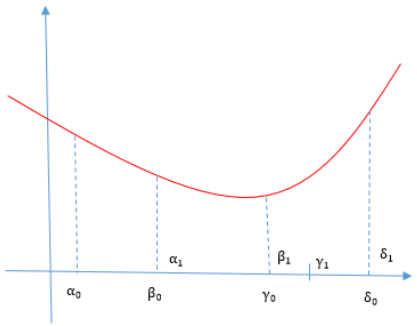

Figure 5.2: Golden Section

2. If $\psi(\beta_k) \geq \psi(\gamma_k)$, then the local minimum is in $[\beta_k, \delta_k]$. We discard α_k and make:

$$\alpha_{k+1} := \beta_k, \ \beta_{k+1} = \gamma_k,$$
$$\delta_{k+1} = \delta_k, \ \gamma_{k+1} = q\delta_k + (1-q)\beta_k.$$

In this way, we maintain that the length of original interval is shrinking at a rate q at each step.

2. Interpolation

The golden section method, although elegant and applicable to all continuous functions, has rather slow convergence. If the function is twice continuously differentiable, we can approximate the minimum much faster by *interpolation*. At iteration k of the method, we have three points

$$\alpha_k < \tau_k < \beta_k$$

such that

$$\psi(\alpha_k) > \psi(\tau_k) < \psi(\beta_k)$$

Then the function $\psi(\cdot)$ has a minimum in the interval $[\alpha_k, \beta_k]$. Next, we interpolate $\psi(\cdot)$ with the second order Lagrange polynomial, determined from the values of $\psi(\cdot)$ at the nodes α_k, τ_k and β_k. The minimum of this polynomial is denoted by γ_k. Simple algebra yields the minimizer γ_k, which is for sure in the interval (α_k, β_k). After that, we evaluate $\psi(\gamma_k)$ and we remove one of the end points, so that out of remaining three points the middle one is the best.

Remark 5.1.1 Notice we can not prove that this method generates a sequence convergent to a local minimum of $\psi(\cdot)$. In practice, the method of interpolation is used with many technical safeguards guaranteeing that the interval $[\alpha_k, \beta_k]$ shrinks sufficiently fast. There are many variations of these techniques, using interpolation with high order polynomials.

5.2 Steepest Descent

5.2.1 Illustration of the method

Now, let's start to consider the first step that is to find a searching direction. In particular, let's discuss the most basic method - *steepest descent*, in which we always choose

$$d^k = -\nabla f(x^k) \tag{5.3}$$

at each iteration.

Suppose d^k is fixed, then we can construct the first order expansion of $f(x^k + \tau d^k)$ about the point x^k:

$$f(x^k + \tau d^k) = f(x^k) + \tau \langle \nabla f(x^k), d^k \rangle + o(\tau)$$

where $\lim_{\tau \downarrow 0} \frac{o(\tau)}{\tau} = 0$. If we wan to keep

$$f(x + \tau d^k) > f(x^k)$$

We need to have, for $\tau^k > 0$,

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Recall that this term is the *directional derivative* $f'(x^k; d^k)$ of $f(\cdot)$ at the point x^k and direction d^k. Thus, it seems to be reasonable to choose a direction and make it as negative as possible. But since directional derivative of a differentiable function is linear in d, there exists no minimum unless $\nabla f(x^k) = 0$. In order to formulate a valid optimization problem, we need to have an upper bound of the length of the direction. Any bound will work here, we choose it to be 1. Thus, we choose d by solving the following optimization problem:

$$\begin{aligned} \min_d \ & \langle \nabla f(x), d \rangle \\ \text{subjec to: } & ||d||^2 \leq 1 \end{aligned}$$

By *Lagrangian*, we have

$$\nabla f(x) + 2\lambda d = 0$$

If $\nabla f(x) \neq 0$, $\lambda > 0$,

$$d = -\nabla f(x)/2\lambda$$

Since the constant really does not matter, we can normalize by setting:

$$d = -\frac{\nabla f(x)}{||\nabla f(x)||}$$

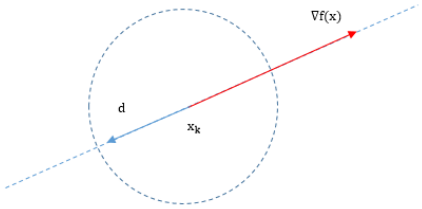

Figure 5.3: Steepest Descent

Graphically, it means (in *figure* 5.3). Let's observe that all points x^k generated by this method belong to the set

$$X_0 = \{x \in \mathbb{R} : f(x) \leq f(x_0)\}$$

because $f(x_{k+1}) \leq f(x_k)$ for all k.

Theorem 5.2.1 The method of *steepest descent* generate a sequence $\{x_k\}$, each of its accumulation point x^* satisfies: $\nabla f(x^*) = 0$.

Proof. By the construction of the method, we have $f(x_{k+1}) \leq f(x_k)$ for all k. Furthermore, all points x_k belong to the set X_0, which is compact. A continuous function on a compact set can achieve its maximum and minimum, so the sequence $\{f(x_k)\}$ is bounded and also monotonically decreasing, thus it is convergent.

Moreover, the sequence $\{x_k\}$ has an accumulation point, since X_0 is bounded. Consider any accumulation point x^* and let $\mathcal{K}$ be the infinite order set of iteration number such that

$$\lim_{k\to\infty,\ k\in\mathcal{K}} x_k = x^*$$

Since the sequence $\{f(x_k)\}$ is convergent, we have

$$f(x^*) = \lim_{n\to\infty} f(x_k) \tag{5.4}$$

We prove the theorem by contradiction. Suppose that the point x, i.e., $\nabla f(x^*) \neq 0$. Consider the direction of steepest descent $-\nabla f(x^*)$ and the points:

$$y(\tau) = x^* - \tau \nabla f(x^*),\ \tau \geq 0$$

Since $\nabla f(x^*) \neq 0$, the problem

$$\min_{\tau\geq 0}\ f(y(\tau))$$

has a solution $\tau_* > 0$, and $f(y(\tau_*)) < f(x^*)$ for every solution τ_*. Of course, the method does not carry out the directional minimization at x^*, but we shall show that the outcome is similar, if the directional minimization is carried out at points which are close to x^*.

Consider the points x_k for $k \in \mathcal{K}$, and the immediately following points x^{k+1}. By the directional minimization condition:

$$f(x_{k+1}) \leq f(x_k - \tau \nabla f(x_k)) \text{ for all } \tau \geq 0 \tag{5.5}$$

when $k \to \infty$, $k \in \mathcal{K}$, then $x_k \to x^*$ and $\nabla f(x_k) \to f(x^*)$. Define $\epsilon = ||\nabla f(x^*)||$. We have

$$x_{k+1} = x_k - \tau_k \nabla f(x_k) \tag{5.6}$$

and $||\nabla f(x_k)|| \geq \epsilon/2$ for all sufficiently large $k \in \mathcal{K}$. Therefore, for all sufficiently large $k \in \mathcal{K}$,

$$0 \leq \tau_k = \frac{||x_{k+1} - x_k||}{||\nabla f(x_k)||} \leq \frac{2}{\epsilon} \text{diam}(X_0)$$

with $\text{diam}(X_0)$ denoting the largest distances between two points of X_0. Consequently, the step sizes τ_k for all sufficiently large $k \in \mathcal{K}$ are uniformly bounded. We can therefore choose an infinite subset $\mathcal{K}_1$ of $\mathcal{K}$ such that the sequence $\{\tau_k\}$, $k \in \mathcal{K}_1$, has a limit. We denote this limit by $\bar{\tau}$. Passing to the limit in (5.6) yields:

$$\lim_{k \to \infty,\ k \in \mathcal{K}_1} x_{k+1} = x^* - \bar{\tau} \nabla f(x^*)$$

