Introduction to Optimization Theory

Lecture Notes

JIANFEI SHEN

SCHOOL OF ECONOMICS
SHANDONG UNIVERSITY
Besides language and music, mathematics is one of the primary manifestations of the free creative power of the human mind.

— Hermann Weyl
# CONTENTS

## 1 Multivariable Calculus
- 1.1 Functions on Euclidean Space ........................................... 2
- 1.2 Directional Derivative and Derivative .................................. 8
- 1.3 Partial Derivatives and the Jacobian .................................... 13
- 1.4 The Chain Rule .............................................................. 13
- 1.5 The Implicit Function Theorem .......................................... 16
- 1.6 Gradient and Its Properties .............................................. 19
- 1.7 Continuously Differentiable Functions .................................. 25
- 1.8 Quadratic Forms: Definite and Semidefinite Matrices ............... 27
- 1.9 Homogeneous Functions and Euler’s Formula ....................... 28

## 2 Optimization in $\mathbb{R}^n$
- 2.1 Introduction ................................................................. 31
- 2.2 Unconstrained Optimization ............................................. 33
- 2.3 Equality Constrained Optimization: Lagrange’s Method .......... 35
- 2.4 Inequality Constrained Optimization: Kuhn-Tucker Theorem ... 42
- 2.5 Envelop Theorem ........................................................... 49

## 3 Convex Analysis in $\mathbb{R}^n$
- 3.1 Convex Sets ................................................................. 55
- 3.2 Separation Theorem ........................................................ 57
- 3.3 Systems of Linear Inequalities: Theorems of the Alternative .... 63
- 3.4 Convex Functions .......................................................... 65
- 3.5 Convexity and Optimization .............................................. 70

**References** ................................................................. 75

**Index** ............................................................ 78
In this chapter we consider functions mapping $\mathbb{R}^m$ into $\mathbb{R}^n$, and we define what we mean by the derivative of such a function. It is important to be familiar with the idea that the derivative at a point $a$ of a map between open sets of (normed) vector spaces is a linear transformation between the vector spaces (in this chapter the linear transformation is represented as a $n \times m$ matrix).

This chapter is based on Spivak (1965, Chapters 1 & 2) and Munkres (1991, Chapter 2)—one could do no better than to study theses two excellent books for multivariable calculus.

Notation

We use standard notation:

$\mathbb{N} := \{1, 2, 3, \ldots\}$ = the set of all natural numbers.
$\mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$ = the set of all integers.
$\mathbb{Q} := \left\{ \frac{n}{m} : n, m \in \mathbb{Z} \text{ and } m \neq 0 \right\}$ = the set of all rational numbers,
$\mathbb{R} :=$ the set of all real numbers.

We also define

$\mathbb{R}_+ := \{ x \in \mathbb{R} : x \geq 0 \}$ and $\mathbb{R}_{++} := \{ x \in \mathbb{R} : x > 0 \}$. 
1.1 Functions on Euclidean Space

Norm, Inner Product and Metric

Definition 1.1 (Euclidean n-space) Euclidean n-space $\mathbb{R}^n$ is defined as the set of all $n$-tuples $(x_1, \ldots, x_n)$ of real numbers $x_i$:

$$\mathbb{R}^n := \{(x_1, \ldots, x_n) : \text{each } x_i \in \mathbb{R}\}.$$  

An element of $\mathbb{R}^n$ is often called a point in $\mathbb{R}^n$, and $\mathbb{R}^1$, $\mathbb{R}^2$, $\mathbb{R}^3$ are often called the line, the plane, and space, respectively.

If $x$ denotes an element of $\mathbb{R}^n$, then $x$ is an $n$-tuple of numbers, the $i^{th}$ one of which is denoted $x_i$; thus, we can write

$$x = (x_1, \ldots, x_n).$$

A point in $\mathbb{R}^n$ is frequently also called a vector in $\mathbb{R}^n$, because $\mathbb{R}^n$, with $x + y = (x_1 + y_1, \ldots, x_n + y_n)$, $x, y \in \mathbb{R}^n$ and $\alpha x = (\alpha x_1, \ldots, \alpha x_n)$, $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$, as operations, is a vector space.

To obtain the full geometric structure of $\mathbb{R}^n$, we introduce three structures on $\mathbb{R}^n$: the Euclidean norm, inner product and metric.

Definition 1.2 (Norm) In $\mathbb{R}^n$, the length of a vector $x \in \mathbb{R}^n$, usually called the norm $\|x\|$ of $x$, is defined by

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$  

Remark 1.3 The norm $\| \cdot \|$ satisfies the following properties: for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

1. $\|x\| \geq 0$,
2. $\|x\| = 0$ iff $x = 0$,
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$,
4. $\|x + y\| \leq \|x\| + \|y\|$ (Triangle inequality).

"iff" is the abbreviation of "if and only if".
Exercise 1.4  Prove that \(|x| - |y|| \leq |x - y|\) for any two vectors \(x, y \in \mathbb{R}^n\) (use the triangle inequality).

Definition 1.5 (Inner Product) Given \(x, y \in \mathbb{R}^n\), the inner product of the vectors \(x\) and \(y\), denoted \(x \cdot y\) or \(\langle x, y \rangle\), is defined as

\[ x \cdot y = \sum_{i=1}^{n} x_i y_i. \]

Remark 1.6 The norm and the inner product are related through the following identity:

\[ \|x\| = \sqrt{x \cdot x}. \]

Theorem 1.7 (Cauchy-Schwartz Inequality) For any \(x, y \in \mathbb{R}^n\) we have

\[ |x \cdot y| \leq \|x\| \|y\|. \]

Proof. We assume that \(x \neq 0\); for otherwise the proof is trivial. For every \(a \in \mathbb{R}\), we have

\[ 0 \leq \|ax + y\|^2 = a^2 \|x\|^2 + 2a(x \cdot y) + \|y\|^2. \]

In particular, let \(a = -(x \cdot y)/\|x\|^2\). Then, from the above display, we get the desired result. \(\square\)

Exercise 1.8 Prove the triangle inequality (use the Cauchy-Schwartz Inequality). Show it holds with equality iff one of the vectors is a nonnegative scalar multiple of the other.

Definition 1.9 (Metric) The distance \(d(x, y)\) between two vectors \(x, y \in \mathbb{R}^n\) is given by

\[ d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \]

The distance function \(d\) is called a metric.

Example 1.10 In \(\mathbb{R}^2\), choose two points \(x^1 = (x^1_1, x^1_2)\) and \(x^2 = (x^2_1, x^2_2)\) with \(x^2_1 - x^1_1 = a\) and \(x^2_2 - x^1_2 = b\). Then Pythagoras tells us that (Figure 1.1)

\[ d(x^1, x^2) = \sqrt{a^2 + b^2} = \sqrt{(x^2_1 - x^1_1)^2 + (x^2_2 - x^1_2)^2}. \]

Remark 1.11 The metric is related to the norm \(\|\cdot\|\) through the identity

\[ d(x, y) = \|x - y\|. \]
Subsets of $\mathbb{R}^n$

- **Definition 1.12** (Open Ball) Let $x \in \mathbb{R}^n$ and $r > 0$. The *open ball* $B(x; r)$ with center $x$ and radius $r$ is given by

$$B(x; r) := \{ y \in \mathbb{R}^n : d(x, y) < r \}.$$  

- **Definition 1.13** (Interior) Let $S \subset \mathbb{R}^n$. A point $x \in S$ is called an *interior point* of $S$ if there is some $r > 0$ such that $B(x; r) \subset S$. The set of all interior points of $S$ is called its *interior* and is denoted $S^o$.

- **Definition 1.14** Let $S \subset \mathbb{R}^n$.

  - $S$ is *open* if for every $x \in S$ there exists $r > 0$ such that $B(x; r) \subset S$.
  
  - $S$ is *closed* if its complement $\mathbb{R}^n \setminus S$ is open.
  
  - $S$ is *bounded* if there exists $r > 0$ such that $S \subset B(0; r)$.
  
  - $S$ is *compact* if (and only if) it is closed and bounded (Heine-Borel Theorem).\(^2\)

- **Example 1.15** On $\mathbb{R}$, the interval $(0, 1)$ is open, the interval $[0, 1]$ is closed. Both $(0, 1)$ and $[0, 1]$ are bounded, and $[0, 1]$ is compact. However, the interval $(0, 1]$ is neither open nor closed. But $\mathbb{R}$ is both open and closed.

\(^2\)This definition does not work for more general metric spaces. See Willard (2004) for details.
Limit and Continuity

**FUNCTIONS** A function from \( \mathbb{R}^m \) to \( \mathbb{R}^n \) (sometimes called a vector-valued function of \( m \) variables) is a rule which associates to each point in \( \mathbb{R}^m \) some point in \( \mathbb{R}^n \). We write

\[ f : \mathbb{R}^m \rightarrow \mathbb{R}^n \]

to indicate that \( f(x) \in \mathbb{R}^n \) is defined for \( x \in \mathbb{R}^m \).

The notation \( f : A \rightarrow \mathbb{R}^n \) indicates that \( f(x) \) is defined only for \( x \) in the set \( A \), which is called the domain of \( f \). If \( B \subset A \), we define \( f(B) \) as the set of all \( f(x) \) for \( x \in B \):

\[ f(B) := \{ f(x) : x \in B \} \]

If \( C \subset \mathbb{R}^n \) we define

\[ f^{-1}(C) := \{ x \in A : f(x) \in C \} \]

The notation \( f : A \rightarrow B \) indicates that \( f(A) \subset B \). The graph of \( f : A \rightarrow B \) is the set of all pairs \( (a, b) \in A \times B \) such that \( b = f(a) \).

A function \( f : A \rightarrow \mathbb{R}^n \) determines \( n \) component functions \( f_1, \ldots, f_n : A \rightarrow \mathbb{R} \) by

\[ f(x) = (f_1(x), \ldots, f_n(x)) \]

**SEQUENCES** A sequence is a function that assigns to each natural number \( n \in \mathbb{N} \) a vector or point \( x_n \in \mathbb{R}^n \). We usually write the sequences as \( (x_n)_{n=1}^{\infty} \) or \( (x_n) \).

**Example 1.16** Examples of sequences in \( \mathbb{R}^2 \) are

(a) \( (x_n) = ((n, n)) \).

(b) \( (x_n) = ((\cos \frac{n\pi}{2}, \sin \frac{n\pi}{2})) \).

(c) \( (x_n) = (((-1)^n/2^n, 1/2^n)) \).

(d) \( (x_n) = (((-1)^n - 1/n, (-1)^n - 1/n)) \).

See Figure 1.2.

**Definition 1.17** (Limit) A sequence \( (x_n) \) is said to have a limit \( x \) or to converge to \( x \) if for every \( \varepsilon > 0 \) there is \( N_\varepsilon \in \mathbb{N} \) such that whenever \( n > N_\varepsilon \), we have \( x_n \in \mathbb{B}(x; \varepsilon) \). We write

\[ \lim_{n \to \infty} x_n = x \quad \text{or} \quad x_n \to x. \]

**Example 1.18** In Example 1.16, the sequences (a), (b) and (d) do not converge, while the sequence (c) converges to \((0,0)\).
(a) \( (x_n) = ((n, n)) \)
(b) \( (x_n) = \left( \cos \frac{n\pi}{2}, \sin \frac{n\pi}{2} \right) \)
(c) \( (x_n) = \{(\frac{n\pi}{2}, 1/2^n)\} \), which is convergent
(d) \( (x_n) = \{((-1)^n - 1/n, (-1)^n - 1/n)\} \)

**Figure 1.2:** Examples of sequences
1.1 FUNCTIONS ON EUCLIDEAN SPACE

Naive Continuity

The simplest way to say that a function \( f : A \to \mathbb{R} \) is continuous is to say that one can draw its graph without taking the pencil off the paper. For example, a function whose graph looks like in Figure 1.3(a) would be continuous in this sense (Crossley, 2005, Chapter 2).

But if we look at the function \( f(x) = \frac{1}{x} \), then we see that things are not so simple. The graph of this function has two parts—one part corresponding to negative \( x \) values, and the other to positive \( x \) values. The function is not defined at 0, so we certainly cannot draw both parts of this graph without taking our pencil off the paper; see Figure 1.3(b). Of course, \( f(x) = \frac{1}{x} \) is continuous near every point in its domain. Such a function deserves to be called continuous. So this characterization of continuity in terms of graph-sketching is too simplistic.

Rigorous Continuity

The notation \( \lim_{x \to a} f(x) = b \) means, as in the one-variable case, that we get \( f(x) \) as close to \( b \) as desired, by choosing \( x \) sufficiently close to, but not equal to, \( a \). In mathematical terms this means that for every number \( \varepsilon > 0 \) there is a number \( \delta > 0 \) such that \( \| f(x) - b \| < \varepsilon \) for all \( x \) in the domain of \( f \) which satisfy \( 0 < \| x - a \| < \delta \).

A function \( f : A \to \mathbb{R}^n \) is called continuous at \( a \in A \) if \( \lim_{x \to a} f(x) = f(a) \), and \( f \) is continuous if it is continuous at each \( a \in A \).
Exercise 1.19  Let

\[ f(x) = \begin{cases} 
  x & \text{if } x \neq 1 \\
  3/2 & \text{if } x = 1.
\end{cases} \]

Show that \( f(x) \) is not continuous at \( a = 1 \).

1.2 Directional Derivative and Derivative

Let us first recall how the derivative of a real-valued function of a real variable is defined. Let \( A \subset \mathbb{R} \); let \( f : A \to \mathbb{R} \). Suppose \( A \) contains a neighborhood of the point \( a \), that is, there is an open ball \( B(a; r) \) such that \( B(a; r) \subset A \). We define the derivative of \( f \) at \( a \) by the equation

\[ f'(a) = \lim_{t \to 0} \frac{f(a + t) - f(a)}{t}, \]

(1.1)

provided the limit exists. In this case, we say that \( f \) is differentiable at \( a \). Geometrically, \( f'(a) \) is the slope of the tangent line to the graph of \( f \) at the point \( (a, f(a)) \).

Definition 1.20  For a function \( f : (a, b) \to \mathbb{R} \), and point \( x_0 \in (a, b) \), if

\[ \lim_{t \to 0} \frac{f(x_0 + t) - f(x_0)}{t} \]

exists and is finite, we denote this limit by \( f'(x_0) \) and call it the left-hand derivative of \( f \) at \( x_0 \). Similarly, we define \( f'(x_0) \) and call it the right-hand derivative of \( f \) at \( x_0 \). Of course, \( f \) is differentiable at \( x_0 \) iff it has left-hand and right-hand derivatives at \( x_0 \) that are equal.

Now let \( A \subset \mathbb{R}^m \), where \( m > 1 \); let \( f : A \to \mathbb{R}^n \). Can we define the derivative of \( f \) by replacing \( a \) and \( t \) in the definition just given by points of \( \mathbb{R}^m \)? Certainly we cannot since we cannot divide a point of \( \mathbb{R}^n \) by a point of \( \mathbb{R}^m \) if \( m > 1 \).