We can now pass to the limit in inequalities (5.5), when $k \to \infty$, $k \in \mathcal{K}_1$. We obtain:

$$f(x^* - \bar{\tau} \nabla f(x^*)) \leq f(x^* - \tau \nabla f(x^*)) \text{ for all } \tau \geq 0$$

We conclude that the step size $\bar{\tau}$ is an optimal solution of the problem. Hence,

$$\lim_{k \to \infty\ k \in \mathcal{K}_1} f(x_{k+1}) = f(y(\tau_*)) < f(x^*)$$

Then also

$$\lim_{k \to \infty} f(x_k) \leq \lim_{k \to \infty\ k \in \mathcal{K}_1} f(x_{k+1}) < f(x^*)$$

which contradicts (5.4). □

5.2.2 Rate of Convergence

Suppose $x_k \to x^*$,

$$\limsup_{k \to \infty} \frac{||x_{k+1} - x^*||}{||x_k - x^*||} = q < 1$$

It is *converging linearly*. If $q = 0$, it is *converging sup-linear*. If

$$\limsup_{k\to\infty} \frac{||x_{k+1} - x^*||}{||x_k - x^*||}$$

is finite, it is called *quadratic convergent*. At $k+1$ iteration, we have value function

$$\psi(\tau_k) = f(x_k - \tau_k \nabla f(x_k))$$

Then, we choose the exact optimal by taking derivative and sending it to zero,

$$\psi'(\tau_k) = -\langle \nabla f(x_k - \tau_k \nabla f(x_k)), \nabla f(x_k) \rangle = 0$$

This implies that in this method of each iteration we are choosing directions that are orthogonal to each other (in *figure* 5.4): Assume $\hat{x}$ is a minimizer, $f(\cdot)$ is a *twice continuously*

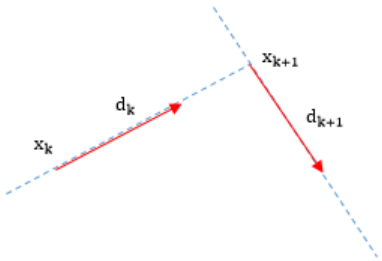

Figure 5.4: Directions are orthogonal

differentiable function, then

$$\nabla f(\hat{x}) = 0$$

Furthermore,

$$Q = \nabla^2 f(\hat{x})$$

is positive definite. We do *second order expansion* of $f(\cdot)$ at $\hat{x}$,

$$f(x) \approx f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \frac{1}{2} \langle x - \hat{x}, \nabla^2 f(\hat{x})(x - \hat{x}) \rangle + o_2(x; \hat{x})$$

where $\lim_{x\to\infty\hat{x}} \frac{o_2(x;\hat{x})}{||x-\hat{x}||^2} = 0$. The second term on the right hand side goes to 0, without loss of generality, we assume that $f(\hat{x}) = 0$, because otherwise we just need to shift it.

Let's consider the quadratic form:

$$f(x) = \frac{1}{2} \langle x, Qx \rangle$$

where Q is positive definite. Apply steepest descent, we define

$$\nabla f(x_k) = g = Qx$$

then, $x_{k+1} = x - \tau g$,

$$\begin{aligned} f(x_{k+1}) &= \frac{1}{2}\langle x - \tau g, Q(x - \tau g)\rangle \\ &= \frac{1}{2}\langle x, Qx\rangle - \tau\langle g, Qx\rangle + \frac{1}{2}\tau^2\langle g, Qg\rangle \\ &= f(x_k) - \tau||g||^2 + \frac{1}{2}\tau^2\langle g, Qg\rangle \end{aligned}$$

Rewrite as a function of τ, and take derivative before sending it to 0:

$$\psi'(\tau) = -||g||^2 + \tau\langle g, Qg\rangle = 0$$

which yields the optimal $\tau = ||g||^2/\langle g, Qg\rangle$. Thus,

$$\begin{aligned} f(x_{k+1}) &= f(x_k) - \frac{||g||^4}{\langle g, Qg\rangle} + \frac{1}{2}\frac{||g||^4}{g, Qg} \\ &= f(x_k) - \frac{1}{2}\frac{||g||^4}{\langle g, Qg\rangle} \end{aligned}$$

Since $x = Q^{-1}g$, $f(x) = \frac{1}{2}\langle g, Q^{-1}g\rangle$. Thus,

$$f(x_{k+1}) = f(x_k) - \frac{1}{2}\frac{\langle g, Q^{-1}g\rangle||g||^4}{\langle g, Q^{-1}g\rangle\langle g, Qg\rangle}$$

As a result,

$$f(x_{k+1}) = f(x_k)\left(1 - \frac{||g||^4}{\langle g, Q^{-1}g\rangle\langle g, Qg\rangle}\right)$$

Since $f(\hat{x}) = 0$, the convergent rate is

$$\frac{f(x_{k+1}) - f(\hat{x})}{f(x_k) - f(\hat{x})} = 1 - \frac{||g||^4}{\langle g, Q^{-1}g\rangle\langle g, Qg\rangle}$$

Because Q is symmetric, thus we can find $z_1, z_2, ..., z_n$ are *orthogonal eigenvectors* of Q, with normalization $||z_i|| = 1$, $\forall i$ and corresponding *eigenvalue*, $\lambda_1, ..., \lambda_n$. Since g is in the range of Q, we can expand g by the orthonormal basis, namely,

$$g = \xi_1 z_1 + \xi_2 z_2 + \cdots + \xi_n z_n$$

Since $||g||^2 = \xi_1^2 + \cdots + \xi_n^2$,

$$Qg = \xi_1 Q z_1 + \cdots + \xi_n Q_n z_n = \xi_1\lambda_1 z_1 + \cdots + \xi_n\lambda_n z_n$$

then,

$$\langle g, Qg\rangle = \sum_{i=1}^{n} \lambda_i \xi_i^2$$

also,

$$Q^{-1}g = \sum_{i=1}^{n} \xi_i Q^{-1} z_i = \sum_{i=1}^{n} \frac{\xi_i}{\lambda_i} z_i$$

This gives us

$$\frac{||g||^4}{\langle g, Q^{-1}g\rangle\langle g, Qg\rangle} = 1 - \frac{(\sum_{i=1}^{n} \xi_i^2)^2}{(\sum_{i=1}^{n} \frac{\xi_i^2}{\lambda_i})(\sum_{i=1}^{n} \xi_i^2 \lambda_i)}$$

Set $\alpha_i = \frac{\xi_i^2}{\sum_{i=1}^{n} \xi_i^2}$, then

$$\frac{f(x_{k+1} - f(\hat{x}))}{f(x_k) - f(\hat{x})} = 1 - \frac{\frac{1}{\sum_{i=1}^{n} \alpha_i - \lambda_i}}{\sum_{i=1}^{n} \alpha_i \frac{1}{\lambda_i}}$$

define $\Phi(\lambda) = \frac{1}{\lambda}$,

$$\frac{f(x_{k+1} - f(\hat{x}))}{f(x_k) - f(\hat{x})} = 1 - \frac{\Phi(\sum_{i=1}^{n} \alpha_i \lambda_i)}{\sum_{i=1}^{n} \alpha_i \Phi(\lambda_i)}$$

Since $\Phi(\cdot)$ is convex,

$$\Phi(\lambda_i) \leq \beta_i \Phi(\lambda_i) + (1 - \beta_i)\Phi(\lambda_n) \tag{5.7}$$

where $\beta_i = \frac{\lambda_n - \lambda_i}{\lambda_n - \lambda_1}$. We can add these inequalities multiplied by α_i to estimate the denominator of (5.7) as follows:

$$\sum_{i=1}^{n} \alpha_i \Phi(\lambda_i) \leq \alpha\Phi(\lambda_1) + (1 - \alpha)\Phi(\lambda_n)$$

where $\alpha = \sim_{i=1}^{n} \alpha_i \beta_i$. The numerator of (5.7) is,

$$(\lambda_n - \lambda_1)\alpha = (\lambda_n - \lambda_1) \sum_{i=1}^{n} \alpha_i \beta_i = \sum_{i=1}^{n} \alpha_i (\lambda_n - \lambda_i) = \lambda_n - \sum_{i=1}^{n} \alpha_i \lambda_i$$

So,

$$\sum_{i=1}^{n} \alpha_i \lambda_i = \lambda_n - (\lambda_n - \lambda_1)\alpha = \alpha\lambda_1 + (1 - \alpha)\lambda_n$$

Hence,

$$(5.7) \geq \frac{\Phi(\alpha\lambda_1 + (1-\alpha)\lambda_n)}{\alpha\Phi(\lambda_1) + (1-\alpha)\Phi(\lambda_n)} = \frac{\lambda_1\lambda_n}{[\alpha\lambda_1 + (1-\alpha)\lambda_n][\alpha\lambda_n + (1-\alpha)\lambda_1]} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}$$

Thus,

$$\frac{f(x_{k+1} - f(\hat{x}))}{f(x_k) - f(\hat{x})} = \frac{(\lambda_n - \lambda_1)^2}{(\lambda_n + \lambda_1)^2}$$

Remark 5.2.2 We choose a simpler version: $f(x) = \frac{1}{2}\langle x, Qx\rangle$ for the convergence analysis, because every twocei differentiable function can be very well approximated by a quadratic function in a neighbourhood of its minimum point. The assumptions that the minimum point is 0, and the minimum value is 0 as well, have been made only for simplicity of notation. They can always be satisfied by shifting the argument and adding a constant to the function. Besides, when we show the convergence rate, we can set λ_1 and λ_n as the smallest and the biggest eigenvalue of Q, we set the ratio

$$\chi = \frac{\lambda_n}{\lambda_1}$$

It is usually called the *condition index*. Because then, the convergence rate q can be expressed in a simple form

$$q = \frac{\chi - 1}{\chi + 1}$$

When it is smaller, that is $\chi \downarrow 1$ and the spectrum of the *Hessian* of Q is concentrated within a relatively short interval, the convergence will become very slow. While if the spectrum is widely spread, $\chi >> 1$, the convergence can be very fast.

5.3 Conjugate Gradient Method

5.3.1 Illustration of the method

Consider following function:

$$f(x) = \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle$$

where Q is positive definite matrix. Then,

$$\nabla f(x) = Qx + c$$

at local minimum $\hat{x}$,

$$Q\hat{x} = -c \;\Rightarrow\; \hat{x} = -Q^{-1}c$$

Now we define following scalar product with respect to:

$$\langle x, y\rangle_Q := \langle x, Qy\rangle$$

so that its induced norm is:

$$||x||_\alpha := \langle x, Qx\rangle^{\frac{1}{2}}$$

Definition 5.3.1 Let Q be a symmetric positive definite matrix of dimension n, vectors $d^1, d^2, ..., d^n$ are called *Q-conjugate* if they are all non-zero and $\langle d^i, d^j\rangle_Q = 0$, for $i \neq j$.

Lemma 5.3.1 Conjugate directions are linearly independent.

Proof. Suppose

$$\gamma_1 d^1 + \cdots + \gamma_n d^n = 0$$

multiplying both sides by Qd^i,

$$\gamma_i \langle d^i, Qd^i\rangle = 0$$

Since d^i is non-zero, $\gamma_i = 0$, we will have $d^1, ..., d^n$ are linearly independent. □

The motivation is that we want those conjugate directions to help to solve the quadratic function below:

$$f(x) = \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle \tag{5.8}$$

efficiently. Because when dimension n becomes very large, minimization can be very difficult, we can't solve $Qx = -c$ easily, so if our conjugate method works it will render us a way to solve this system easily. On the other hand, a twice continuously differentiable function can be well approximated by quadratic model in neighbourhood of its local minimum.

Theorem 5.3.2 Assume that $d^1, d^2, ..., d^n$ are Q-conjugate and that the sequence $x^1, x^2, ..., x^{n+1}$ is obtained by successive minimization of function (5.8) in direction d^k, $k = 1, ..., n$:

1. $x^{k+1} = x^k + \tau_k d^k$;
2. $f(x^{k+1}) = \min_{\tau\in\mathbb{R}} f(x^k + \tau d^k)$

then for every $k = 1, 2, ..., n$ the point x^{k+1} is the minimum of (5.8) in the linear manifold

$$L_k = x^1 + \text{lin}\{d^1, ..., d^k\}$$

Proof. Let's consider the matrix

$$D = [d^1 \ d^2 \ \cdots \ d^n]$$

and the linear transform of variables

$$x = x^1 + Dy = x^1 + d^1 y_1 + d^2 y_2 + \cdots + d^n y_n$$

As the direction d^i are conjugate,

$$\langle Dy, QDy \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle d^i, Qd^j \rangle y_i y_j = \sum_{i=1}^{n} \langle d^i, Qd^i \rangle (y_i)^2$$

Using Taylor expansion of $f(\cdot)$ at x^1, which is exact, we obtain

$$f(x) = f(x^1 + Dy) = f(x^1) + \frac{1}{2}\langle Dy, QDy \rangle + \langle Qx^1 + c, Dy \rangle = f(x^1) + \sum_{i=1}^{n} f_i(y_i)$$

where

$$f_i(y_i) = \frac{1}{2}\langle d^i, Qd^i \rangle (y_i)^2 + \langle Qx^1 + c, d^i \rangle y_i$$

The minimum of $f(\cdot)$ w.r.t y can be calculated independently for each y_i. The best value for each y_i is equal to the step size τ_i used at iteration i if our method. Consequently, each direction d^i need to be used only once, and each x^{k+1} is the minimum of $f(\cdot)$ over L_k. □

Corollary 5.3.3 The minimum of (5.8) can be found in no more than n steps.

Proof. The directions d^k, $k = 1, ..., n$ are linearly independent, so $L_n = \mathbb{R}^n$, and the result follows from above theorem. □

Conjugate directions are very good search directions for a quadratic function, but it seems that one needs to know the Hessian to construct them. We shall show that this is not necessary at all. Assume temporarily that we know the Hessian Q. We can construct a sequence of conjugate directions $d^1, d^2, ..., d^n$ and a sequence of points $x^1, x^2, ..., x^{n+1}$ by the process of successive orthogonalization of the gradients $\nabla f(x^1), \nabla f(x^2), ..., \nabla f(x^n)$.