Directional Derivative

The following is our first attempt at a definition of “derivative”. Let \( A \subset \mathbb{R}^m \) and let \( f : A \to \mathbb{R}^n \). We study how \( f \) changes as we move from a point \( a \in A^\circ \) (the interior of \( A \)) along a line segment to a nearby point \( a + t u \), where \( u \neq 0 \). Each point on the segment can be expressed as \( a + t u \), where \( t \in \mathbb{R} \). The vector \( u \) describes the direction of the line segment. Since \( a \) is an interior point of \( A \), the line segment joining \( a \) to \( a + t u \) will lie in \( A \) if \( t \) is small enough.
Definition 1.21 (Directional Derivative) Let \( A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R}^n \). Suppose \( A \) contains a neighborhood of \( a \). Given \( u \in \mathbb{R}^m \) with \( u \neq 0 \), define
\[
f'(a; u) = \lim_{t \to 0} \frac{f(a + tu) - f(a)}{t},
\]
provided the limit exists. This limit is called the directional derivative of \( f \) at \( a \) with respect to the vector \( u \).

Remark 1.22 In calculus, one usually requires \( u \) to be a unit vector, i.e., \( \|u\| = 1 \), but that is not necessary.

Example 1.23 Let \( f : \mathbb{R}^2 \to \mathbb{R} \) be given by the equation \( f(x_1, x_2) = x_1 x_2 \). The directional derivative of \( f \) at \( a = (a_1, a_2) \) with respect to the vector \( u = (u_1, u_2) \) is
\[
f'(a; u) = \lim_{t \to 0} \frac{f(a + tu) - f(a)}{t} = \lim_{t \to 0} \frac{(a_1 + tu_1)(a_2 + tu_2) - a_1 a_2}{t}
= u_2 a_1 + u_1 a_2.
\]

Example 1.24 Suppose the directional derivative of \( f \) at \( a \) with respect to \( u \) exists. Then for \( c \in \mathbb{R} \),
\[
f'(a; cu) = \lim_{t \to 0} \frac{f(a + tu) - f(a)}{t} = c \lim_{t \to 0} \frac{f(a + tu) - f(a)}{tc}
= c \lim_{s \to 0} \frac{f(a + su) - f(a)}{s}
= cf'(a; u).
\]

Remark 1.25 Example 1.24 shows that if \( u \) and \( v \) are collinear vectors in \( \mathbb{R}^m \), then \( f'(a; u) \) and \( f'(a; v) \) are collinear in \( \mathbb{R}^n \).

However, directional derivative is not the appropriate generalization of the notion of “derivative”. The main problems are:

Problem 1. Continuity does not follow from this definition of “differentiability”. There exists functions such that \( f'(a; u) \) exists for all \( u \neq 0 \) but are not continuous.

Example 1.26 Define \( f : \mathbb{R}^2 \to \mathbb{R} \) by setting
\[
f(x, y) = \begin{cases} 
0 & \text{if } (x, y) = (0, 0) \\
\frac{x^2 y}{x^4 + y^2} & \text{if } (x, y) \neq (0, 0).
\end{cases}
\]
We show that all directional derivatives of \( f \) exist at \( 0 \), but that \( f \) is not continuous at \( 0 \). Let \( u = (h, k) \neq 0 \). Then

\[
\frac{f(0 + tu) - f(0)}{t} = \frac{(th)^2(tk)}{(th)^4 + (ik)^2} = \frac{h^2k}{t^2h^4 + k^2},
\]

so that

\[
f'(0; u) = \begin{cases} 
\frac{h^2}{k} & \text{if } k \neq 0, \\
0 & \text{if } k = 0.
\end{cases}
\]

However, the function \( f \) takes the value \( 1/2 \) as each point of the parabola \( y = x^2 \) (except at \( 0 \)), so \( f \) is not continuous at \( 0 \) since \( f(0) = 0 \).

**Problem 2.** Composites of “differentiable” functions may not differentiable.

**Derivative**

To give the right generalization of the notion of “derivative”, let us reconsider (1.1). In fact, if \( f'(a) \) exists, let \( R_a(t) \) denote the difference

\[
R_a(t) := \begin{cases} 
\frac{f(a + t) - f(a)}{t} - f'(a) & \text{if } t \neq 0 \\
0 & \text{if } t = 0.
\end{cases}
\]  

(1.2)

From (1.2) we see that \( \lim_{t \to 0} R_a(t) = 0 \). Then we have

\[
f(a + t) = f(a) + f'(a)t + R_a(t)t. 
\]

(1.3)

Note that (1.3) also holds for \( t = 0 \). This is called the first-order Taylor formula for approximating \( f(a + t) - f(a) \) by \( f'(a)t \). The error committed is \( R_a(t)t \). See Figure 1.4. It is this idea leads to the following definition:

**Definition 1.27** (Differentiability) Let \( A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R}^n \). Suppose \( A \) contains a neighborhood of \( a \). We say that \( f \) is differentiable at \( a \) if there is an \( n \times m \) matrix \( B_a \) such that

\[
f(a + h) = f(a) + B_a \cdot h + \|h\| R_a(h).
\]

where \( \lim_{h \to 0} R_a(h) = 0 \). The matrix \( B_a \), which is unique, is called the derivative of \( f \) at \( a \); it is denoted \( Df(a) \).
1.2 DIRECTIONAL DERIVATIVE AND DERIVATIVE

**Figure 1.4:** $f'(a)t$ is the linear approximation to $f(a + t) - f(a)$ at $a$.

**Figure 1.5:** $Df(a)$ is the linear part of $f$ at $a$.

**Remark 1.28** [1] Notice that $h$ is a point of $\mathbb{R}^m$ and $f(a + h) - f(a) - B_a \cdot h$ is a point of $\mathbb{R}^n$, so the norm signs are essential.

[2] The derivative $Df(a)$ depends on the point $a$ as well as the function $f$. We are not saying that there exists a $B$ which works for all $a$, but that for a fixed $a$ such a $B$ exists.

[3] Here is how to visualize $Df$. Take $m = n = 2$. The function $f : A \to \mathbb{R}^2$ distorts shapes nonlinearly; its derivative describes the linear part of the distortion. Circles are sent by $f$ to wobbly ovals, but they become ellipses under $Df(a)$ (here we treat $Df(a)$ as the matrix that represents a linear operator; see, e.g., Axler 1997.) Lines are sent by $f$ to curves, but they become straight lines under $Df(a)$. See Figure 1.5.

**Example 1.29** Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be defined by the equation

$$f(x) = A \cdot x + b.$$
where \( A \) is an \( n \times m \) matrix, and \( a \in \mathbb{R}^n \). Then
\[
f(a + h) = A \cdot (a + h) + b = A \cdot a + b + A \cdot h
\]
\[= f(a) + A \cdot h.\]

Hence, \( R_a(h) = [f(a + h) - f(a) - A \cdot h]/\|h\| = 0 \); that is, \( Df(a) = A \).

We now show that the definition of derivative is stronger than directional derivative. In particular, we have:

**Theorem 1.30** Let \( A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R}^n \). If \( f \) is differentiable at \( a \), then \( f \) is continuous at \( a \).

**Proof.** Differentiability at \( a \) implies that
\[
\|f(a + h) - f(a)\| = \|Df(a) \cdot h + \|h\| R_a(h)\|
\]
\[\leq \|Df(a)\| \cdot \|h\| + \|R_a(h)\| \cdot \|h\|
\]
\[\to 0,
\]
as \( a + h \to a \), where the inequality follows from the Triangle Inequality and Cauchy-Schwartz Inequality (Theorem 1.7).

However, there is a nice connection between directional derivative and derivative.

**Theorem 1.31** Let \( A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R}^n \). If \( f \) is differentiable at \( a \), then all the directional derivatives of \( f \) at \( a \) exist, and
\[
f'(a; u) = Df(a) \cdot u.
\]

**Proof.** Fix any \( u \in \mathbb{R}^m \) and take \( h = tu \). Then
\[
\frac{f(a + tu) - f(a)}{t} = \frac{Df(a) \cdot (tu) + \|tu\| R_a(tu)}{t}
\]
\[= \frac{Df(a) \cdot u + \|u\| R_a(tu)}{t}.
\]
The last term converges to zero as \( t \to 0 \), which proves that \( f'(a; u) = Df(a) \cdot u. \)

**Exercise 1.32** Define \( f : \mathbb{R}^2 \to \mathbb{R} \) by setting
\[
f(x, y) = \begin{cases} 
0 & \text{if } (x, y) = (0, 0) \\
x^2y & \text{if } (x, y) \neq (0, 0).
\end{cases}
\]
Show that \( f \) is not differentiable at \( (0, 0) \).

**Exercise 1.33** Let \( f : \mathbb{R}^2 \to \mathbb{R} \) be defined by \( f(x, y) = \sqrt{|xy|} \). Show that \( f \) is not differentiable at \( (0, 0) \).
1.3 Partial Derivatives and the Jacobian

We now introduce the notion of the “partial derivatives” of a real-valued function. Let \( e_1, \ldots, e_m \) be the standard basis of \( \mathbb{R}^m \), i.e.,
\[
\begin{align*}
  e_1 &= (1, 0, 0, \ldots, 0), \\
  e_2 &= (0, 1, 0, \ldots, 0), \\
  \vdots \\
  e_m &= (0, 0, \ldots, 0, 1).
\end{align*}
\]

**Definition 1.34** (Partial Derivatives) Let \( A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R} \). We define the \( j \)th partial derivative of \( f \) at \( a \) to be the directional derivative of \( f \) at \( a \) with respect to the vector \( e_j \), provided this derivative exists; and we denote it by \( D_j f(a) \). That is,
\[
D_j f(a) = \lim_{t \to 0} \frac{f(a + t e_j) - f(a)}{t}.
\]

**Remark 1.35** It is important to note that \( D_j f(a) \) is the ordinary derivative of a certain function; in fact, if \( g(x) = f(a_1, \ldots, a_{j-1}, x, a_{j+1}, \ldots, a_m) \), then \( D_j f(a) = g'(a_j) \). This means that \( D_j f(a) \) is the slope of the tangent line at \((a, f(a))\) to the curve obtained by intersecting the graph of \( f \) with the plane \( x_i = a_i \) with \( i \neq j \). See Figure 1.6.

The following theorem relates partial derivatives to the derivative in the case where \( f \) is a real-valued function.

**Theorem 1.36** Let \( A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R} \). If \( f \) is differentiable at \( a \), then
\[
D f(a) = \begin{bmatrix} D_1 f(a) & D_2 f(a) & \cdots & D_m f(a) \end{bmatrix}.
\]

**Proof.** If \( f \) is differentiable at \( a \), then \( D f(a) \) is a \((1 \times m)\)-matrix. Let
\[
D f(a) = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_m \end{bmatrix}.
\]

It follows from Theorem 1.31 that
\[
D_j f(a) = f'(a; e_j) = D f(a) \cdot e_j = \lambda_j.
\]

1.4 The Chain Rule

We extend the familiar chain rule to the current setting.
Theorem 1.37 (Chain Rule) Let $A \subset \mathbb{R}^m$; let $B \subset \mathbb{R}^n$. Let $f: A \to \mathbb{R}^n$ and $g: B \to \mathbb{R}^p$, with $f(A) \subset B$. Suppose $f(a) = b$. If $f$ is differentiable at $a$, and if $g$ is differentiable at $b$, then the composite function $g \circ f: A \to \mathbb{R}^p$ is differentiable at $a$. Furthermore,

$$D(g \circ f)(a) = Dg(b) \cdot Df(a).$$


Here is an application of the Chain Rule.

Theorem 1.38 Let $A \subset \mathbb{R}^m$; let $f: A \to \mathbb{R}^n$. Suppose $A$ contains a neighborhood of $a$. Let $f_i: A \to \mathbb{R}$ be the $i^{\text{th}}$ component function of $f$, so that

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}.$$  

(a) The function $f$ is differentiable at $a$ if and only if each component function $f_i$ is differentiable at $a$. 

Figure 1.6: $D_1 f(a, b)$. 

$\square$
(b) If \( f \) is differentiable at \( a \), then its derivative is the \((n \times m)\)-matrix whose \( i \)th row is the derivative of the function \( f_i \). That is,

\[
\begin{bmatrix}
Df_1(a) \\
\vdots \\
Df_n(a)
\end{bmatrix} =
\begin{bmatrix}
D_1f_1(a) & \cdots & D_mf_1(a) \\
D_1f_2(a) & \cdots & D_mf_2(a) \\
\vdots & \ddots & \vdots \\
D_1f_n(a) & \cdots & D_mf_n(a)
\end{bmatrix}.
\]

**Proof.** (a) Assume that \( f \) is differentiable at \( a \) and express the \( i \)th component of \( f \) as

\[ f_i = \pi_i \circ f, \]

where \( \pi_i : \mathbb{R}^n \to \mathbb{R} \) is the projection that sends a vector \( x = (x_1, \ldots, x_n) \) to \( x_i \). Notice that we can write \( \pi_i \) as \( \pi_i(x) = A \cdot x \), where \( A \) is a \( 1 \times n \) matrix such that

\[
A = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix},
\]

where the number 1 appears at the \( i \)th place. Then \( \pi_i \) is differentiable and \( D\pi(x) = A \) for all \( x \in A \) (see Example 1.29). By the Chain Rule, \( f_i \) is differentiable at \( a \) and

\[
Df_i(a) = D(\pi_i \circ f)(a) = D\pi_i(f(a)) \cdot Df(a) = A \cdot Df(a). \tag{1.4}
\]

Now suppose that each \( f_i \) is differentiable at \( a \). Let

\[
B := \begin{bmatrix}
Df_1(a) \\
\vdots \\
Df_n(a)
\end{bmatrix}.
\]

We show that \( Df(a) = B \).

\[
\|h\| R_a(h) = f(a + h) - f(a) - B \cdot h = \begin{bmatrix}
f_1(a + h) - f_1(a) - Df_1(h) \\
\vdots \\
f_n(a + h) - f_n(a) - Df_n(h)
\end{bmatrix} = \|h\| \begin{bmatrix}
R_a(h; f_1) \\
\vdots \\
R_a(h; f_n)
\end{bmatrix},
\]

where \( R_a(h; f_i) \) is the Taylor remainder for \( f_i \). It is clear that \( \lim_{h \to 0} R_a(h) = 0 \), and which proves that \( Df(a) = B \).
(b) This claim follows from the previous part and Theorem 1.36.

Remark 1.39 Theorem 1.38 implies that there is little loss of generality assuming \( n = 1 \), i.e., that our functions are real-valued. Multidimensionality of the \textit{domain}, not the \textit{range}, is what distinguished multivariable calculus from one-variable calculus.

Definition 1.40 (Jacobian Matrix) Let \( A \subseteq \mathbb{R}^m \); let \( f : A \to \mathbb{R}^n \). If the partial derivatives of the component functions \( f_i \) of \( f \) exist at \( a \), then one can form the matrix that has \( \frac{D_j f_i(a)}{D_a} \) as its entry in row \( i \) and column \( j \). This matrix, denoted by \( J f(a) \), is called the \textit{Jacobian matrix} of \( f \). That is,

\[
J f(a) = \begin{bmatrix}
D_1 f_1(a) & \cdots & D_m f_1(a) \\
\vdots & \ddots & \vdots \\
D_1 f_n(a) & \cdots & D_m f_n(a)
\end{bmatrix}.
\]

Remark 1.41 [1] The Jacobian encapsulates all the essential information regarding the linear function that best approximates a differentiable function at a particular point. For this reason it is the Jacobian which is usually used in practical calculations with the derivative

[2] If \( f \) is differentiable at \( a \), then \( J f(a) = D f(a) \). However, it is possible for the partial derivatives, and hence the Jacobian matrix, to exist, without it following that \( f \) is differentiable at \( a \) (see Exercise 1.32).