Gram-Schmidt Orthogonalization procedure

1. *Step 0:* set $k = 1$;
2. *Step 1:* calculate $v^k = -\nabla f(x^k)$, if $v^k = 0$, then stop, otherwise continue;
3. *Step 2:* make the vector v^k Q-orthogonal to directions $d^1, d^2, ..., d^{k-1}$ by the formula:

$$d^k := v^k - \sum_{i=1}^{k-1} \frac{\langle v^k, Qd^i \rangle}{\langle d^i, Qd^i \rangle} d^i \tag{5.9}$$

4. *Step 3:* Calculate the next point

$$x^{k+1} = x^k + \tau_k d^k$$

such that

$$f(x^{k+1}) = \min_{\tau \in \mathbb{R}} f(x^k + \tau d^k)$$

5. *Step 4:* increase k by 1 and go to *Step 1.*

Suppose the first $k-1$ directions $d^1, ..., d^{k-1}$ are Q-orthogonal. Then by *theorem* 5.3.2, the point x^k is the minimum of $f(\cdot)$ in the manifold:

$$L_{k-1} = x^1 + \text{lin } \{d^1, ..., d^{k-1}\}$$

Consequently, the vector $v^k = -\nabla f(x^k)$ satisfies the relation:

$$v^k \perp L_{k-1} \tag{5.10}$$

Since each direction d^i is a linear combination of the previously observed gradients v^j, for

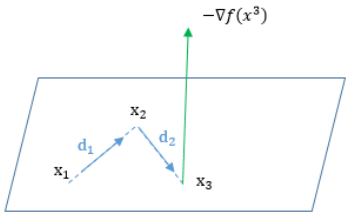

Figure 5.5: Manifold and minimum

$j \leq i$, we can represent the manifold L_{k-1} as follows:

$$L_{k-1} = x^1 + \text{lin } \{v^1, v^2, ..., v^{k-1}\} = x^1 + \text{lin } \{v^1, v^2 - v^1, ..., v^k - v^{k-1}\}.$$

Therefore, (**??**) implies that

$$v^k \perp (v^j - v^{j-1}) \text{ for } 2 \leq j \leq k-1$$

Using the definition of v^j we obtain:

$$v^j - v^{j-1} = -Q(x^j - x^{j-1}) = -\tau_{j-1} Q d^{j-1}$$

From (5.9) for $j-1$ it follows that d^{j-1} is a linear combination of $-\nabla f(x^{j-1})$ and of directions d^i, for $i \leq j-2$. Since $\nabla f(x^{j-1}) \perp d^i$, for $i \leq j-2$, we conclude that

$$\langle \nabla f(x^{k-1}), d^{j-1} \rangle = -||\nabla f(x^{j-1})||^2,$$

which is negative, if the algorithm did not stop at x^{j-1}. As the direction d^{j-1} is a direction of descent, we conclude that $\tau_{j-1} > 0$. It follows that

$$v^k \perp Q d^{j-1}, \text{ for } 2 \leq j \leq k-1$$

Therefore, all but the last components of them sum in (5.9) vanish and we get

$$d^k = -\nabla f(x^k) + a_k d^{k-1}$$

with

$$a_k = -\frac{\nabla f(x^k), Q d^{k-1}\rangle}{\langle d^{k-1}, Q d^{k-1}\rangle}$$

Using the equality $Q d^{k-1} = (\nabla f(x^k) - \nabla f(x^{k-1}))/\tau_{k-1}$, we can transform the last equation as follows:

$$a_k = -\frac{\nabla f(x^k), \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\langle d^{k-1}, f(x^k) - \nabla f(x^{k-1})\rangle}$$

We can further write it as:

$$a_k = -\frac{||\nabla f(x^k)||^2}{||\nabla f(x^{k-1})||^2} \tag{5.11}$$

This convince us that the algorithm can be implemented without the knowledge of Q. It is known as the *conjugate gradient method* of *Hastenes* and *Stiefel*.

Then the *conjugate-gradient algorithm* can be described as follows:

1. *Step 0:* set $k = 1$;
2. *Step 1:* calculate $\nabla f(x^k)$. If $\nabla f(x^k) = 0$, then stop; otherwise, continue;
3. *Step 2:* Calculate:
$$d^k = \begin{cases} -\nabla f(x^k) & \text{if } k = 1, \\ -\nabla f(x^k) + a_k d^{k-1} & \text{if } k > 1. \end{cases}$$
with a_k defined in (5.11).
4. *Step 3:* Calculate the next point:
$$x^{k+1} = x^k + \tau_k d^k$$
such that
$$f(x^{k+1}) = \min_{\tau \geq 0} f(x^k + \tau d^k)$$
5. *Step 4:* increase k by 1 and go to *Step 1*.

Now, if f is not quadratic but just a twice continuously differentiable function. We need to reinitialize at some iteration k, theoretically, $k = n, 2n, 3n, ...$ But in practice, we usually restart after each $15 \sim 20$ steps. Much more relevant is the analysis of the first k-steps of the method for small k. It turns out the location of eigenvalues of the Hessian has a decisive impact on the quality of early iteration of conjugate gradient method.

5.3.2 Rate of convergence

We know that x^{k+1} is the minimum of $x^1 + \text{lin}\ \{d^1, d^2, ..., d^k\}$ which is equivalent to $x^1 + \text{lin}\ \{\nabla f(x^1), \nabla f(x^2), ..., \nabla f(x^k)\}$, or

$$x^1 + \text{lin}\ \{\nabla f(x^1) - \nabla f(x^2), ..., \nabla f(x^k) - \nabla f(x^{k-1})\}$$

Observe that $\nabla f(x^1) - \nabla f(x^2) = Q(x^2 - x^1)$ and it is proportional to $Q\nabla f(x^1)$, also

$$\nabla f(x^3) - \nabla f(x^2) = Q(x^3 - x^2) = \tau_2 d^2$$

where d^2 is a linear combination of $\nabla f(x^1)$ and $Q\nabla f(x^1)$, thus, equivalent to span $\{Q\nabla f(x^1), Q^2\nabla f(x^1)\}$. In general, we have

$$x^1 + \text{lin}\ \{\nabla f(x^1), Q\nabla f(x^1), Q^2\nabla f(x^1), ..., Q^{k-1}\nabla f(x^1)\}$$

Define $g = \nabla f(x^1)$, that is, $S_{k-1} = \text{lin}\ \{g, Qg, Q^2g, ..., Q^{k-1}g\}$. Assume Q has k distinct eigenvalues, $1 \leq k \leq n$, then

$$g = \sum_{j=1}^{k} \theta_j z^j$$

where z^j are eigenvectors of λ_j's. Thus,

$$Qg = \sum_{j=1}^{k} \theta_j \lambda_j z^j \in \text{lin}\ \{z^1, ..., z^k\}$$

Indeed,

$$S_m \subset \text{lin}\ \{z^1, ..., z^k\},\ m = 1, 2, ...$$

No more than k distinct nested subspaces can have this property, and therefore, $S_{k+1} = S_k$. As $\nabla f(x^{k+1}) \perp S_k$, it implies that $\nabla f(x^{k+1}) = 0$.