1.5 The Implicit Function Theorem

Let \( U \subseteq \mathbb{R}^k \times \mathbb{R}^n \) be open. Let \( f : U \to \mathbb{R}^n \). Fix a point \( (a, b) \in U \) and write \( f(a, b) = 0 \). Our goal is to solve the equation

\[
f(x, y) = 0
\]

near \( (a, b) \). More precisely, we hope to show that the set of points \( (x, y) \) nearby \( (a, b) \) at which \( f(x, y) = 0 \), the level-set of \( f \) through \( 0 \), is the graph of a function \( y = g(x) \). If so, \( g \) is the \textit{implicit function} defined by (1.5). See Figure 1.7.

Under various hypotheses we will show that \( g \) exists, is unique, and is differentiable. Let us first consider an example.

Example 1.42 Consider the function \( f : \mathbb{R}^2 \to \mathbb{R} \) defined by

\[
f(x, y) = x^2 + y^2 - 1.
\]
If we choose \((a, b)\) with \(f(a, b) = 0\) and \(a \neq \pm 1\), there are (Figure 1.8) open intervals \(A\) containing \(a\) and \(B\) containing \(b\) with the following property: if \(x \in A\), there is a unique \(y \in B\) with \(f(x, y) = 0\). We can therefore define a function \(g: A \to \mathbb{R}\) by the condition \(g(x) \in B\) and \(f(x, g(x)) = 0\) (if \(b > 0\), as indicated in Figure 1.8, then \(g(x) = \sqrt{1 - x^2}\)). For the function \(f\) we are considering there is another number \(b_1\) such that \(f(a, b_1) = 0\). There will also be an interval \(B_1\) containing \(b_1\) such that, when \(x \in A\), we have \(f(x, g_1(x)) = 0\) for a unique \(g_1(x) \in B_1\) (here \(g_1(x) = -\sqrt{1 - x^2}\)). Both \(g\) and \(g_1\) are differentiable. These functions are said to be defined implicitly by the equation \(f(x, y) = 0\).

If we choose \(a = 1\) or \(-1\) it is impossible to find any such function \(g\) defined in an open interval containing \(a\).

**Theorem 1.43** (Implicit Function Theorem) Let \(U \subset \mathbb{R}^{k+n}\) be open; let \(f: U \to \mathbb{R}^n\) be of class \(C^r\). Write \(f\) in the form \(f(x, y)\) for \(x \in \mathbb{R}^k\) and \(y \in \mathbb{R}^n\). Suppose that \((a, b)\) is a point of \(U\) such that \(f(a, b) = 0\). Let \(M\) be the \(n \times n\) matrix

\[
M = \begin{bmatrix}
D_{k+1} f_1(a, b) & D_{k+2} f_1(a, b) & \cdots & D_{k+n} f_1(a, b) \\
D_{k+1} f_2(a, b) & D_{k+2} f_2(a, b) & \cdots & D_{k+n} f_2(a, b) \\
\vdots & \vdots & \ddots & \vdots \\
D_{k+1} f_n(a, b) & D_{k+2} f_n(a, b) & \cdots & D_{k+n} f_n(a, b)
\end{bmatrix}.
\]

If \(\det(M) \neq 0\), then near \((a, b)\), \(L_f(0)\) is the graph of a unique function \(y = g(x)\). Besides, \(g\) is \(C^r\).

**Proof.** The proof is too long to give here. You can find it from, e.g., Spivak (1965, Theorem 2-12), Rudin (1976, Theorem 9.28), Munkres (1991, Theorem 2.9.2), or Pugh (2002, Theorem 5.22). \(\square\)
Example 1.44  Let \( f : \mathbb{R}^2 \to \mathbb{R} \) be given by the equation
\[
  f(x, y) = x^2 - y^3.
\]
Then \((0,0)\) is a solution of the equation \( f(x, y) = 0 \). Because \( \partial f(0,0)/\partial y = 0 \), we do not expect to be able to solve this equation for \( y \) in terms of \( x \) near \((0,0)\). But in fact, we can; and furthermore, the solution is unique! However, the function we obtain is not differentiable at \( x = 0 \). See Figure 1.9.

Example 1.45  Let \( f : \mathbb{R}^2 \to \mathbb{R} \) be given by the equation
\[
  f(x, y) = -x^4 + y^2.
\]
Then \((0,0)\) is a solution of the equation \( f(x, y) = 0 \). Because \( \partial f(0,0)/\partial y = 0 \), we do not expect to be able to solve for \( y \) in terms of \( x \) near \((0,0)\). In fact, however, we can do so, and we can do so in such a way that the resulting function is differentiable. However, the solution is not unique. See Figure 1.10.

Now the point \((1,1)\) is also a solution to \( f(x, y) = 0 \). Because \( \partial f(1,1)/\partial y = 2 \), one can solve this equation for \( y \) as a continuous function of \( x \) in a neighborhood of \( x = 1 \). See Figure 1.10.
Remark 1.46 We will use the Implicit Function Theorem in Theorem 2.12. The theorem will also be used to derive comparative statics for economic models, which we perhaps do not have time to discuss.

1.6 Gradient and Its Properties

In this section we investigate the significance of the gradient vector, which is defined as follows:

Definition 1.47 (Gradient) Let \( A \subseteq \mathbb{R}^m \); let \( f : A \to \mathbb{R} \). Suppose \( A \) contains a neighborhood of \( a \). The gradient of \( f \), denoted by \( \nabla f(a) \), is defined by

\[
\nabla f(a) := \begin{bmatrix} D_1 f(a) & D_2 f(a) & \cdots & D_m f(a) \end{bmatrix}.
\]

Remark 1.48 [1] It follows from Theorem 1.36 that if \( f \) is differentiable at \( a \), then \( \nabla f(a) = D f(a) \). The inverse does not hold; see Remark 1.41[2].
[2] With the notation of gradient, we can write the Jacobian of \( f = (f_1, \ldots, f_n) \) as

\[
J_f(a) = \begin{bmatrix}
\nabla f_1(a) \\
\vdots \\
\nabla f_n(a)
\end{bmatrix}.
\]

Let \( f : \mathbb{R}^m \to \mathbb{R}^n \) with \( f = (f_1, \ldots, f_n) \). Recall that the level set of \( f \) through \( 0 \) is given by

\[
L_f(0) := \{ x \in \mathbb{R}^m : f(x) = 0 \} = \bigcap_{i=1}^n L_{f_i}(0).
\]  

Given a point \( a \in L_f(0) \), it is intuitively clear what it means for a plane to be tangent to \( L_f(0) \) at \( a \). Figure 1.12 shows some examples of tangent planes. A formal definition of tangent plane will be given later. In this section we show that the gradient \( \nabla f(a) \) is orthogonal to the tangent plane of \( L_f(0) \) at \( a \) and, under some conditions, the tangent plane of \( L_f(0) \) at \( a \) can be characterized by the vectors that are orthogonal to \( \nabla f(a) \). We begin with the simplest case that \( f : \mathbb{R}^2 \to \mathbb{R} \). Here is an example.

\[\textbf{Example 1.49}\] Let \( f(x, y) = x^2 + y^2 \). The level set \( L_f(10) \) is given by

\[
L_f(10) = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 10\}.
\]

Calculus yields

\[
\frac{dy}{dx} \bigg|_{\text{along } L_f(10) \text{ at } (1,3)} = -\frac{1}{3}.
\]
Figure 1.12: Examples of tangent planes.
Hence, the tangent plane at \((1, 3)\) is given by \(y = 3 - (x - 1)/3\). Since \(\nabla f(1, 3) = (2, 6)\), the result follows immediately; see Figure 1.11(a).

The result in Example 1.49 can be explained as follows. If we change \(x_1\) and \(x_2\), and are to remain on \(L_f(0)\), then \(dx_1\) and \(dx_2\) must be such as to leave the value of \(f\) unchanged at 0. They must therefore satisfy

\[
f'(x; (dx_1, dx_2)) = D_1 f(x) \, dx_1 + D_2 f(x) \, dx_2 = 0.
\]

By solving (1.7) for \(dx_2/dx_1\), the slope of the level set through \(x\) will be (see Figure 1.11(b))

\[
\frac{dx_2}{dx_1} = -\frac{D_1 f(x)}{D_2 f(x)}.
\]

Since the slope of the vector \(\nabla f(x) = (D_1 f(x), D_2 f(x))\) is \(D_2 f(x)/D_1 f(x)\), we obtain the desired result.

We then present the general result. For simplicity, we assume throughout this section that each \(f_i \in C^1\). For \(L_f(0)\) defined in (1.6), obtaining an explicit representation for the tangent plane is a fundamental problem that we now address. First we define curves on \(L_f(0)\) and the tangent plane at some point \(x \in \mathbb{R}^m\). You may want to refer some Differential Geometry textbooks, e.g., O’Neill (2006), Spivak (1999), or Lee (2009), for understanding some of the following concepts better.

One can picture a curve in \(\mathbb{R}^m\) as a trip taken by a moving point \(c\). At each “time” \(t\) in some interval \([a, b] \subset \mathbb{R}\), \(c\) is located at the point

\[
c(t) = (c_1(t), \ldots, c_m(t)) \in \mathbb{R}^m.
\]

In rigorous terms then, \(c\) is a function from \([a, b]\) to \(\mathbb{R}^m\), and the component functions \(c_1, \ldots, c_m\) are its Euclidean coordinate functions. We define the function \(c\) to be differentiable provided its component functions are differentiable.

\begin{itemize}
  \item \textbf{Example 1.50} A helix \(c: \mathbb{R} \to \mathbb{R}^3\) is obtained through the formula
    \[
    c(t) = (a \cos t, a \sin t, bt),
    \]
    where \(a, b > 0\). See Figure 1.13.
  \item \textbf{Definition 1.51} A curve on \(L_f(0)\) is a continuous curve \(c: [a, b] \to L_f(0)\). A curve \(c(t)\) is said to pass through the point \(a \in L_f(0)\) if \(a = c(t^*)\) for some \(t^* \in [a, b]\).
\end{itemize}
1.6 GRADIENT AND ITS PROPERTIES

Figure 1.13: The Helix.

Definition 1.52 The tangent plane at \( a \in L_f(0) \), denoted \( T_f(a) \), is defined as the collection of the derivatives at \( a \) of all differentiable curves on \( L_f(0) \) passing through \( a \).

Ideally, we would like to express the tangent plane defined in Definition 1.52 in terms of derivatives of functions \( f_i \) that defines the surface \( L_f(0) \) (see Example 1.49). We introduce the subspace

\[
M := \{ x \in \mathbb{R}^m : \nabla f_i(a) \cdot x = 0 \}
\]  

and investigate under what conditions \( M \) is equal to the tangent plane at \( a \).

The following result shows that \( \nabla f_i(a) \) is orthogonal to the tangent plane \( T_f(a) \) for all \( a \in L_f(0) \).

Theorem 1.53 For each \( a \in L_f(0) \), the gradient \( \nabla f_i(a) \) is orthogonal to the tangent plane \( T_f(a) \).

Proof. We establish this result by showing \( T_f(a) \subset M \) for each \( a \in L_f(0) \). Every curve \( c(t) \) passing through \( a \) at \( t = t^* \) satisfies \( f(x(t^*)) = 0 \), and so

\[
\nabla f_i(c(t^*)) \cdot Dc(t^*) = 0.
\]

That is, \( Dc(t^*) \in M \).

\[ \square \]

Definition 1.54 A point \( a \in L_f(0) \) is said to be a regular point if the gradient vectors \( (\nabla f_1(a), \ldots, \nabla f_n(a)) \) are linearly independent.

In general, at regular points it is possible to characterize the tangent plane in terms of \( \nabla f_i(a), \ldots, \nabla f_n(a) \).
Theorem 1.55  At a regular point \(a\) of \(L_f(0)\), the tangent plane \(T_f(a)\) is equal to \(M\)

Proof. We show that \(M \subset T_f(a)\). Combining this result with Theorem 1.53, we have \(T_f(a) = M\).

To show \(M \subset T_f(a)\), we must show that if \(x \in M\) then there exists a curve on \(L_f(a)\) passing through \(a\) with derivative \(x\). To construct such a curve we consider the equations

\[
f \left( a + tx + Df(a)^T \cdot u(t) \right) = 0,
\]

where for fixed \(t\) we consider \(u(t) \in \mathbb{R}^n\) to be the unknown. This is a nonlinear system of \(n\) equations and \(n\) unknowns, parametrized continuously by \(t\). At \(t = 0\) there is a solution \(u(0) = 0\). The Jacobian matrix of the system with respect to \(u\) at \(t = 0\) is the \(n \times n\) matrix

\[
Df(a) \cdot Df(a)^T,
\]

which is nonsingular, since \(Df(a)\) is of full rank if \(a\) is a regular point. Thus, by the Implicit Function Theorem (Theorem 1.43) there is a continuously differentiable solution \(u(t)\) in some region \(t \in [-a, a]\).

The curve

\[
c(t) := a + tx + Df(a)^T \cdot u(t)
\]

is thus a curve on \(L_f(0)\). By differentiating the system (1.9) with respect to \(t\) at \(t = 0\) we obtain

\[
Df(a) \cdot \left[ x + Df(a)^T \cdot Du(0) \right] = 0.
\]

By definition of \(x\) we have \(Df(a) \cdot x = 0\) and thus, again since \(Df(a) \cdot Df(a)^T\) is nonsingular, we conclude from (1.10) that

\[
Du(0) = \left[ Df(a) \cdot Df(a)^T \right]^{-1} \cdot 0 = 0.
\]

Therefore,

\[
Dc(0) = x + Df(a)^T \cdot Du(0) = x,
\]

and the constructed curve has derivative \(x\) at \(a\).

Example 1.56  In \(\mathbb{R}^2\) let \(f(x_1, x_2) = x_1\). Then \(L_f(0)\) is the \(x_2\)-axis, and every point on that axis is regular since \(\nabla f(0, x_2) = (1, 0)\). In this case, \(T_f((x_1, 0)) = M\), and which is the \(x_2\)-axis.
In $\mathbb{R}^2$ let $f(x_1, x_2) = x_1^2$. Again, $L_f(0)$ is the $x_2$-axis, but now no point 
on the $x_2$-axis is regular: $\nabla f(0, x_2) = (0, 0)$. Indeed in this case $M = \mathbb{R}^2$, 
while the tangent plane is the $x_2$-axis.

We close this section by providing another property of gradient vectors. 
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $a \in \mathbb{R}^n$,
where $\nabla f(a) \neq 0$. Suppose that we want to determine the direction in which $f$ 
increases most rapidly at $a$. By a "direction" here we mean a unit vector $u$. Let $\theta_u$ denote 
the angle between $u$ and $\nabla f(a)$. Then

\[ f'(a; u) = \nabla f(a) \cdot u = \|\nabla f(a)\| \cos \theta_u. \]

But $\cos \theta_u$ attains its maximum value of 1 when $\theta_u = 0$, that is, when $u$ 
and $\nabla f(a)$ are collinear and point in the same direction. We conclude that 
$\|\nabla f(a)\|$ is the maximum value of $f'(a; u)$ for $u$ a unit vector, and that this 
maximum value is attained with $u = \nabla f(a)/\|\nabla f(a)\|$.

1.7 Continuously Differentiable Functions

We know that mere existence of the partial derivatives does not imply differentiability (see Exercise 1.32). If, however, we impose the additional condition that these partial derivatives are continuous, then differentiability is assured.

**Theorem 1.57** Let $A$ be open in $\mathbb{R}^m$. Suppose that the partial derivatives 
$D_j f_i(x)$ of the component functions of $f$ exist at each point $x \in A$ and are continuous on $A$. Then $f$ is differentiable at each point of $A$.

A function satisfying the hypotheses of this theorem is often said to be 
continuously differentiable, or of class $\mathcal{C}^1$, on $A$.

**Proof of Theorem 1.57.** It suffices to show that each component function of $f$ is differentiable by Theorem 1.38. Therefore we may restrict ourselves 
to the case of a real-valued function $f : A \to \mathbb{R}$. Let $a \in A$. We claim that 
$D f(a) = \nabla f(a)$ when $f \in \mathcal{C}^1$.

Recall that $(e_1, \ldots, e_m)$ is the standard basis of $\mathbb{R}^m$. Then every $h = 
(h_1, \ldots, h_m) \in \mathbb{R}^m$ can be written as $h = \sum_{i=1}^{m} h_i e_i$. For each $i = 1, \ldots, m$, let 

\[ p_i := a + \sum_{k=1}^{i} h_k e_k = p_{i-1} + h_i e_i, \]

where $p_0 := a$. Figure 1.14 illustrates the case where $m = 3$ and all $h_i$ are positive. For each $i = 1, \ldots, m$, we also define a function $\sigma_i : [0, 1] \to \mathbb{R}^m$ by
letting
\[ \sigma_i(t) = p_{i-1} + th_i e_i. \]
So, \( \sigma_i \) is a segment from \( p_{i-1} \) to \( p_i \).

By the one-dimensional chain rule and mean value theorem applied to the differentiable real-valued function \( g : [0,1] \to \mathbb{R} \) defined by
\[ g(t) = (f \circ \sigma_i)(t), \]
there exists \( \tilde{t}_i \in (0,1) \) such that
\[
\begin{align*}
    f(p_i) - f(p_{i-1}) &= g(1) - g(0) \\
    &= g' \left( \tilde{t}_i \right) \\
    &= \frac{df \left( h_i^{l-1}, h_i^{l-1}, \ldots, h_i^{l-1} \right)}{dt} \bigg|_{t=\tilde{t}_i} \\
    &= D_i f \left( \sigma_i \left( \tilde{t}_i \right) \right) \cdot h_i.
\end{align*}
\]
Telescoping \( f(a + h) - f(a) \) along \( (\sigma_1, \ldots, \sigma_m) \) gives
\[
\begin{align*}
    f(a + h) - f(a) - \nabla f(a) \cdot h &= \sum_{i=1}^{m} \left[ f \left( p_i \right) - f \left( p_{i-1} \right) \right] - \nabla f(a) \cdot h \\
    &= \sum_{i=1}^{m} \left[ D_i f \left( \sigma_i \left( \tilde{t}_i \right) \right) - D_i f(a) \right] \cdot h_i.
\end{align*}
\]
Continuity of the partials implies that \( \mathbf{D}_i f(\sigma_i(\mathbf{h})) - \mathbf{D}_i f(a) \to 0 \) as \( \|\mathbf{h}\| \to 0 \).

**Remark 1.58** It follows from Theorem 1.57 that \( \sin(xy) \) and \( xy^2 + ze^{xy} \) are both differentiable since they are of class \( \mathcal{C}^1 \).

Let \( A \subset \mathbb{R}^m \) and \( f : A \to \mathbb{R}^n \). Suppose that the partial derivative \( \mathbf{D}_i f_i \) of the component functions of \( f \) exist on \( A \). These then are functions from \( A \) to \( \mathbb{R} \), and we may consider their partial derivatives, which have the form

\[
\mathbf{D}_k (\mathbf{D}_i f_i) =: \mathbf{D}_{jk} f_i
\]

and are called the *second-order partial derivatives of \( f \).* Similarly, one defines the third-order partial derivatives of the functions \( f_i \), or more generally the *partial derivatives of order \( r \)* for arbitrary \( r \).

**Definition 1.59** If the partial derivatives of the function \( f_i \) of order less than or equal to \( r \) are continuous on \( A \), we say \( f \) is of class \( \mathcal{C}^r \) on \( A \). We say \( f \) is of class \( \mathcal{C}^1 \) on \( A \) if the partials of the functions \( f_i \) of all orders are continuous on \( A \).

**Definition 1.60** (Hessian) Let \( a \in A \subset \mathbb{R}^m \); let \( f : A \to \mathbb{R} \) be twice-differentiable at \( a \). The \( m \times m \) matrix representing the second derivative of \( f \) is called the *Hessian* of \( f \), denoted \( \mathbf{H} f(a) \):

\[
\mathbf{H} f(a) = \begin{bmatrix}
\mathbf{D}_{11} f(a) & \mathbf{D}_{12} f(a) & \cdots & \mathbf{D}_{1m} f(a) \\
\mathbf{D}_{21} f(a) & \mathbf{D}_{22} f(a) & \cdots & \mathbf{D}_{2m} f(a) \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{D}_{m1} f(a) & \mathbf{D}_{m2} f(a) & \cdots & \mathbf{D}_{mm} f(a)
\end{bmatrix} = \mathbf{D}(\nabla f).
\]

**Remark 1.61** If \( f : A \to \mathbb{R} \) is of class \( \mathcal{C}^2 \), then the Hessian of \( f \) is a symmetric matrix, i.e., \( \mathbf{D}_{ij} f(a) = \mathbf{D}_{ji} f(a) \) for all \( i, j = 1, \ldots, m \) and for all \( a \in A \). See Rudin (1976, Corollary to Theorem 9.41, p. 236).

**Exercise 1.62** Find the Hessian of the Cobb-Douglas function

\[
f(x, y) = x^\alpha y^\beta.
\]

### 1.8 Quadratic Forms: Definite and Semidefinite Matrices

**Definition 1.63** (Quadratic Form) Let \( \mathbf{A} \) be a symmetric \( n \times n \) matrix. A *quadratic form* on \( \mathbb{R}^n \) is a function \( Q_{\mathbf{A}} : \mathbb{R}^n \to \mathbb{R} \) of the form

\[
Q_{\mathbf{A}}(\mathbf{x}) = \mathbf{x} \cdot \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.
\]
Since the quadratic form $Q_A$ is completely specified by the matrix $A$, we henceforth refer to $A$ itself as the quadratic form. Observe that if $f$ is of class $C^2$, then the Hessian $Hf$ of $f$ defines a quadratic form; see Remark 1.61.

**Definition 1.64** A quadratic form $A$ is said to be
- *positive definite* if we have $x \cdot Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$;
- *positive semidefinite* if we have $x \cdot Ax \geq 0$ for all $x \in \mathbb{R}^n$;
- *negative definite* if we have $x \cdot Ax < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$;
- *negative semidefinite* if we have $x \cdot Ax \leq 0$ for all $x \in \mathbb{R}^n$.

### 1.9 Homogeneous Functions and Euler's Formula

**Definition 1.65** (Homogeneous Function) A function $f : \mathbb{R}^n \to \mathbb{R}$ is *homogeneous of degree* $r$ (for $r = \ldots, -1, 0, 1, \ldots$) if for every $t > 0$ we have

$$f(tx_1, \ldots, tx_n) = t^r f(x_1, \ldots, x_n).$$

**Exercise 1.66** The function

$$f(x, y) = Ax^\alpha y^\beta, \quad A, \alpha, \beta > 0,$$

is known as the *Cobb-Douglas* function. Check whether this function is homogeneous.

**Theorem 1.67** (Euler's Formula) Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is homogeneous of degree $r$ (for some $r = \ldots, -1, 0, 1, \ldots$) and differentiable. Then at any $x^* \in \mathbb{R}^n$ we have

$$\nabla f(x^*) \cdot x^* = rf(x^*).$$

**Proof.** By definition we have

$$f(tx^*) - t^r f(x^*) = 0.$$

Differentiating with respect to $t$ using the chain rule, we have

$$\nabla f(tx^*) \cdot x^* = rt^{r-1} f(x^*).$$

Evaluating at $t = 1$ gives the desired result.

**Lemma 1.68** If $f$ is homogeneous of degree $r$, its partial derivatives are homogeneous of degree $r - 1$. 


Exercise 1.69  Prove Lemma 1.68.

Exercise 1.70  Let \( f(x, y) = A x^\alpha y^\beta \) with \( \alpha + \beta = 1 \) and \( A > 0 \). Show that Theorem 1.67 and Lemma 1.68 hold for this function.
2

OPTIMIZATION IN $\mathbb{R}^N$


2.1 Introduction

An optimization problem in $\mathbb{R}^n$, or simply an optimization problem, is one where the values of a given function $f : \mathbb{R}^n \to \mathbb{R}$ are to be maximized or minimized over a given set $X \subset \mathbb{R}^n$. The function $f$ is called the objective function and the set $X$ the constraint set. Notationally, we will represent these problems by

Maximize $f(x)$ subject to $x \in X$,

and

Minimize $f(x)$ subject to $x \in X$,

respectively. More compactly, we shall also write

$max \{ f(x) : x \in X \}$ and $min \{ f(x) : x \in X \}$.

\textbf{Example 2.1}  \textbf{[a]} Let $X = [0, \infty)$ and $f(x) = x$. Then the problem $max \{ f(x) : x \in X \}$ has no solution; see Figure 2.1(a).

\textbf{[b]} Let $X = [0, 1]$ and $f(x) = x(1-x)$. Then the problem $max \{ f(x) : x \in X \}$ has exactly one solution, namely $x = 1/2$; see Figure 2.1(b).
[c] Let $X = [-1, 1]$ and $f(x) = x^2$. Then the problem $\max \{ f(x) : x \in X \}$ has two solutions, namely $x = -1$ and $x = 1$; see Figure 2.1(c).

Example 2.1 suggests that we shall talk of the set of solutions of the optimization problem, which is denoted

$$\arg\max \{ f(x) : x \in X \} = \{ x \in X : f(x) \geq f(y) \text{ for all } y \in X \} ,$$

and

$$\arg\min \{ f(x) : x \in X \} = \{ x \in X : f(x) \leq f(y) \text{ for all } y \in X \} .$$

We close this section by considering an optimization problem in economics.

\textbf{Example 2.2 } There are $n$ commodities in an economy. There is a consumer whose \textit{utility} from consuming $x_i \geq 0$ units of commodity $i$ ($i = 1, \ldots, n$) is given by $u(x_1, \ldots, x_n)$, where $u: \mathbb{R}_+^n \to \mathbb{R}$ is the consumer’s \textit{utility function}. The consumer’s income is $I > 0$, and faces the price vector $p = (p_1, \ldots, p_n)$. His budget set is given by (see Figure 2.2)

$$B(p, I) := \{ x \in \mathbb{R}_+^n : p \cdot x \leq I \} .$$

The consumer’s objective is to maximize his utility over the budget set, i.e.,

$$\text{Maximize } u(x) \text{ subject to } x \in B(p, I).$$

Here we introduce an important fact we will make frequent use of.
Lemma 2.3 Let $f: \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then for all $h \in \mathbb{R}^n$ satisfying $h \cdot \nabla f(x) > 0$, we have

$$f(x + \varepsilon h) > f(x) \text{ for all } \varepsilon > 0 \text{ sufficiently small.}$$

Proof. We approximate $f$ by its Taylor series expansion (see Definition 1.27):

$$f(x + \varepsilon h) = f(x) + \varepsilon h \cdot \nabla f(x) + \|\varepsilon h\| R_x(\varepsilon h),$$

where $\lim_{\varepsilon \to 0} R_x(\varepsilon h) = 0$. Then

$$\lim_{\varepsilon \to 0} \frac{f(x + \varepsilon h) - f(x)}{\varepsilon} = \frac{h \cdot \nabla f(x)}{\|h\|} + \|h\| \cdot \lim_{\varepsilon \to 0} R_x(\varepsilon h) = \frac{h \cdot \nabla f(x)}{\|h\|} > 0;$$

therefore, there is $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$,

$$\frac{f(x + \varepsilon h) - f(x)}{\varepsilon} > 0,$$

i.e., $f(x + \varepsilon) - f(x) > 0$ for all $\varepsilon \in (0, \bar{\varepsilon})$. □

2.2 Unconstrained Optimization

Definition 2.4 (Maximum) Given $X \subset \mathbb{R}^n$, $f: X \to \mathbb{R}$ and $x \in X$, we say $x$ is a maximum of $f$ if

$$f(x) = \max \{ f(y) : y \in X \}.$$
2.2.1 First-Order Necessary Conditions

Recall that $X^*$ is the interior of $X \subset \mathbb{R}^n$ (Definition 1.13), and $f'(x; u)$ is the directional derivative of $f$ at $x$ with respect to $u$ (Definition 1.21).

**Theorem 2.5** Let $X \subset \mathbb{R}^n$ and $x \in X^*$; let $f : X \to \mathbb{R}$ be differentiable at $x$. If $x$ is a local maximum of $f$, then $\nabla f(x) = 0$.

**Proof.** Suppose that $\nabla f(x) \neq 0$. Let $h = \nabla f(x)$ (we can do this since $x \in X^*$). Then $h \cdot \nabla f(x) > 0$. Hence $f(x + \varepsilon h) > f(x)$ for all $\varepsilon > 0$ sufficiently small by Lemma 2.3. This contradicts local optimality of $x$. □

**Definition 2.6** (Critical Point) A vector $x \in \mathbb{R}^n$ such that $\nabla f(x) = 0$ is called a critical point.

**Example 2.7** Let $X = \mathbb{R}^2$ and $f(x, y) = xy - 2x^4 - y^2$. The first order condition is

$$\nabla f(x, y) = (y - 8x^3, x - 2y) = (0, 0).$$

Thus, the critical points are $(x, y) = (0, 0), (1/4, 1/8), (-1/4, -1/8)$.

2.2.2 Second-Order Sufficient Conditions

The first-order conditions for unconstrained local optima do not distinguish between maxima and minima (see the following Example 2.9). To obtain such a distinction in the behavior of $f$ at an optimum, we need to examine the behavior of the Hessian $Hf$ of $f$ (see Definition 1.60).

**Theorem 2.8** Suppose $f$ is of class $C^2$ on $X \subset \mathbb{R}^n$, and $x \in X^*$.

(a) If $f$ has a local maximum at $x$, then $Hf(x)$ is negative semidefinite.

(b) If $f$ has a local minimum at $x$, then $Hf(x)$ is positive semidefinite.

(c) If $\nabla f(x) = 0$ and $Hf(x)$ is negative definite at some $x$, then $x$ is a strict local maximum of $f$ on $X$.

(d) If $\nabla f(x) = 0$ and $Hf(x)$ is positive definite at some $x$, then $x$ is a strict local minimum of $f$ on $X$.

**Proof.** See Sundaram (1996, Section 4.6). □

**Example 2.9** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = 2x^3 - 3x^2$. It is easy to check that $f \in C^2$ on $\mathbb{R}$ and there are two critical points: $x = 0$ and $x = 1$. Invoking the second-order conditions, we get $f''(0) = -6$ and $f''(1) = 6.$
Thus, the point $x = 0$ is a strict local maximum of $f$ on $\mathbb{R}$, and the point $x = 1$ is a strict local minimizer of $f$ on $\mathbb{R}$; see Figure 2.3.

However, there is nothing in the first- or second-order conditions that will help determine whether these points are global optima. In fact, they are not: global optima do not exist in this example, since $\lim_{x \to +\infty} f(x) = +\infty$ and $\lim_{x \to -\infty} f(x) = -\infty$.