5.3.3 Pre-conditioning

The speed of convergence of the method of steepest descent is determined by the condition index:

$$\chi = \frac{\lambda_{max}}{\lambda_{min}}$$

The conjugate gradient method is also negatively affected by the condition index. Suppose we know a symmetric positive definite matrix V which is a good approximation of the inverse of the Hessian matrix,

$$V \approx Q^{-1}$$

Let $V^{\frac{1}{2}}$. Change the variable:

$$x = V^{\frac{1}{2}} y$$

we can express $f(x) = \frac{1}{2}\langle x, Qx\rangle + \langle c, x\rangle$ as

$$h(y) = f(V^{\frac{1}{2}} y) = \langle y, V^{\frac{1}{2}} Q V^{\frac{1}{2}} y\rangle + \langle c, V^{\frac{1}{2}} y\rangle$$

The Hessian of the function in the new coordinates equals:

$$\nabla^2 h(y) = V^{\frac{1}{2}} Q V^{\frac{1}{2}} \tag{5.12}$$

If V were exactly equal to Q^{-1}, we would obtain $\nabla^2 h(y) = I$, and the condition index would be perfectly 1. But knowing Q^{-1} means knowing the minimum $\hat{x} = -Q^{-1}c$, and there is no need for optimization at all. We cannot assume that we know the inverse of the Hessian. Instead, we assume that we can find a matrix V which is close to Q^{-1} in the sense that the condition index of $V^{\frac{1}{2}} Q V^{\frac{1}{2}}$ is much closer to 1 than that of Q. The change of variable with such a matrix V is called *pre-conditioning*.

Let's denote by r^k the direction used by the conjugate gradient method for (5.12) at iteration k. Applying conjugate gradient method yields the relation

$$r^k = \begin{cases} -\nabla h(y^k), & \text{if } k = 1, \\ -\nabla h(y^k) + a_k r^{k-1}, & \text{if } k > 1. \end{cases}$$

where

$$a_k = \frac{\langle \nabla h(y^k), \nabla h(y^k) - \nabla h(y^{k-1})\rangle}{||\nabla h(y^{k-1})||^2}$$

also,

$$y^{k+1} = y^k + \tau_k r^k, \;\; h(y^{k+1}) = \min_{\tau \geq 0} h(y^k + \tau r^k)$$

Define $x^k = V^{\frac{1}{2}} y^k$, we know

$$\nabla h(y^k) = V^{\frac{1}{2}} \nabla f(x^k)$$

then

$$d^k = V^{\frac{1}{2}} r^k$$

Thus,

$$d^k = \begin{cases} -V\nabla f(x^k), & \text{if } k = 1, \\ -\nabla f(x^k) + a_k d^{k-1}, & \text{if } k > 1. \end{cases}$$

Furthermore,

$$a_k = \frac{\langle \nabla f(x^k), V[\nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\langle \nabla f(x^{k-1}), V\nabla f(x^{k-1})\rangle}$$

The directional minimization condition for y becomes the direction minimization condition for x,

$$f(x^{k+1}) = \min_{\tau \geq 0} f(x^k + \tau d^k)$$

In fact, we do not need to have any explicit form of the pre-conditioner V. The only requirement is to be able to multiply vectors by V.

The conjugate gradient method with pre-conditioning can be formally applied to any continuously differentiable function. Of course, the concept of Q-orthogonality makes little sense there. Nevertheless, good convergence properties of the method for quadratic functions are indicative for the performance in the non-quadratic case.

Chapter 6

Constrained Optimization

In this chapter, we will focus on the set-constrained problem:

$$\min_{x \in X} \ f(x), \tag{6.1}$$

with a continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$, and with a convex closed set $X \subset \mathbb{R}^n$.

6.1 The reduced Gradient Method

We consider non-linear optimization problem with linear constraints:

$$\begin{aligned} \min . \ & f(x) && (6.2)\\ \text{subject to: } & Ax = b, && (6.3)\\ & x \geq 0 && (6.4) \end{aligned}$$

Here $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a continuously differentiable function. The matrix A of dimension $m \times n$, and the vector $b \in \mathbb{R}^m$ are given. We also assume that the rank of A is m.

The idea of *reduced conjugate gradient method* is to generate a sequence of feasible points $\{x^k\}$ by moving within the facets of the feasible polyhedron of 6.2. At current x^k we split the decision vector x into three sub-vectors:

1. *non-basic variables x_N:* they are assumed to be fixed at 0;
2. *super-basic variables x_S:* they are non-negative and are considered as independent variables; and
3. *basic variables x_B:* their values are calculated to satisfy the equation $Ax = b$.

The partition of x into these three sub-vectors corresponds to the division of A into three sub-matrices: N, S and B. The choice of the sub-vector of basic variables should be such that the matrix B is square and has rank m. We use the index sets: I_B, I_S and I_N, to denote the indices of basic, super-basic, and non-basic variables, correspondingly. After rearranging the components of x and columns of A in such a way that the components (columns) corresponding to I_B come first, those corresponding to I_S after them, and the components of columns associated with I_N last, we may write:

$$x = [x_B \; x_S \; x_N]^\top, \; A = [B \; S \; N]. \tag{6.5}$$

The equality constraints take on the form:

$$Bx_B + Sx_S + Nx_N = b,$$

By fixing the non-basic variables at zero, we choose a facet of the feasible set of problem (6.2) defined by the relations:

$$x_B = B^{-1}(b - S_{X_S}), \tag{6.6}$$
$$x_B \geq 0, \tag{6.7}$$
$$x_S \geq 0, \tag{6.8}$$
$$x_N = 0 \tag{6.9}$$

We shall try to minimize $f(\cdot)$ within this facet, by treating x_S as independent variables, and x_B as dependent variables determined via (6.6). Using the convention that

$$f(x) = f(x_B, x_S, x_N)$$

We can represent the objective function within this facet as

$$\psi(x_S) = f(B^{-1}(b - Sx_S), x_S, 0).$$

Let us denote by $\nabla_{x_B} f(x)$, $\nabla_{x_S} f(x)$ and $\nabla_{x_N} f(x)$ the three sub-vectors of the gradient of $f(\cdot)$ corresponding to the partition of the vectors x. The gradient of $\psi(\cdot)$ can now be calculated by chain rules of multivariate calculus:

$$\nabla\psi(x_S) = \nabla_{x_S} f(x) - S^\top [B^{-1}]^\top \nabla_{x_B} f(x)$$

We do not need to evaluate the inverse of the matrix B to calculate the gradient, rather, we just need to solve the system of equations:

$$B^\top \pi = \nabla_{x_B} f(x)$$

and we set

$$\nabla\psi(x_S) = \nabla_{x_S} f(x) - S^\top \pi$$

The vector $\nabla\psi(x_S)$ is called the *reduced gradient* of the objective function (6.2).

Since we have an easy way to calculate the gradient $\psi(\cdot)$, we can apply an efficient unconstrained optimization method to the problem:

$$\min \ \psi(x_S) \tag{6.10}$$

To be specific, let us assume that it is the conjugate gradient method. Of course, the minimization of $f(\cdot)$ over the facet is not equivalent to problem (6.10), because we have ignored the inequality constraints. Therefore, we introduce to the conjugate gradient method a modification within the step size selection rule (the second step used to be that we just optimize over τ, but now we should have more restrictions).