### 2.3 Equality Constrained Optimization: Lagrange’s Method

In this section we consider problems with equality constraints of the form

$$\begin{align*}
\max & \quad f(x) \\
\text{s.t.} & \quad g_i(x) = 0, \quad i = 1, \ldots, k \leq n.
\end{align*}$$

We assume that $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, k$, are continuously differentiable functions. For notational convenience, we introduce the constraint function $g : \mathbb{R}^n \to \mathbb{R}^k$, where

$$g = (g_1, \ldots, g_k).$$

We can then write the constraints in the more compact form $g(x) = 0$. Also recall that a point $x^*$ such that $g_i(x^*) = 0$ for all $i = 1, \ldots, k$ is called a regular point if the gradient vectors

$$\nabla g_1(x^*), \ldots, \nabla g_k(x^*)$$
are linearly independent (see Definition 1.54).

2.3.1 First-Order Necessary Conditions

Since the representation of the tangent plane is known from Theorem 1.55, it is fairly simple to derive the necessary and sufficient conditions for a point to be a local extremum point subject to equality constraints.

**Lemma 2.10** Let \( x^* \) be a regular point of the constrains \( g(x) = 0 \) and a local maximum of \( f \) subject to these constraints. Then all \( x \in \mathbb{R}^n \) satisfying

\[
Dg(x^*) \cdot x = 0
\]  

must also satisfy

\[
\nabla f(x^*) \cdot x = 0.
\]  

**Proof.** Take an arbitrary point \( x \in M := \{ y \in \mathbb{R}^n : Dg(x^*) \cdot y = 0 \} \). Since \( x^* \) is regular, it follows from Theorem 1.55 that the tangent plane \( T_g(x^*) \) coincides with \( M \). Therefore, there exists a differentiable curve \( c : [a, b] \to \mathbb{R}^n \) on \( L_g(x^*) \) passing through \( x^* \) with \( c(t^*) = x^* \) and \( Dc(t^*) = x \) for some \( t^* \in [a, b] \).

Since \( x^* \) is a constrained local maximum of \( f \), we have

\[
\frac{df(c(t))}{dt} \bigg|_{t=t^*} = 0,
\]

or equivalently,

\[
\nabla f(x^*) \cdot x = 0.
\]

**Lemma 2.10** says that \( \nabla f(x^*) \) is orthogonal to the tangent plane \( T_g(x^*) \). Here is an example.

**Example 2.11** Consider the following problem

\[
\begin{align*}
\max & \quad f(x_1, x_2) = -x_1/2 + x_2/2 \\
\text{s.t.} & \quad g(x_1, x_2) = x_1^2 + x_2^2 = 2.
\end{align*}
\]

At the local maximum \( x^* = (-1, 1) \), the gradient \( \nabla f(x^*) \) is orthogonal to the tangent plane of the constraint surface. See Figure 2.4.

We now conclude that **Lemma 2.10** implies that \( \nabla f(x^*) \) is a linear combination of the gradients of \( g \) at \( x^* \).
0

Figure 2.4: Example 2.11.

Theorem 2.12 (Lagrange’s Theorem) Let \( x^* \) be a local maximum of \( f \) subject to \( g(x) = 0 \), and assume that \( x^* \) is a regular point of these constraints. Then, there exists a unique vector \( \lambda^* = (\lambda_1^*, \ldots, \lambda_k^*) \) such that

\[
\nabla f(x^*) = \sum_{i=1}^{k} \lambda_i^* \cdot \nabla g_i(x^*).
\]

Proof. Since \( x^* \) is a regular point of the constraints, the vectors \( \nabla g_1(x^*), \ldots, \nabla g_k(x^*) \) consists of a basis of \( \mathbb{R}^n \) if \( k = n \), and in this case it is obvious to see that the theorem holds. So we assume that \( k < n \).

Let \( U \) be the subspace spanned by \( \nabla g_1(x^*), \ldots, \nabla g_k(x^*) \). Then, \( \mathbb{R}^n \) can be represented as the direct sum of \( U \) and \( U^\perp \), the orthogonal complement of \( U \), i.e., \( \mathbb{R}^n = U \oplus U^\perp \). It follows from Theorem 1.55 that \( U^\perp = M \).

Therefore, \( \dim(\mathbb{R}^n) = \dim(U) + \dim(M) \), or equivalently,

\[
\dim(M) = n - k. \quad (2.3)
\]

Let \( V \) be the subspace spanned by \( \nabla g_1(x^*), \ldots, \nabla g_k(x^*), \nabla f(x^*) \). Suppose there does not exist \( \lambda_1, \ldots, \lambda_k \) such that \( \nabla f(x^*) = \sum_{i=1}^{k} \lambda_i \nabla g_i(x^*) \). Then \( \nabla g_1(x^*), \ldots, \nabla g_k(x^*), \nabla f(x^*) \) are linearly independent, and so \( \dim(V) = k + 1 \). However, Lemma 2.10 says that the

\[
M \subset V^\perp,
\]

and which means that We then have

\[
\dim(M) \leq \dim(V^\perp) = n - \dim(V) = n - k + 1. \quad (2.4)
\]
which contradicts (2.3). We thus conclude that $\nabla f(x^*) \in U$, i.e., there exists a unique vector $\lambda^* = (\lambda_1^*, \ldots, \lambda_k^*)$ such that

$$\nabla f(x^*) = \sum_{i=1}^{k} \lambda_i^* \nabla g_i(x^*).$$

\[\Box\]

\[\diamond\] Example 2.13 \ Consider the problem

$$\begin{align*}
\max & \quad x_1 x_2 + x_2 x_3 + x_1 x_3 \\
\text{s.t.} & \quad x_1 + x_2 + x_3 = 3.
\end{align*}$$

The necessary conditions become

$$\begin{align*}
x_2 + x_3 &= \lambda \\
x_1 + x_3 &= \lambda \\
x_1 + x_2 &= \lambda.
\end{align*}$$

These three equations together with the one constraint equation give four equations that can be solved for the four unknowns $x_1, x_2, x_3, \lambda$. Solution yields $x_1 = x_2 = x_3 = 1, \lambda = 2$.

**The Regularity** \ If $x^*$ is not a regular point, then Theorem 2.12 can fail. Consider the following example:

\[\diamond\] Example 2.14 \ Let $f(x) = (x + 1)^2$ and $g(x) = x^2$. Consider the problem of maximizing $f$ subject to $g(x) = 0$. The maximum is clearly $x = 0$. But $Dg(0) = 0$ and $Df(0) = 2$, so there is no $\lambda$ such that $Df(0) = \lambda Dg(0)$.

**The Lagrange Multipliers** \ The vector $\lambda^* = (\lambda_1^*, \ldots, \lambda_k^*)$ in Theorem 2.12 is called the vector of *Lagrange multipliers* corresponding to the local optimum $x^*$. The $i$th multiplier $\lambda_i^*$ measures the sensitivity of the value of the objective function at $x^*$ to a small relaxation of the $i$th constraint $g_i$.

To clarify the notion of “relaxation of a constraint”, let us consider the following 1-constraint case:

$$\begin{align*}
\max & \quad f(x) \\
\text{s.t.} & \quad \tilde{g}(x) - c = 0,
\end{align*}$$

where $c \in \mathbb{R}$. Then a relaxation of the constraint may be thought of as an increase in $c$. 
Suppose that for each \( c \in \mathbb{R} \) there exists a global maximum \( x^*(c) \). Suppose further that \( x^*(c) \) is a regular point for each \( c \), so there exists \( \lambda^*_*(c) \in \mathbb{R} \) such that
\[
\nabla f(x^*(c)) = \lambda^*_*(c) \cdot \nabla g_i(x^*(c)).
\]

Suppose further that \( x^*(c) \) is differentiable with respect to \( c \). Then
\[
\frac{df(x^*(c))}{dc} = \nabla f(x^*(c)) \cdot Dx^*(c)
= \lambda^*_*(c) \cdot \left[ \nabla g(x^*(c)) \cdot Dx^*(c) \right].
\]
Since \( \tilde{g}(x^*(c)) - c = 0 \), we also have \( \nabla \tilde{g}(x^*(c)) \cdot Dx^*(c) = 1 \). Therefore,
\[
\frac{df(x^*(c))}{dc} = \lambda^*_*(c).
\]
In sum, the Lagrange multipliers \( \lambda^*_*(c) \) tells us that a small relaxation in the constraint will raise the maximized value of the objective function by \( \lambda^*_*(c) \). For this reason, \( \lambda^*_*(c) \) is also called the shadow price of the constraint.

**Lagrange's Method** We now describe a procedure for using Theorem 2.12 to solve (ECP).

**Step 1** Set up a function \( \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \), called the *Lagrangian*, defined by
\[
\mathcal{L}(x, \lambda) = f(x) - \sum_{i=1}^{k} \lambda_i g_i(x).
\]
The vector \( \lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k \) is called the vector of *Lagrange multipliers*.

**Step 2** Find all critical points of \( \mathcal{L}(x, \lambda) \):
\[
\frac{\partial \mathcal{L}}{\partial x_i}(x, \lambda) = D_i f(x) - \sum_{\ell=1}^{k} \lambda_\ell D_\ell g_\ell(x) = 0, \quad i = 1, \ldots, n \tag{2.5}
\]
\[
\frac{\partial \mathcal{L}}{\partial \lambda_j}(x, \lambda) = g_j(x) = 0, \quad j = 1, \ldots, k. \tag{2.6}
\]
Define
\[
Y := \left\{(x, \lambda) \in \mathbb{R}^{n+k} : (x, \lambda) \text{ satisfies } (2.5) \text{ and } (2.6)\right\}.
\]
Step 3 Evaluate $f$ at each point $x$ in the set

$$\{x \in \mathbb{R}^n : \exists \lambda \text{ such that } (x, \lambda) \in Y\}.$$ 

Thus, we see that Lagrange's method is a clever way of converting a maximization problem with constraints, to another maximization problem without constraint, by increasing the number of variables.

Why the Lagrange's Method typically succeeds in identifying the desired optima? This is because the set of all critical points of $\mathcal{L}$ contains the set of all local maximums and minimums of (ECP) when the regularity condition is met. That is, if $x^*$ is a local maximum or minimum of $f$ subject to $g(x^*) = 0$, and if $x^*$ is regular, then there exists $\lambda^*$ such that $(x^*, \lambda^*)$ is a critical point of $\mathcal{L}$.

We are not going to explain why the Lagrange's method could fail (but see Sundaram 1996, Section 5.4 for details).

Example 2.15 Consider the problem

$$\max \quad -x^2 - y^2$$

$$\text{s.t.} \quad x + y - 1 = 0.$$ 

First, form the Lagrangian,

$$\mathcal{L}(x, y, \lambda) = -x^2 - y^2 - \lambda(x + y - 1).$$

Then set all of its first-order partials equal to zero:

$$\frac{\partial \mathcal{L}}{\partial x} = -2x - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial y} = -2y - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = x + y - 1 = 0.$$ 

So the critical points of $\mathcal{L}$ is

$$(x^*, y^*, \lambda^*) = (1/2, 1/2, -1).$$

Hence, $f(x^*, y^*) = -1/2$; see Figure 2.5.

Exercise 2.16 A consumer purchases a bundle $(x, y)$ to maximize utility. His income is $I > 0$ and prices are $p_x > 0$ and $p_y > 0$. His utility function is

$$u(x, y) = x^a y^b,$$

where $a, b > 0$. Find his optimal choice $(x^*, y^*)$. 

2.3 EQUALITY CONSTRAINED OPTIMIZATION: LAGRANGE’S METHOD

Figure 2.5: Lagrange’s method.

Lagrange’s Theorem Is Not Sufficient Lagrange’s Theorem (Theorem 2.12) only gives us a necessary—not a sufficient—condition for a constrained local maximum. To see why the first order condition is not generally sufficient, consider the following example.

Example 2.17 Consider the problem

\[
\begin{align*}
\text{max} & \quad x + y^2 \\
\text{s.t.} & \quad x - 1 = 0.
\end{align*}
\]

Observe that \((x^*, y^*) = (1, 0)\) satisfies the constraint \(g(x^*, y^*) = 0\), and \((1, 0)\) is regular. Furthermore, the first-order condition from Lagrange’s Theorem is satisfied at \((1, 0)\). This is because \(\nabla f(1, 0) = \nabla g(1, 0) = (1, 0)\). Hence, by letting \(\lambda = 1\) we have \(\nabla f(1, 0) = \lambda \nabla g(1, 0)\).

However, \((1, 0)\) is not a constrained local maximum: for \(\varepsilon > 0\), we have \(g(1, \varepsilon) = 0\) and \(f(1, \varepsilon) = 1 + \varepsilon^2 > 1 = f(1, 0)\). See Figure 2.6.

2.3.2 Second-Order Analysis

We probably do not have time to discuss the second-order conditions. See Jehle and Reny (2011, Section A2.3.4) and Sundaram (1996, Section 5.3).
2.4 Inequality Constrained Optimization: Kuhn-Tucker Theorem

We now consider a problem of the form

\[
\begin{align*}
\text{max} & \quad f(x) \\
\text{s.t.} & \quad h_1(x) \leq 0, \ldots, h_\ell(x) \leq 0, \\
\end{align*}
\] (ICP)

where \( f \) and \( h_i \) are continuously differentiable functions from \( \mathbb{R}^n \) to \( \mathbb{R} \). More succinctly, we can write this problem as

\[
\begin{align*}
\text{max} & \quad f(x) \\
\text{s.t.} & \quad h(x) = 0,
\end{align*}
\]

where \( h : \mathbb{R}^n \to \mathbb{R}^\ell \) is the function \( h = (h_1, \ldots, h_\ell) \).

For any feasible point \( x \), the set of active inequality constraints is denoted by

\[
A(x) := \{i : h_i(x) = 0\}.
\] (2.7)

If \( i \notin A(x) \), we say that the \( i^{\text{th}} \) constraint is inactive at \( x \). We note that if \( x^* \) is a local maximum of (ICP), then \( x^* \) is also a local maximum for a problem identical to (ICP) except that the inactive constraints at \( x^* \) have been discarded. Thus, inactive constraints at \( x^* \) do not matter; they can be ignored in the statement of optimality conditions.
Definition 2.18 Let \( x^* \) be a point satisfying the constraint \( h(x^*) \leq 0 \). Then \( x^* \) is said to be \textit{regular} if \( \nabla h_i(x^*) \), \( i \in A(x^*) \), are linearly independent.

2.4.1 First-Order Analysis

Example 2.19 Figure 2.7 illustrates a problem with two inequality constraints and depicts three possibilities, depending on whether none, one, or two constraints are active.

- In the first case, we could have a constrained local maximum such as \( x \), for which both constraints are inactive. Such a vector must be a critical point of the objective function.

- In the second case, only a single constraint is active at a constrained local maximum such as \( y \), and here the gradients of the objective and constraint are collinear. As we will see, these gradients actually point in the same direction.

- Lastly, we could have a constrained local maximum such as \( z \), where both constraints are active. Here, the gradient of the objective is not collinear with the gradient of either constraint, and it may appear that
no gradient restriction is possible. But in fact, $\nabla f(z)$ can be written as a linear combination of $\nabla h_1(z)$ and $\nabla h_2(z)$ with non-negative weights.

The restrictions evident in Figure 2.7 are formalized in the next theorem.

- **Theorem 2.20** (Kuhn-Tucker Theorem) Let $x^*$ be a local maximum of (ICP), and assume that $x^*$ is regular. Then there exists a vector $\lambda^* = (\lambda_1^*, \ldots, \lambda_\ell^*)$ such that

$$\lambda_i^* \geq 0 \text{ and } \lambda_i^* h_i(x^*) = 0, \quad i = 1, \ldots, \ell.$$ (KT-1)

$$\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i \nabla h_i(x^*).$$ (KT-2)

**Proof.** See Sundaram (1996, Section 6.5).

**Remark 2.21** [1] Geometrically, the first-order condition from the Kuhn-Tucker Theorem means that the gradient of the objective function, $\nabla f(x^*)$, is contained in the “semi-positive cone” generated by the gradients of binding constraints, i.e., it is contained in the set

$$\left\{ \sum_{i=1}^{\ell} \alpha_i \nabla h_i(x^*) : \alpha_1, \ldots, \alpha_\ell \geq 0 \right\},$$

depicted in Figure 2.8.

[2] Condition (KT-1) in Theorem 2.20 is called the condition of complementary slackness: if $h_i(x^*) < 0$ then $\lambda_i^* = 0$; if $\lambda_i^* > 0$ then $h_i(x^*) = 0$.

The Kuhn-Tucker Multipliers The vector $\lambda^*$ in Theorem 2.20 is called the vector of Kuhn-Tucker multipliers corresponding to the local maximum $x^*$. The Kuhn-Tucker multipliers measure the sensitivity of the objective function at $x^*$ to relaxations of the various constraints:
If \( h_i(x^*) < 0 \), then the \( i^{th} \) constraint is already slack, so relaxing it further will not help raise the value of the objective function, and \( \lambda_i^* \) must be zero.

If \( h_i(x^*) = 0 \), then relaxing the \( i^{th} \) constraint may help increase the value of the maximization exercise, so we have \( \lambda_i^* \geq 0 \).

**Two Differences** There are two important differences from the case of equality constraints (see Theorem 2.12 and Theorem 2.20):

- The regularity condition now holds only for the gradients of binding constraints.
- The multipliers are non-negative. This difference comes from the fact that now only the inequality \( h_i(x) \leq 0 \) needs to be maintained, so relaxing the constraint never hurts.

**The Regularity Condition** As with the analogous condition in Theorem 2.12 (see Example 2.14), here we show that the regularity condition in Theorem 2.20 is essential.

**Example 2.22** Consider the following maximization problem

\[
\text{max } f(x_1, x_2) = x_1
\]

s.t. \( h_1(x_1, x_2) = -(1 - x_1)^3 + x_2 \leq 0 \)

\( h_2(x_1, x_2) = -x_1 \leq 0 \)

\( h_3(x_1, x_2) = -x_2 \leq 0 \).

See Figure 2.9. Clearly the solution is \((x_1^*, x_2^*) = (1, 0)\). At this point we have

\[
\nabla h_1(1, 0) = (0, 1), \quad \nabla h_3(1, 0) = (0, -1) \quad \text{and} \quad \nabla f(1, 0) = (1, 0).
\]

Since \( x_1^* > 0 \), it follows from the complementary slackness condition (KT-1) that \( \lambda_2^* = 0 \). But now (KT-2) fails: for any \( \lambda_1 \geq 0 \) and \( \lambda_3 \geq 0 \), we have

\[
\lambda_1 \nabla h_1(1, 0) + \lambda_3 \nabla h_3(1, 0) = (0, \lambda_1 - \lambda_3) \neq \nabla f(1, 0).
\]

This is because \((1, 0)\) is not regular: There are two binding constraints at \((1, 0)\), namely \( h_1 \) and \( h_3 \), and the gradients \( \nabla h_1(1, 0) \) and \( \nabla h_3(1, 0) \) are colinear. Certainly \( \nabla f(1, 0) \) cannot be contained in the cone generated by \( \nabla h_1(1, 0) \) and \( \nabla h_3(1, 0) \); see Remark 2.21.
Figure 2.9: The constraint qualification fails at $(1, 0)$.

**The Lagrangian**  As with equality constraints, we can define the Lagrangian $\mathcal{L}: \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R}$ by

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i=1}^{\ell} \lambda_i h_i(x),$$

and then condition (KT-2) from Theorem 2.20 is the requirement that $x$ is a critical point of the Lagrangian given multipliers $\lambda_1, \ldots, \lambda_\ell$.

Let us consider a numerical example.

**Example 2.23**  Consider the problem

$$\begin{align*}
\max & \quad x^2 - y \\
\text{s.t.} & \quad x^2 + y^2 - 1 \leq 0.
\end{align*}$$

Set the Lagrangian:

$$\mathcal{L}(x, y, \lambda) = x^2 - y - \lambda(x^2 + y^2 - 1).$$
The critical points of $\mathcal{L}$ are the solutions $(x, y, \lambda)$ to

\[
\begin{align*}
2x - 2\lambda x &= 0 \quad (2.8) \\
-1 - 2\lambda y &= 0 \quad (2.9) \\
\lambda &\geq 0 \quad (2.10) \\
x^2 + y^2 - 1 &\leq 0 \quad (2.11) \\
\lambda(x^2 + y^2 - 1) &= 0 \quad (2.12)
\end{align*}
\]

For (2.8) to hold, we must have $x = 0$ or $\lambda = 1$.

- If $\lambda = 1$, then (2.9) implies $y = -1/2$, and (2.12) implies $x = \pm \sqrt{3}/2$. That is,

$$(x, y, \lambda) = \left( \pm \frac{\sqrt{3}}{2}, -\frac{1}{2}, 1 \right).$$

We thus have $f(x, y) = 5/4$; see Figure 2.10.

- If $x = 0$, then $\lambda > 0$ by (2.9). Hence $h$ is binding: $x^2 + y^2 = 1 = 0$, and so $y = \pm 1$. Since (2.9) implies that $y = 1$ is impossible, we have

$$(x, y, \lambda) = \left( 0, -1, \frac{1}{2} \right).$$

At this critical point, we have $f(0, -1) = 1 < 5/4$, which means that $(0, -1, 1/2)$ cannot be a solution. Since there are no other critical points, it follows that there are exactly two solutions to the maximization problem, namely $(x^*, y^*) = (\pm \sqrt{3}/2, -1/2)$. 
Exercise 2.24  Let $U = \mathbb{R}^2$; let $f(x, y) = (x - 1)^2 + y^2$, and let $h(x, y) = 2kx - y^2 \leq 0$, where $k > 0$. Solve the maximization problem

$$\max \{ f(x, y) : h(x, y) \leq 0 \}.$$  

Example 2.25  Consider a consumer’s problem:

$$\max_{(x, y) \in \mathbb{R}^2} \{ u(x, y) = x + y \}$$

s.t.  

$$h_1(x, y) = -x \leq 0$$
$$h_2(x, y) = -y \leq 0$$
$$h_3(x, y) = px + py - I \leq 0,$$

where $p_x, p_y, I > 0$.

We first identify all possible combinations of constraints that can, in principle, be binding at the optimum. There are eight combinations to be check:

$\emptyset, h_1, h_2, h_3, (h_1, h_2), (h_1, h_3), (h_2, h_3)$, and $(h_1, h_2, h_3)$.

Of these, the last one can be ruled out, since $h_1 = h_2 = 0$ implies that $h_3 < 0$. Moreover, since $u$ is strictly increasing in both arguments, it is obvious that $h_3 = 0$. So we only need to check three combinations: $(h_1, h_3)$, $(h_2, h_3)$, and $h_3$.

- If the optimum occurs at a point where only $h_1$ and $h_3$ are binding, then

$$\nabla h_1(x, y), \nabla h_3(x, y) = [-1, 0], (p_x, p_y)$$

is linear independent. So the constraint qualification holds at such a point.

- Similarly, the constraint qualification holds if only $(h_2, h_3)$ bind.

- If $h_3$ is the only binding constraint, then $\nabla h_3(x, y) = (p_x, p_y) \neq 0$; that is, the constraint qualification holds.

Exercise 2.26  Solve the problem of (2.13).

2.4.2 Second-Order Analysis

We probably do not have time to discuss the second-order conditions. See Duggan (2010, Section 6.3).
2.5 Envelop Theorem

Let \( A \subseteq \mathbb{R} \). The graph of a real-valued function \( f \) on \( A \) is a curve in the \( \mathbb{R}^2 \) plane, and we shall also refer to the curve itself as \( f \). Given a one-dimensional parametrized family of curves \( f_\alpha: A \to \mathbb{R} \), where \( \alpha \) runs over some interval, the curve \( h: A \to \mathbb{R} \) is the \emph{envelope of the family} if

- each point on the curve \( h \) is tangent to the graph of one of the curves \( f_\alpha \) and
- each curve \( f_\alpha \) is tangent to \( h \).

(See, e.g., Apostol 1974, p. 342 or Zorich 2004, p. 252 for this definition.) That is, for each \( \alpha \), there is some \( q \) and also for each \( q \), there is some \( \alpha \), satisfying

\[
\begin{aligned}
f_\alpha(q) &= h(q) \\
f'_\alpha(q) &= h'(q).
\end{aligned}
\]

We may regard \( h \) as a function of \( \alpha \) if the correspondence between curves and points on the envelope is one-to-one.

2.5.1 An Envelopment Theorem for Unconstrained Maximization

Consider now an unconstrained parametrized maximization problem. Let \( x^*(q) \) be the value of the control variable \( x \) that maximizes \( f(x, q) \), where \( q \) is our parameter of interest. For some fixed \( x \), the function

\[
\phi_x(q) := f(x, q)
\]

defines a curve. We also define the \emph{value function}

\[
V(q) := f(x^*(q), q) = \max_x \phi_x(q).
\]

Under appropriate conditions, the graph of the value function \( V \) will be the envelope of the curves \( \phi_x \). “Envelope theorems” in maximization theory are concerned with the tangency conditions this entails.

\textbf{Example 2.27} Let

\[
f(x, q) = q - (x - q)^2 + 1, \quad \text{where } x, q \in [0, 2].
\]

Then given \( q \), the maximizing \( x \) is given by \( x^*(q) = q \), and \( V(q) = q + 1 \).
Figure 2.11: The graph of \( V \) is the envelope of the family of graphs of the functions \( \phi_x \).

For each \( x \), the function \( \phi_x \) is given by

\[
\phi_x(q) = q - (x - q)^2 + 1.
\]

The graphs of these functions and of \( V \) are shown for selected values of \( x \) in Figure 2.11. Observe that the graph of \( V \) is the envelope of the family of graphs of the functions \( \phi_x \). Consequently the slope of \( V \) is the slope of the \( \phi_x \) to which it is tangent, that is,

\[
V'(q) = \frac{\partial \phi_x}{\partial q} \bigg|_{x=x^*(q)=q} = \frac{\partial f}{\partial q} \bigg|_{x=x^*(q)=q} = 1 + 2(x - q)\big|_{x=x^*(q)=q} = 1.
\]

This last observation is one version of the Envelope Theorem.

2.5.2 An Envelope Theorem for Constrained Maximization

Consider the maximization problem,

\[
\max_{x \in \mathbb{R}^n} f(x, q) \quad \text{s.t.} \quad g_i(x, q) = 0, \quad i = 1, \ldots, m.
\]  (2.14)
where \( x \) is a vector of choice variables, and \( q = (q_1, \ldots, q_\ell) \in \mathbb{R}^\ell \) is a vector of parameters that may enter the objective function, the constraints, or both.

Suppose that for each \( q \) there exists a unique solution \( x(q) \). Furthermore, we assume that the objective function \( f: \mathbb{R}^n \to \mathbb{R} \), constraints \( g_i: \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R} \) \((i = 1, \ldots, m)\), and solutions \( x: \mathbb{R}^\ell \to \mathbb{R}^n \) are differentiable in the parameter \( q \).

Then, for every parameter \( q \), the maximized value of the objective function is \( f(x(q), q) \). This defines a new function, \( V: \mathbb{R}^\ell \to \mathbb{R} \), called the value function. Formally,

\[
V(q) := \max_{x \in \mathbb{R}^n} \{ f(x, q) : g_i(x, q) = 0, i = 1, \ldots, m \}.
\]

(2.15)

**Theorem 2.28** (Envelope Theorem) Consider the value function \( V(q) \) for the problem (2.14). Let \((\lambda_1, \ldots, \lambda_m)\) be values of the Lagrange multipliers associated with the maximum solution \( x(\bar{q}) \) at \( \bar{q} \). Then for each \( k = 1, \ldots, \ell \),

\[
\frac{\partial V(q)}{\partial q_k} = \frac{\partial f(x(q), \bar{q})}{\partial q_k} - \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(x(q), \bar{q})}{\partial q_k}.
\]

(2.16)

**Proof.** By definition, \( V(q) = f(x(q), q) \) for all \( q \). Using the chain rule, we have

\[
\frac{\partial V(q)}{\partial q_k} = \sum_{i=1}^{n} \left[ \frac{\partial f(x(q), \bar{q})}{\partial x_i} \frac{\partial x_i(q)}{\partial q_k} \right] + \frac{\partial f(x(q), \bar{q})}{\partial q_k}.
\]

It follows from Theorem 2.12 that

\[
\frac{\partial f(x(q), \bar{q})}{\partial x_i} = \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(x(q), \bar{q})}{\partial x_i}.
\]

Hence,

\[
\frac{\partial V(q)}{\partial q_k} = \sum_{i=1}^{n} \left[ \frac{\partial f(x(q), \bar{q})}{\partial x_i} \frac{\partial x_i(q)}{\partial q_k} \right] + \frac{\partial f(x(q), \bar{q})}{\partial q_k}
\]

\[
= \sum_{i=1}^{n} \left[ \left( \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(x(q), \bar{q})}{\partial x_i} \right) \frac{\partial x_i(q)}{\partial q_k} \right] + \frac{\partial f(x(q), \bar{q})}{\partial q_k}
\]

\[
= \sum_{j=1}^{m} \lambda_j \sum_{i=1}^{n} \left[ \frac{\partial g_j(x(q), \bar{q})}{\partial x_i} \frac{\partial x_i(q)}{\partial q_k} \right] + \frac{\partial f(x(q), \bar{q})}{\partial q_k}.
\]

Finally, since \( g_j(x(q), q) = 0 \) for all \( q \), we have

\[
\sum_{i=1}^{n} \left[ \frac{\partial g_j(x(q), \bar{q})}{\partial x_i} \frac{\partial x_i(q)}{\partial q_k} \right] + \frac{\partial g_j(x(q), \bar{q})}{\partial q_k} = 0.
\]
Combining, we get (2.16).

\[52\]

\[\square\]

\[\textbf{Example 2.29} \quad \text{We are given the problem}\]

\[
\max_{(x,y)\in\mathbb{R}^2} \{f((x,y),q) = xy\} \quad \text{s.t.} \quad g((x,y),q) = 2x + 4y - q = 0.
\]

Forming the Lagrangian, we get

\[
\mathcal{L} = xy - \lambda(2x + 4y - q),
\]

with first-order conditions:

\[
y - 2\lambda = 0 \\
x - 4\lambda = 0 \\
q - 2x - 4y = 0.
\]

These solve for \(x(q) = q/4, y(q) = q/8\) and \(\lambda(q) = q/16\). Thus,

\[
V(q) = x(q)y(q) = \frac{q^2}{32}.
\]

Differentiating \(V(q)\) with respect to \(q\) we get

\[
V'(q) = \frac{q}{16}.
\]

Now let us verify this using the Envelope Theorem. The theorem tells us that

\[
V'(q) = \frac{\partial f((x(q),y(q)),q)}{\partial q} - \lambda(q) \frac{\partial g((x(q),y(q)),q)}{\partial q} = \lambda(q) = \frac{q}{16}.
\]

\[\textbf{Example 2.30} \quad \text{Consider a consumer whose utility function} \quad u: \mathbb{R}^n_+ \rightarrow \mathbb{R} \quad \text{is strictly increasing in every commodity} \quad i = 1, \ldots, n. \ \text{Then this consumer's problem is}\]

\[
\max u(x) \quad \text{s.t.} \quad \sum_{i=1}^{n} x_i p_i = I.
\]

The Lagrangian is

\[
\mathcal{L}(x,\lambda) = u(x) - \lambda \left( I - \sum_{i=1}^{n} x_i p_i \right).
\]

It follows from \textbf{Theorem 2.28} that

\[
V'(I) = \frac{\partial \mathcal{L}(x,\lambda)}{\partial I} = \lambda.
\]

That is, \(\lambda\) measures the \textit{marginal utility of income}. 
2.5.3 Integral Form Envelope Theorem

The Envelope theorems we introduced so far rely on assumptions that are not satisfactory for applications, e.g., mechanism design. Unfortunately, it is too technique to develop the more advanced treatment of the Envelope Theorem. We refer the reader to Milgrom and Segal (2002) and Milgrom (2004, Chapter 3) for the integral form Envelope Theorem.
CONVEX ANALYSIS IN $\mathbb{R}^N$

Rockafellar (1970) is the classical reference for finite-dimensional convex analysis. As for infinite-dimensional convex analysis, Luenberger (1969) is an excellent text.