Given a direction of descent d_S^k for $\psi(\cdot)$ in the space of super-basic variables,

$$d_B^k = x_B^{new} - x_B = B^{-1}(b - S(x_S + d_S^k)) - B^{-1}(b - Sx_S) = -B^{-1}Sd_S^k.$$

The non-basic variables remain fixed at $x_N^k = 0$, thus the resulting direction of change of the vector x is

$$d^k = [d_B^k \ d_S^k \ 0]$$

The step size value τ_k is the solution of the problem:

$$\begin{aligned} \min . \ & f(x_B^k + \tau d_B^k, x_S^k + \tau d_S^k, x_N^k) \\ \text{subjec to: } & x_B^k + \tau d_B^k \geq 0, \\ & x_S^k + \tau d_S^k \geq 0, \\ & \tau \geq 0 \end{aligned} \tag{6.11}$$

Consider the point $x^{k+1} = x^k + \tau_k d^k$ obtained. If

$$\langle \nabla f(x^{k+1}), d^k \rangle = 0,$$

the point x^{k+1} is the same as in an unconstrained version of (6.11). In this case, we can continue the operation of the conjugate gradient method in the subspace of the super-basic variables x_S. If some of the bounds in (6.11) are active, that is

$$\langle \nabla f(x^{k+1}), d^k \rangle < 0,$$

we cannot continue with the unconstrained optimization algorithm in the same sub-space (conjugate gradient require new $\nabla f(x^{k+1})$ orthogonal to the manifold generated by previous direction). We then select a basic or super-basic variable x, that hit its bound. It can be identified by the conditions:

$$x_r^{k+1} = 0 \text{ and } d_r^k < 0. \tag{6.12}$$

The variable x_r is reclassified to the set of non-basic variables. If $r \in I_S$ then we simply set:

$$I_S^{new} := I_S \setminus \{r\}, \ I_N^{new} := I_N \cup \{r\},$$

and we continue. If x_r is a basic variable, we move the index r from the set I_B to the set I_N of non-basic variables. In this case, however, we also need augment the et of basic variables, in order to keep basis matrix B square and non-singular. To this end, we choose among the super-variables a variable x_e to be reclassified to the set of basic variables:

$$I_B^{new} := I_B \setminus \{r\} \cup \{e\},\ I_S^{new} := I_S \setminus \{e\},\ I_N^{new} := I_N \cup \{r\}.$$

The new basis matrix will have column $\{a^i : i \in I_B^{new}\}$. We must ensure that it is non-singular.

Lemma 6.1.1 Suppose (6.12) holds, true for a basic variable x_r. Then we can find a super-basic variable x_e such that the column a^i, $i \in I_B^{new}$, are linearly independent.

It follows that it is possible to introduce into the basis matrix a column of S and keep the basis matrix non-singular. The choice of a particular column, and testing whether its introduction yields a non-singular matrix, can be done by specialized techniques of numerical linear algebra, which are well established in linear programming. Usually, the current matrix B is maintained in a factorized form:

$$B = LU$$

with some lower-triangular matrix L and upper triangular matrix U. An exchange of a column of B is equivalent to an exchange of the corresponding U. Then, by dedicated re-factorization techniques, U is brought back to an upper triangular form, which change Las well. If this process is successful and yields non-singular factors

$$B^{new} = L^{new}U^{new}$$

the update is completed. Otherwise, another column of S has to be tried. Above *lemma* guarantees that a successful update will eventually occur.

Returning to the reduced gradient method, we notice that every encounter with a non-negative bound for basic or super-basic variables results in an increase of cardinality of the set of non-basic and a decrease of the cardinality of the set of super-basic variables. After such a change, the conjugate gradient method has to be re-initialized in the new facet to of the feasible set.

Since the cardinality of the set of super-basic variables can be decreased only finitely many times, after finitely many iterations it remains unchanged. As the feasible set is compact, every accumulation point of the sequence generated by the conjugate gradient method with re-initialization satisfies the necessary condition of local minimum of unconstrained optimization on $\psi(x_S)$, associated with the last classification into basic, super-basic, and non-basic variables. In practical computations an approximation to such a point will be found, but we are not going to introduce these technical details into our analysis. The mechanism of the method is best explained if we assume that a point $\hat{x}_S$ is found such that

$$\nabla\psi(\hat{x}_S) = 0.$$

The corresponding values of basic variables are

$$\hat{x}_B = B^{-1}(b - S\hat{x}_S),$$

and the non-basic variables are $\hat{x}_N = 0$. If the last partition has no super-basic variables at all, we set $\hat{x}_B = B^{-1}b$. We shall call the point $\hat{x} = (\hat{x}_B, \hat{x}_S, \hat{x}_N)$ a *semi-stationary point*.

The question to be addressed now is whether a semi-stationary point satisfies optimality condition of a local minimum of the original problem. It can be answered by analysing the vector

$$\bar{g}_N = \nabla_{x_N} f(\hat{x}) - N^\top \hat{\pi},$$

where $\hat{\pi}$ calculated at $\hat{x}$ by

$$B^\top \hat{\pi} = \nabla_{x_B} f(\hat{x}).$$

Lemma 6.1.2 If $\bar{g}_N \geq 0$, then the point $\hat{x} = (\hat{x}_B, \hat{x}_S, \hat{x}_N)$ satisfies the necessary condition of optimality for original problem.

Remark 6.1.3 There are further technicality in this method, we refer the reader to the book.

6.2 Penalty Method

6.2.1 general idea

The idea of *penalty methods* is to approximate a constrained optimization problem by an unconstrained optimization problem or by a problem with simple constraints. Consider problem

$$\min_{x \in X \cap X_0} \; f(x) \tag{6.13}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$, and X and X_0 are closed subsets of $\mathbb{R}^n$. We represent the feasible set as an intersection of two sets to allow treating "easy" constraints $x \in X_0$ directly.

We construct a continuous function $P : \mathbb{R}^n \mapsto \mathbb{R}$ associated with the set X that has following two properties:

$$P(x) \begin{cases} = 0, & \text{if } x \in X; \\ > 0, & \text{if } x \notin X. \end{cases}$$

We call this *penalty function*. It can be used to formulate an auxiliary problem with simple constraints:

$$\min_{x \in X_0} \; [\psi_\rho(x) := f(x) + \rho P(x)] \tag{6.14}$$

Here $\rho > 0$ is called *penalty parameter*. The idea of problem (6.14) is that the term $\rho P(\cdot)$, which is added to the objective function, introduces a "penalty" for violating the constraints $x \in X$. We hope that if ρ is sufficiently large, the solution of (6.14) should be close to a solution of (6.13).

Lemma 6.2.1 If problem (6.14) for some $\rho \geq 0$ has a solution x^* which is an element of the set X, then x^* is an optimal solution of problem (6.13).

Proof. As x^* solves (6.14), for every $x \in X_0 \cap X$ we have

$$f(x^*) + \rho P(x^*) \leq f(x) + \rho P(x)$$

Since both x^* and x are feasible, $P(x^*) = P(x) = 0$. Hence x^* is optimal. □

In general, we cannot expect to obtain a feasible minimizer of (6.14) for a finite value of the penalty parameter. Usually, we consider a sequence of problems (6.14), where the penalty parameter ρ is increased to $+\infty$. Still, under fairly general conditions, convergence to a solution of (6.13) occurs.

Theorem 6.2.2 Assume that problem (6.13) has an optimal solution. Let $\rho_k \to \infty$, as $k \to \infty$, and assume that problem (6.14) has a solution x^k for $\rho = \rho_k$. Then every accumulation point of sequence $\{x^k\}$ is an optimal solution of problem (6.13).

Proof. Suppose $\hat{x}$ is optimal solution of (6.13). As it is feasible for (6.14),

$$f(x^k) + \rho_k P(x^k) \leq f(\hat{x}) + \rho P(\hat{x}) = f(\hat{x}),\ k = 1, 2, ... \tag{6.15}$$

Hence,

$$P(x^k) \leq \frac{f(\hat{x}) - f(x^k)}{\rho_k}$$

Consider a convergent subsequence $\{x^k\}$, $k \in \mathcal{K}$. Let x^∞ be its limit. Passing to the limit with $k \to \infty$, $k \in \mathcal{K}$, in the last inequality we conclude that $P(x^k) \to 0$ as $k \to \infty$, $k \in \mathcal{K}$. It follows from the continuity of $P(\cdot)$ that $P(x^\infty) = 0$, and thus $x^\infty \in X$.

Furthermore, inequality (6.15) implies that

$$f(x^\infty) = \lim_{k \to \infty, k \in \mathcal{K}} f(x^k) \leq \lim_{k \to \infty, k \in \mathcal{K}} [f(x^k) + \rho_k P(x^k)] \leq f(\hat{x})$$

The point x^∞ is an optimal solution of (6.13). □

To apply this theorem, we need additional conditions guaranteeing that the sequence $\{x^k\}$ indeed has accumulation points. The easiest is the condition that the set X_0 is compact. More generally, it is sufficient that a feasible point x^0 such that the set $\{x \in X_0 : f(x) \leq f(x^0)\}$ is bounded.

6.2.2 Quadratic Penalty

We can now apply the general ideas of penalty methods to the nonlinear optimization problem:

$$\begin{aligned} \min \ & f(x) \\ \text{subjec to: } & g_i(x) \leq 0, \ i = 1, ..., m, \\ & h_i(x) = 0, \ i = 1, ..., p, \\ & x \in X_0. \end{aligned} \tag{6.16}$$

All functions $f : \mathbb{R}^n \mapsto \mathbb{R}$, $g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., m$, and $h_i : \setminus^n \mapsto \mathbb{R}$, $i = 1, ..., p$, are assumed to be continuously differentiable. The set X_0 is convex and closed.

The following function

$$P_2(x) = \frac{1}{2} \sum_{i=1}^{m} [\max(0, g_i(x))]^2 + \frac{1}{2} \sum_{i=1}^{p} [h_i(x)]^2$$

is called the *quadratic penalty function* for problem (6.19). For the set X defined by the inequality and equality constraints of problem (6.19), the function $P_2(\cdot)$ satisfies the general conditions of the penalty function. Thus, under the condition that X_0 is compact, *theorem* in the above section can be applied.

The main advantage of the penalty function $P_2(\cdot)$ is that it is continuously differentiable, as a composition of the continuously differentiable functions $g_i(\cdot)$ and $h_i(\cdot)$ with function $t \mapsto [\max(0, t)]^2$ and $t \mapsto t^2$, which are continuously differentiable as well. Consider the sequence of problems:

$$\min_{x \in X_0} \ \{\psi_{\rho_k}(x) := f(x) + \rho_k P_2(x)\} \tag{6.17}$$

with $\rho_k \to \infty$. The corresponding solutions of problem (6.17) are denoted by x^k. If the sequence $\{x^k\}$ is bounded, then it has accumulation points, and theorem above implies that each accumulation point is a solution of problem (6.19). Therefore, to focus attention, we can assume that the entire sequence $\{x^k\}$ is convergent to some solution $\hat{x}$ of problem (6.19).

Let us recall *Robinson's constraint qualification condition* for problem (6.19). It is convenient to introduce the set $I^0(\hat{x})$ of active inequality constraints:

$$I^0(\hat{x}) = \{1 \leq i \leq m : g_i(\hat{x}) = 0\}$$

Now, we consider the mapping $\bar{g}(x)$ with coordinates equal to $g_i(x)$, $i \in I^0(\hat{x})$. *Robinson's condition* takes on the form:

$$0 \in \text{int}\{[\bar{g}'(\hat{x})d + y \ h'(\hat{x})d]^\top : d \in K_{X_0}(\hat{x}), y \geq 0\}.$$

The set $\bar{K}_{X_0}(\hat{x})$ is the set of feasible directions for X_0 at $\hat{x}$. We know *Robinson's condition* guarantees the existence of *Lagrangian multipliers* at the point $\hat{x}$: vector $\hat{\lambda} \in \mathbb{R}^m_+$ and $\hat{\mu} \in \mathbb{R}^p$

such that

$$0 \in \nabla f(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{x}) + N_{X_0}(\hat{x})$$

Moreover, $\hat{\lambda}_i g(\hat{x}) = 0,\ i = 1, ..., m$. It turns out that the quadratic penalty method can approximate the *Lagrange multipliers* very well.

Theorem 6.2.3 Assume that $\rho_k \to \infty$,as $k \to \infty$, and that the sequence $\{x^k\}$ of solutions of problem (6.17) is convergent to some solution $\hat{x}$ of problem (6.19). Furthermore, assume that *Robinson's condition* is satisfied at $\hat{x}$. Then the sequences

$$\begin{aligned} \lambda_i^k &= \rho_k \max(0, g_i(x^k)),\ i = 1, ..., m \\ \mu_i^k &= \rho_k h_i(x^k),\ i = 1, ..., p, \end{aligned}$$

are bounded, and each accumulation point $(\hat{\lambda}, \hat{\mu})$ of the sequence $\{(\lambda^k, \mu^k)\}$ is the vector of *Lagrangian multipliers* satisfying the necessary condition of optimality at $\hat{x}$.

Remark 6.2.4 This is an interesting observation that the values of the Lagrangian multipliers are obtained as a by-product of a constrained optimization method. A disadvantage of the quadratic penalty method is the increased difficulty of problem (6.17) for large values of the penalty parameter ρ. Because as ρ increases, the condition index will tend to ∞ making the problem very ill-conditioned.

6.3 Dual Method

We still consider the optimization problem of the following form:

$$\begin{aligned} \min\ & f(x) \\ \text{subjec to: } & g_i(x) \le 0,\ i = 1, ..., m, \\ & h_i(x) = 0,\ i = 1, ..., p, \\ & x \in X_0. \end{aligned} \tag{6.18}$$

all functions $f : \mathbb{R}^n \mapsto \mathbb{R}$, $g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., m$, and $h_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., p$, are assumed to be continuously differentiable. The set X_0 is convex and closed.

We consider the *Lagrangian*,

$$L(x, \lambda, \mu) = f(x) + \langle \lambda g(x) \rangle + \langle \mu, h(x) \rangle$$

and the dual function:

$$L_D(\lambda, \mu) = \inf_{x \in X_0} L(x, \lambda, \mu)$$

The dual function is defined as follows:

$$\max_{(\lambda,\mu)\in\Lambda_0} \; L_D(\lambda,\mu)$$

with $\Lambda_0 = \mathbb{R}^m_+ \times \mathbb{R}^p$.

Assume that the function $f(\cdot)$ and $g_i(\cdot)$, $i = 1, ..., m$ are convex, the function $h_i(\cdot)$ are affine, and *Slater's condition* is satisfied. Then, according to theorem in chapter 4, problem (6.19) has an optimal solution if and only if the dual problem has an optimal solution. Also, the solutions of the dual problem are the *lagrangian multipliers* satisfying the necessary and sufficient conditions of optimality for the primal problem, at each of its optimal solutions.

The idea of the dual method is to solve the dual problem by an iterative method for optimization with simple constraints, and then to recover the primal solution (theorem in chapter 4).

We focus on the case that $L_D(\cdot)$ is continuously differentiable. Also, we make the following assumptions:

1. The set X_0 is bounded;
2. The function $f(\cdot)$ is strictly convex.

These conditions ensure that for every (λ, μ) a solution $\hat{x}(\lambda, \mu)$ of dual problem exists and is unique. Furthermore, we know that the dual function is continuously differentiable, and its gradient is given by

$$L_D(\lambda,\mu) = [g(\hat{x}(\lambda,\mu))\ h(\hat{x}(\lambda,\mu))]^\top$$

Thus, we can apply to dual problem any method for problems with simple constraints.