To understand the material what follows, it is necessary that the reader have a good background in Multivariable Calculus (Chapter 1) and Linear Algebra (Axler, 1997).

In this chapter, we will exclusively consider convexity in $\mathbb{R}^n$ for concreteness, but much of the discussion here generalizes to infinite dimensional vector spaces. You may want to consult Lay (1982), Hiriart-Urruty and Lemaréchal (2001), Berkovitz (2002), Bertsekas, Nedić and Ozdaglar (2003), Bertsekas (2009), Ok (2007, Chapter G) and Royden and Fitzpatrick (2010, Chapter 6).

3.1 Convex Sets

Definition 3.1 (Convex Set) A subset $C \subset \mathbb{R}^n$ is convex if for every pair of points $x_1, x_2 \in C$, the line segment

$$[x_1, x_2] := \{x : x = \lambda x_1 + (1 - \lambda)x_2, 0 \leq \lambda \leq 1\}$$

belongs to $C$.

Exercise 3.2 Sketch the following sets in $\mathbb{R}^2$ and determine from figure which sets are convex and which are not:

(a) $\{(x, y) : x^2 + y^2 \leq 1\}$,

(b) $\{(x, y) : 0 < x^2 + y^2 \leq 1\}$,

(c) $\{(x, y) : y \geq x^2\}$,
Figure 3.1: \( \Delta^2 \) in \( \mathbb{R}^3 \).

(d) \( \{(x, y) : |x| + |y| \leq 1\} \), and

(e) \( \{(x, y) : y \geq 1/1 + x^2\} \).

**Lemma 3.3** Let \( \{C_\alpha\} \) be a collection of convex sets such that \( C := \bigcap \alpha C_\alpha \neq \emptyset \). Then \( C \) is convex.

**Proof.** Let \( x_1, x_2 \in C \). Then \( x_1, x_2 \in C_\alpha \) for all \( \alpha \). Since \( C_\alpha \) is convex, we have \( [x_1, x_2] \subset C_\alpha \) for all \( \alpha \). Hence, \( [x_1, x_2] \subset C \), so \( C \) is convex. \( \square \)

For each positive integer \( n \), define

\[
\Delta^{n-1} := \left\{ (\lambda_1, \ldots, \lambda_n) \in [0, 1]^n : \sum_{i=1}^{n} \lambda_i = 1 \right\}.
\]

For \( n = 1 \), the set \( \Delta^0 \) is the singleton \{1\}. For \( n = 2 \), the set \( \Delta^1 \) is the closed line segment joining \((0, 1)\) and \((1, 0)\). For \( n = 3 \), the set \( \Delta^2 \) is the closed triangle with vertices \((1, 0, 0)\), \((0, 1, 0)\) and \((0, 0, 1)\) (see Figure 3.1).

**Definition 3.4** A point \( x \in \mathbb{R}^n \) is a **convex combination** of points \( x_1, \ldots, x_k \) if there exists \( \lambda \in \Delta^{k-1} \) such that

\[
x = \sum_{i=1}^{k} \lambda_i x_i.
\]

**Lemma 3.5** A set \( C \subset \mathbb{R}^n \) is convex iff every convex combination of points in \( C \) is also in \( C \).
3.2 SEPARATION THEOREM

Proof. The “if” part is evident. So we shall prove the “only if” statement by induction on \(k\). It holds for \(k = 2\) by definition. Suppose the statement is true for \(k = n\). Now consider \(k = n + 1\). Let \(x_1, \ldots, x_{k+1} \in C\) and \(\lambda \in \Delta^n\) with \(\lambda_{k+1} \in (0, 1)\). Then

\[
x = \sum_{i=1}^{n+1} \lambda_i x_i
\]

\[
= \left(\sum_{j=1}^{n} \lambda_j\right) \left[\sum_{i=1}^{n} \left(\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}\right) x_i\right] + \lambda_{k+1} x_{k+1}
\]

\[
\in C. \quad \square
\]

Let \(A \subset \mathbb{R}^n\), and let \(\mathcal{A}\) be the class of all convex subsets of \(\mathbb{R}^n\) that contain \(A\). We have \(\mathcal{A} \neq \emptyset\)—after all, \(\mathbb{R}^n \in \mathcal{A}\). Then, by Lemma 3.3, \(\bigcap \mathcal{A}\) is a convex set in \(\mathbb{R}^n\) that contains \(A\). Clearly, this set is the smallest (that is, \(\subset\)-minimum) convex subset of \(\mathbb{R}^n\) that contains \(A\).

Definition 3.6 The convex hull of \(A\), denoted by \(\text{cov}(A)\), is the intersection of all convex sets containing \(A\).

Exercise 3.7 For a given set \(A\), let \(K(A)\) denote the set of all convex combinations of points in \(A\). Show that \(K(A)\) is convex and \(A \subset K(A)\).

Theorem 3.8 Let \(A \subset \mathbb{R}^n\). Then \(\text{cov}(A) = K(A)\).

Proof. Let \(\mathcal{A}\) be the family of convex sets containing \(A\). Since \(\text{cov}(A) = \bigcap \mathcal{A}\) and \(K(A) \in \mathcal{A}\) (Exercise 3.7), we have \(\text{cov}(A) \subset K(A)\).

To prove the reverse inclusion relation, take an arbitrary \(C \in \mathcal{A}\). Then \(A \subset C\). It follows from Lemma 3.5 that \(K(A) \subset C\). Hence \(K(A) \subset \bigcap \mathcal{A} = \text{cov}(A)\). \(\square\)

3.2 Separation Theorem

This section is devoted to the establishment of separation theorems. In some sense, these theorems are the fundamental theorems of optimization theory. For simplicity, we restrict our analysis on \(\mathbb{R}^n\).

Definition 3.9 A hyperplane \(H_a^\beta\) in \(\mathbb{R}^n\) is defined to be the set of points that satisfy the equation \((a, x) = \alpha\). Thus,

\[
H_a^\beta := \{x \in \mathbb{R}^n : (a, x) = \beta\}
\]

The vector \(a\) is said to be a normal to the hyperplane.
Remark 3.10  Geometrically, a hyperplane $H_a^\beta$ in $\mathbb{R}^n$ is a translation of an $(n-1)$-dimensional subspace (an affine manifold). Algebraically, it is a level set of a linear functional. For an excellent explanation about hyperplanes, see Luenberger (1969, Section 5.12).

A hyperplane $H_a^\beta$ divides $\mathbb{R}^n$ into two half spaces, one on each side of $H_a^\beta$. The set $\{x \in \mathbb{R}^n : \langle a, x \rangle \geq \beta \}$ is called the half-space above the hyperplane $H_a^\beta$, and the set $\{x \in \mathbb{R}^n : \langle a, x \rangle \leq \beta \}$ is called the half-space below the hyperplane $H_a^\beta$; see Figure 3.2.

It therefore seems natural to say that two sets $X$ and $Y$ are separated by a hyperplane $H_a^\beta$ if they are contained in different half spaces determined by $H_a^\beta$. We will introduce two separation theorems.

\begin{itemize}
  \item **Theorem 3.11 (A First Separation Theorem)** Let $C$ be a closed and convex subset of $\mathbb{R}^n$; let $y \in \mathbb{R}^n \sim C$. Then there exists a vector $a \in \mathbb{R}^n$ with $a \neq 0$, and a scalar $\beta \in \mathbb{R}$ such that $\langle a, y \rangle > \beta$ and $\langle a, x \rangle < \beta$ for all $x \in C$.
\end{itemize}

To motivate the proof, we argue heuristically from Figure 3.3, where $C$ is assumed to have a tangent at each boundary point. Draw a line from $y$ to $x^*$, the point of $C$ that is closest to $y$. The vector $y - x^*$ is orthogonal to $C$ in the sense that $y - x^*$ is orthogonal to the tangent line at $x^*$. The tangent line, which is exactly the set $\{z \in \mathbb{R}^n : \langle y - x^*, z - x^* \rangle = 0 \}$, separates $y$ and $C$. The point $x^*$ is characterized by the fact that $\langle y - x^*, x - x^* \rangle \leq 0$ for all $x \in C$. If we move the tangent line parallel to itself so as to pass through a point $x_0 \in (x^*, y)$, we are done. We now justify these steps in a series of claims.

\begin{itemize}
  \item **Claim 1** Let $C$ be a convex subset of $\mathbb{R}^n$ and let $y \in \mathbb{R}^n \sim C$. If there exists a point in $C$ that is closest to $y$, then it is unique.
\end{itemize}
Figure 3.3: The separating hyperplane theorem.

**Proof.** Suppose that there were two points \( x_1 \) and \( x_2 \) of \( C \) that were closest to \( y \). Then \( (x_1 + x_2)/2 \in C \) since \( C \) is convex, and so

\[
    d(y, C) \leq \left\| \frac{x_1 + x_2}{2} - y \right\|
\]

\[
    = \left\| \frac{1}{2} [(x_1 - y) + (x_2 - y)] \right\|
\]

\[
    \leq \frac{1}{2} \|x_1 - y\| + \frac{1}{2} \|x_2 - y\|
\]

\[
    = d(y, C).
\]

Hence the triangle inequality holds with equality. It follows from Exercise 1.8 that there exists \( \kappa \geq 0 \) such that \( x_1 - y = \kappa(x_2 - y) \). Clearly, \( \kappa \neq 0 \); for otherwise \( x_1 - y = 0 \) implies that \( y = x_1 \in C \). Then \( \kappa = 1 \) since \( \|x_1 - y\| = \|x_2 - y\| = d(y, C) \). But then \( x_1 = x_2 \).

**Claim 2** Let \( C \) be a closed subset of \( \mathbb{R}^n \) and let \( y \in \mathbb{R}^n \setminus C \). Then there exists a point \( x^* \in C \) that is closest to \( y \).

**Proof.** Take an arbitrary point \( x_0 \in C \). Let \( r > \|x_0 - y\| \). Then \( C_1 := \overline{B}(y; r) \cap C \) is nonempty (at least \( x_0 \) is in the intersection), closed, and bounded and hence is compact. The function \( x \mapsto \|x - y\| \) is continuous on \( C_1 \) and so attains its minimum at some point \( x^* \in C_1 \), i.e.,

\[
    \|x^* - y\| \leq \|x - y\| \quad \text{for all} \; x \in C_1.
\]

---

For a set \( X \subset \mathbb{R}^n \), the function \( d(y, X) \) is the distance from \( y \) to \( X \) defined by

\[
    d(y, X) = \inf_{x \in X} \|y - x\|.
\]
For every \( x \in C \setminus C_1 \), we have
\[
\|x - y\| > r > \|x_0 - y\| \geq \|x^* - y\|,
\]
since \( x_0 \in C_1 \).

**Claim 3** Let \( C \) be a convex subset of \( \mathbb{R}^n \) and let \( y \in \mathbb{R}^n \setminus C \). Then \( x^* \in C \) is a closest point in \( C \) to \( y \) iff
\[
\langle y - x^*, x - x^* \rangle \leq 0 \quad \text{for all } x \in C.
\]

**Proof.** Let \( x^* \in C \) be a closest point to \( y \) and let \( x \in C \). Since \( C \) is convex, we have
\[
[x^*, x] := \{ z(t) \in \mathbb{R}^n : z(t) = x^* + t(x - x^*), t \in [0, 1] \} \subset C.
\]
Let
\[
g(t) := \|z(t) - y\|^2 = \langle x^* + t(x - x^*) - y, x^* + t(x - x^*) - y \rangle
\]
\[
= \sum_{i=1}^{n} [x_i^* - y_i + t(x_i - x_i^*)]^2.
\]
Observe that \( g(0) = x^* \). Since \( g \) is continuously differentiable on \( (0, 1] \) and \( x^* \in \text{argmin}_{x \in C} \|x - y\| \), we have \( g_+(0) \geq 0 \). Since
\[
g'(t) = 2 \sum_{i=1}^{n} [x_i^* - y_i + t(x_i - x_i^*)] (x_i - x_i^*)
\]
\[
= 2 \left[ -\sum_{i=1}^{n} (y_i - x_i^*) (x_i - x_i^*) + t \sum_{i=1}^{n} (x_i - x_i^*)^2 \right]
\]
\[
= 2 \left[ -\langle y - x^*, x - x^* \rangle + t\|x - x^*\|^2 \right].
\]
(3.3)

Letting \( t \downarrow 0 \) we get (3.2).

Conversely, suppose that (3.2) holds. Take an arbitrary \( x \in C \setminus \{x^*\} \). It follows from (3.3) that if \( t \in (0, 1] \) then
\[
g'(t) = 2 \left[ -\langle y - x^*, x - x^* \rangle + t\|x - x^*\|^2 \right] \geq 2t\|x - x^*\|^2 > 0.
\]
That is, \( g \) is strictly increasing on \([0, 1]\). Thus, \( g(1) = \|x - y\| > \|x^* - y\| = g(0) \).
3.2 SEPARATION THEOREM

Proof of Theorem 3.11. We now can complete the proof of Theorem 3.11.
Let \( x^* \in C \) be the closest point to \( y \) (by Claim 1 and Claim 2). Let \( a = y - x^* \).
Then for all \( x \in C \), we have \( \langle a, x - x^* \rangle \leq 0 \) (by Claim 3), i.e., \( \langle a, x \rangle \leq \langle a, x^* \rangle \),
with equality occurring when \( x = x^* \). Hence,
\[
\max_{x \in C} \langle a, x \rangle = \langle a, x^* \rangle.
\]
On the other hand, \( \langle a, y - x^* \rangle = \|a\|^2 > 0 \), so
\[
\langle a, y \rangle = \langle a, x^* \rangle + \|a\|^2 > \langle a, x^* \rangle.
\]
Finally, take an arbitrary \( \beta \in (\langle a, x^* \rangle, \langle a, y \rangle) \). We thus have \( \langle a, y \rangle > \beta \)
and \( \langle a, x \rangle \leq \langle a, x^* \rangle < \beta \) for all \( x \in C \).
\( \square \)

Theorem 3.12 (Separation Theorem) Let \( X \) and \( Y \) be two disjoint convex
subsets of \( \mathbb{R}^n \). Then there exists \( a \in \mathbb{R}^n \) with \( a \neq 0 \) and a scalar \( \beta \in \mathbb{R} \)
that \( \langle a, x \rangle \geq \beta \) for all \( x \in X \) and \( \langle a, y \rangle \leq \beta \) for all \( y \in Y \). That is, there is a
hyperplane \( H^\beta_a \) that separates \( X \) and \( Y \).