Let's consider the projection method. Since

$$\Pi_{\Lambda_0}(\lambda,\mu) = (\max(0,\lambda),\mu)$$

the projection method takes on the form

$$\begin{aligned}\lambda^{k+1} &= \max\left(0, \lambda^k + \tau g(\hat{x}(\lambda^k,\mu^k))\right),\\ \mu^{k+1} &= \mu^k + \tau h(\hat{x}(\lambda^k,\mu^k)),\ k = 1, 2, ...\end{aligned}$$

Note that the dual problem is a maximization problem and therefore we make steps in the direction of the gradient, rather than its negative. Unfortunately, it is hard to determine the right value of the step size τ in this method.

We can also apply the reduced gradient method to the dual problem. It has a very simple form: some of the multipliers λ_i are non-basic and fixed at 0, and the others are super-basic

and are free to move. There are no-basic variables at all, and the reduced gradient is just the gradient of the dual function with respect to the super-=basic multipliers. Here the difficulty is that the directional maximization requires the evaluation of the dual function at many points, and this maybe expensive.

6.4 Augmented Lagrangian Method

Again, we consider

$$\begin{aligned} \min \ & f(x) \\ \text{subjec to: } & g_i(x) \leq 0, \ i = 1, ..., m, \\ & h_i(x) = 0, \ i = 1, ..., p, \\ & x \in X_0. \end{aligned} \tag{6.19}$$

The functions $f : \mathbb{R}^n \mapsto \mathbb{R}$, $g_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., m$, and $h_i : \mathbb{R}^n \mapsto \mathbb{R}$, $i = 1, ..., p$, are assumed to be twice continuously differentiable. The set X_0 is convex and closed.

The *augmented Lagrangian* for the problem (6.19) has the following form:

$$\begin{aligned} L_\rho(x, \lambda, \mu) = f(x) + \frac{\rho}{2} \sum_{i=1}^{m} [\max(0, g_i(x) + \frac{\lambda_i}{\rho})]^2 \\ + \frac{\rho}{2} \sum_{i=1}^{p} [h_i(x) + \frac{\mu_i}{\rho}]^2 - \frac{1}{2\rho} \sum_{i=1}^{m} \lambda_i^2 - \frac{1}{2\rho} \sum_{i=1}^{p} \mu_i^2 \end{aligned} \tag{6.20}$$

Assume that $\hat{x}$ is a local minimum of problem (6.19) satisfying *Robinson's condition*. We also assume that the strong second order sufficient condition is satisfied with *Lagrange multipliers* $(\hat{\lambda}, \hat{\mu})$. By previous theorem, there exists ρ_0 such that for all $\rho > \rho_0$ the pair $\hat{x}$ and $(\hat{\lambda}, \hat{\mu})$ is a local saddle point of the augmented Lagrangian. If we knew the optimal Lagrange multipliers $(\hat{\lambda}, \hat{\mu})$, we could, for a sufficiently large ρ, minimize the augmented Lagrangian to obtain the local minimum $\hat{x}$. However, the multipliers are not known, and need to be estimated.

The theorem in the penalty method section gives us a way to estimate it. Suppose we minimize the augmented Lagrangian with respect to x in a certain neighbourhood U of the optimal point, with given values of multipliers (λ^k, μ^k). If (λ^k, μ^k) are close to the optimal values $(\hat{\lambda}, \hat{\mu})$, the minimum of the augmented *Lagrangian* will be close to $\hat{x}$. The terms corresponding to $i \notin I^0(\hat{x})$ become 0, and in the neighbourhood of $\hat{x}$ the augmented

Lagrangian simplifies:

$$\begin{aligned} L_\rho(x,\lambda^k,\mu^k) &= f(x) + \frac{\rho}{2}\sum_{i\in I^0(\hat{x})}[g_i(x)+\frac{\lambda_i^k}{\rho}]^2 + \frac{\rho}{2}\sum_{i=1}^{p}[h_i(x)+\frac{\mu_i^k}{\rho}]^2 \\ &- \frac{1}{2\rho}\sum_{i=1}^{m}(\lambda_i^k)^2 - \frac{1}{2\rho}\sum_{i=1}^{p}(\mu_i^k)^2 \\ &= f(x) + \sum_{i\in I^0(\hat{x})}\left(\lambda_i^k g_i(x) + \frac{\rho}{2}[g_i(x)]^2\right) \\ &+ \sum_{i=1}^{p}\left(\mu_i^k h_i(x) + \frac{\rho}{2}[h_i(x)]^2\right) - \frac{1}{2\rho}\sum_{i\in I^0(\hat{x})}(\lambda_i^k)^2 \end{aligned} \tag{6.21}$$

This is equal (after a translation by the last sum) to the quadratic penalty function for the problem:

$$\begin{aligned} \min\ & f(x) + \sum_{i\in I^0(\hat{x})}\lambda_i^k g_i(x) + \sum_{i=1}^{p}\mu_i^k h_i(x) \\ \text{subject to: } & g_i(x) = 9,\ i\in I^0(\hat{x}), \\ & h_i(x) = 0,\ i = 1,...,p, \\ & x\in X_0 \end{aligned} \tag{6.22}$$

We know that at the minimum x^k of the quadratic penalty function, the quantities $\rho g_i(x^k)$, $i\in I^0(\hat{x})$, and $\rho h_i(x^k)$, $i=1,...,p$, become very close to the Lagrange multipliers associated with the constraints of problem (6.22). Since the gradient of the objective function of (6.22) has the form:

$$\nabla f(x) + \sum_{i\in I^0(\hat{x})}\lambda_i^k\nabla g_i(x) + \sum_{i=1}^{p}\mu_i^k\nabla h_i(x)$$

We conclude that $\lambda_i^k + \rho g_i(x^k)$ and $\mu_i^k + \rho h_i(x^k)$ are good approximations of the Lagrange multipliers in the original problem (6.19). As $\hat{\lambda}\geq 0$, we mayu take the positive part of $\lambda_i^k + \rho g_i(x^k)$.

We construct, therefore, the following iterative process. At iteration k, for given values of Lagrange multipliers $(\lambda^k,\mu^k)\in\Lambda_0$, we solve the problem

$$\min_{x\in X_0}\ L_\rho(x,\lambda^k,\mu^k)$$

After obtaining the solution x^k, we update the multipliers by the formula:

$$\begin{aligned} \lambda_i^{k+1} &= \max(0,\lambda_i^k + \rho g_i(x^k)),\ i=1,...,m, \\ \mu_i^{k+1} &= \mu_i^k + \rho h_i(x^k),\ i=1,...,p, \end{aligned}$$

and the iteration continues. This is called the *augmented Lagrangian method or the method of multipliers*.

The multiplier method is closely related to the duality theory for augmented Lagrangian. By the strong second order condition, if $\rho > \rho_0$, the function $L_\rho(\cdot, \lambda, \mu)$ is locally strictly convex about $\hat{x}$, if (λ, μ) is close to $(\hat{\lambda}, \hat{\mu})$. We can thus consider the local dual function

$$L_{\rho\ D}(\lambda, \mu) = min_{x \in X_0 \cap U} L_\rho(x, \lambda, \mu),$$

where U is a small neighbourhood of $\hat{x}$. The minimizer is unique and it follows from that

$$\nabla_\lambda L_\rho(\lambda^k, \mu^k) = g(x^k),$$
$$\nabla_\mu L_\rho(\lambda^k, \mu^k) = h(x^k).$$

Thus the method of multipliers may be regarded as a version of the method of steepest ascent with projection applied to the dual problem. An interesting observation is that the penalty parameter ρ is a good value of the step size in this method.