Proof. First note that
\[
X \cap Y = \emptyset \iff 0 \notin X - Y.
\]
Since \( X \) and \( Y \) are convex, so is \( X - Y \). Hence, there exists an \( a \neq 0 \) such that
\[
\langle a, x - y \rangle \leq 0 \quad \text{for all } x - y \in X - Y.
\]
Therefore,
\[
\langle a, x \rangle \leq \langle a, y \rangle \quad \text{for all } x \in X \text{ and } y \in Y. \tag{3.4}
\]
Let \( \alpha := \sup \{ \langle a, x \rangle : x \in X \} \) and \( \gamma := \inf \{ \langle a, y \rangle : y \in Y \} \). Then from (3.4)
we get that \( \alpha \) and \( \gamma \) are finite, and for any number \( \beta \in [\alpha, \gamma] \),
\[
\langle a, x \rangle \leq \beta \leq \langle a, y \rangle \quad \text{for all } x \in X \text{ and } y \in Y.
\]
Thus the hyperplane \( H^\beta_a \) separates \( X \) and \( Y \).
\( \square \)

Theorem 3.13 (Proper Separation Theorem) Let \( X \) and \( Y \) be two convex
sets such that \( X^\circ \neq \emptyset \) and \( X^\circ \cap Y = \emptyset \). Then there exists a hyperplane \( H^\beta_a \)
that properly separates \( \overline{X^\circ} \) and \( \overline{Y} \).

Proof. Since \( X^\circ \) is convex and disjoint from \( Y \), it follows from Theorem 3.12
that there exists a hyperplane \( H^\beta_a \) such that
\[
\langle a, x \rangle \leq \beta \leq \langle a, y \rangle \quad \text{for all } x \in X^\circ \text{ and all } y \in Y. \tag{3.5}
\]
Let $x \in \overline{X}$. Then $x \in \overline{X^*}$ since $\overline{X} = \overline{X^*}$. So there exists a sequence $(x_n)$ in $X^*$ such that $x_n \to x$. It then follows from (3.5) and the continuity of the inner product that (3.5) holds for all $x \in \overline{X}$ and all $y \in \overline{Y}$. Hence, the hyperplane $H^\beta_a$ separates $\overline{X}$ and $\overline{Y}$. Since $(\overline{X})^* = X^*$, we know that $(\overline{X})^* \neq \emptyset$. Furthermore, $\langle a, x \rangle < \beta$ for $x \in (\overline{X})^*$, so the separation is proper.

**Theorem 3.14 (Strong Separation Theorem)** Let $K$ be a compact convex set and let $C$ be a closed convex set such that $K \cap C = \emptyset$. Then $K$ and $C$ can be strongly separated.

**Proof.** For every $\varepsilon > 0$, define

$$K_\varepsilon := K + \mathbb{B}(0; \varepsilon) = \bigcup_{x \in K} \{ x + \mathbb{B}(0; \varepsilon) \}$$

$$C_\varepsilon := C + \mathbb{B}(0; \varepsilon) = \bigcup_{y \in C} \{ y + \mathbb{B}(0; \varepsilon) \}.$$ 

Clear, both $K_\varepsilon$ and $C_\varepsilon$ are open and convex.

We now show that there exists $\varepsilon > 0$ such that $K_\varepsilon \cap C_\varepsilon = \emptyset$. If the assertion were false, there would exist a sequence $(\varepsilon_n)$ with $\varepsilon_n > 0$ and $\varepsilon_n \to 0$ and a sequence $(w_n)$ such that $w_n \in K_\varepsilon \cap C_\varepsilon$ for each $n$. Since

$$w_n = \begin{cases} x_n + z_n & \text{with } x_n \in K \text{ and } \|z_n\| < \varepsilon_k \\ y_n + z'_n & \text{with } y_n \in C \text{ and } \|z'_n\| < \varepsilon_k, \end{cases}$$

we have a sequence $(x_k)$ in $K$ and a sequence $(y_n)$ in $C$ such that

$$\|w_n - x_n\| < \varepsilon_k,$$

$$\|w_n - y_n\| < \varepsilon_k.$$ 

Hence,

$$\|x_n - y_n\| = \| (x_n - w_n) + (w_n - y_n) \| \leq \|x_n - w_n\| + \|y_n - w_n\| < 2\varepsilon_k.$$ (3.6)

Since $K$ is compact, there exists a subsequence $(x_{n_k})$ that converges to $x_0 \in K$. It follows from (3.6) that

$$\|y_{n_k} - x_0\| \leq \|x_{n_k} - x_0\| + \|x_{n_k} - y_{n_k}\| \to 0,$$

that is, $y_{n_k} \to x_0$. Since $C$ is closed, $x_0 \in C$. This contradicts the assumption that $C \cap K = \emptyset$, and so the assertion is true.

We have shown that there exists $\varepsilon > 0$ such that $K_\varepsilon \cap C_\varepsilon = \emptyset$. Hence, by Theorem 3.13 there is a hyperplane that properly separates $K_\varepsilon$ and $C_\varepsilon$. Since both $K_\varepsilon$ and $C_\varepsilon$ are open, the separation is strict. According to the definition, this says that $K$ and $C$ are strongly separated. \(\square\)
3.3 Systems of Linear Inequalities: Theorems of the Alternative

Let $A$ be an $m \times n$ matrix and $b \neq 0$ be a vector in $\mathbb{R}^m$. We focus on finding a non-negative $x \in \mathbb{R}^n_+$ such that $A \cdot x = b$ or show that no such $x$ exists. The problem can be framed as follows: can $b$ be expressed as a non-negative linear combination of the columns of $A$?

**Definition 3.15** The set of all non-negative linear combinations of the columns of $A$ is called the finite cone generated by the columns of $A$. It is denoted $\text{cone}(A)$.

**Lemma 3.16** Let $A$ be an $m \times n$ matrix, then $\text{cone}(A)$ is a convex set.

**Proof.** Take any two $y, y' \in \text{cone}(A)$. Then there exist $x, x' \in \mathbb{R}^n_+$ such that

$$y = A x \quad \text{and} \quad y' = A x'.$$

For any $\lambda \in [0, 1]$ we have

$$\lambda y + (1 - \lambda) y' = \lambda A x + (1 - \lambda) A x' = A \left[ \lambda x + (1 - \lambda) x' \right].$$

Since $\lambda x + (1 - \lambda) x' \in \mathbb{R}^n_+$ it follows that $\lambda y + (1 - \lambda) y' \in \text{cone}(A)$. \hfill \Box

**Lemma 3.17** Let $A$ be an $m \times n$ matrix, then $\text{cone}(A)$ is a closed set.

**Proof.** We prove this lemma by considering two cases.

**Case 1: The columns of $A$ are linearly independent.** Let $(w_n)$ be a convergent sequence in $\text{cone}(A)$ with limit $w$. We show that $w \in \text{cone}(A)$. For each $w_n$ there exists $x_n \in \mathbb{R}^n_+$ such that $w_n = A x_n$. We use the fact that $(A x_n)$ converges to show that $(x_n)$ converges. \hfill \Box

**Theorem 3.18** (Farkas' Lemma) Let $A$ be an $m \times n$ matrix and let $b \in \mathbb{R}^m - \{0\}$. Then one and only one of the systems

\[
\begin{align*}
Ax &= b, \quad x \geq 0, \quad x \in \mathbb{R}^n, \quad (F_I) \\
yA &\geq 0, \quad y \cdot b < 0, \quad y \in \mathbb{R}^m. \quad (F_{II})
\end{align*}
\]

has a solution.
**Proof.** We first suppose that \((F_I)\) has a solution \(x^*\). We shall show that \((F_{II})\) has no solution. Suppose there exists \(y_0 \in \mathbb{R}^m\) such that \(A^T y_0 > 0\). Then
\[
y_0 \cdot b = x^* \cdot (A^T y_0) \geq 0.
\]
Hence \((F_{II})\) has no solution.

We now suppose that \((F_I)\) has no solution; that is,
\[
b \notin \text{cone}(A).
\]
Since \(\text{cone}(A)\) is convex and closed, there exists a hyperplane \(H^\beta_{y_0}\) that strongly separates \(b\) and \(\text{cone}(A)\). Thus, \(y_0 \neq 0\) and
\[
\langle y_0, b \rangle > \beta > \langle y_0, z \rangle \quad \text{for all } z \in \text{cone}(A).
\]
Since \(0 \in \text{cone}(A)\), we get that \(\beta > 0\). Thus,
\[
\langle y_0, b \rangle > 0. \tag{3.7}
\]
For all \(z \in \text{cone}(A)\) we have
\[
\langle y_0, z \rangle = \langle y_0, Ax \rangle = \langle x, A^T y_0 \rangle < \alpha.
\]
the last inequality now holding for all \(x \in \mathbb{R}^n_+\). \(\square\)

*Example 3.19* We use the Farkas' Lemma to decide if the following system has a non-negative solution:
\[
\begin{bmatrix} 4 & 1 & -5 \\ 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.
\]
The Farkas alternative is
\[
\begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} 4 & 1 & -5 \\ 1 & 0 & 2 \end{bmatrix} \geq \begin{bmatrix} 0 & 0 & 0 \end{bmatrix},
\]
\[
y_1 + y_2 < 0.
\]
Or equivalently,
\[
\begin{align*}
4y_1 + y_2 & \geq 0, \\
y_1 & \geq 0, \\
-5y_1 + 2y_2 & \geq 0, \\
y_1 + y_2 & < 0.
\end{align*} \tag{3.8}
\]
There has no solution to the system (3.8). The second inequality requires that $y_1 \geq 0$. Combining this with the last inequality we conclude that $y_2 < 0$. But $y_1 \geq 0$ and $y_2 < 0$ contradict the third inequality of (3.8). So, the original system has a non-negative solution.

**Example 3.20** We use the Farkas’ Lemma to decide the solvability of the system:

\[
\begin{bmatrix}
1 & 1 & 0 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= \begin{bmatrix}
2 \\
2 \\
2 \\
1
\end{bmatrix}.
\]

We are interested in non-negative solutions of this system. The Farkas alternative is

\[
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4
\end{bmatrix}
\begin{bmatrix}
1 & 1 & 0 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4
\end{bmatrix}
\geq
\begin{bmatrix}
0 & 0 & 0
\end{bmatrix}.
\]

\[2y_1 + 2y_2 + 2y_3 + y_4 < 0.\]

One solution is $y_1 = y_2 = y_3 = -1/2$ and $y_4 = 1$, implying that the given system has no solution.

## 3.4 Convex Functions

Throughout this section we will assume the subset $C \subset \mathbb{R}^n$ is convex and $f$ is a real-valued function defined on $C$, that is, $f : C \rightarrow \mathbb{R}$. When we take $x_1, x_2 \in C$, we will let $x_t := tx_1 + (1-t)x_2$, for $t \in [0, 1]$, denote the convex combination of $x_1$ and $x_2$.

**Definitions**

**Definition 3.21** A function $f : C \rightarrow \mathbb{R}$ is **convex** if for all $x_1, x_2 \in C$ and $t \in [0, 1]$,

\[f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).\]

The function $f$ is **strictly convex** if the above inequality holds strictly.

A function $f : C \rightarrow \mathbb{R}$ is **concave** if for all $x_1, x_2 \in C$ and $t \in [0, 1]$,

\[f(tx_1 + (1-t)x_2) \geq tf(x_1) + (1-t)f(x_2).\]

The function $f$ is **strictly concave** if the above inequality holds strictly.
Definition 3.22  A function $f : C \to \mathbb{R}$ is quasi-convex if, for all $x_1, x_2 \in C$ and $t \in [0, 1],$

$$f [tx_1 + (1 - t)x_2] \leq \max \{f(x_1), f(x_2)\}.$$ 

The function $f$ is strictly quasi-convex function if the above inequality holds strictly.

A function $f : C \to \mathbb{R}$ is quasi-concave if, for all $x_1, x_2 \in C$ and $t \in [0, 1],$

$$f [tx_1 + (1 - t)x_2] \geq \min \{f(x_1), f(x_2)\}.$$ 

The function $f$ is strictly quasi-concave if the above inequality holds strictly.

Geometric Interpretation

Given a function $f : C \to \mathbb{R}$ and $y_0 \in \mathbb{R}$, let us define
The **epigraph** of \( f \): \( \text{epi}(f) := \{ (x, y) \in C \times \mathbb{R} : f(x) \leq y \} \);

The **subgraph** of \( f \): \( \text{sub}(f) := \{ (x, y) \in C \times \mathbb{R} : f(x) \geq y \} \);

The **superior set** for level \( y_0 \): \( S(y_0) := \{ x \in C : f(x) \geq y_0 \} \);

The **inferior set** for level \( y_0 \): \( I(y_0) := \{ x \in C : f(x) \leq y_0 \} \).

We then have (see Figure 3.4)

- \( f \) is convex \( \iff \) \( \text{epi}(f) \) is convex.
- \( f \) is concave \( \iff \) \( \text{sub}(f) \) is convex.
- \( f \) is quasi-convex \( \iff \) \( I(y_0) \) is convex.
- \( f \) is quasi-concave \( \iff \) \( S(y_0) \) is convex.

**Convexity and Quasi-convexity**

It is a simple matter to show that concavity (convexity) implies quasi-concavity (quasi-convexity).

*Theorem 3.23* Let \( C \subseteq \mathbb{R}^n \) and \( f : C \to \mathbb{R} \). If \( f \) is concave on \( C \), it is also quasi-concave on \( C \). If \( f \) is convex on \( C \), it is also quasi-convex on \( C \).

**Proof.** We only prove the first claim, and leave the second as an exercise. Suppose \( f \) is concave on \( C \). Take any \( x, y \in C \) and \( t \in [0, 1] \). Without loss of generality, we let

\[
f(x) \geq f(y).
\]

By the definition of concavity, we have

\[
f \left[ tx + (1-t)y \right] \geq tf(x) + (1-t)f(y) \\
= t \left[ f(x) - f(y) \right] + f(y) \\
\geq f(y) \\
= \min \{ f(x), f(y) \}.
\]

Hence, \( f \) is quasi-concave. \qed

*Exercise 3.24* Prove the second claim in *Theorem 3.23*: if \( f \) is convex, then it is also quasi-convex.
Concavity and Hessian

We now characterize concavity of a function using the Hessian matrix.

**Theorem 3.25** Let $A \subset \mathbb{R}^n$. The (twice continuously differentiable) function $f : A \rightarrow \mathbb{R}$ is concave if and only if $H_f(x)$ is negative semidefinite for every $x \in A$.

**Proof.** See Mas-Colell et al. (1995, Theorem M.C.2).

**Example 3.26** Let $A := (0, \infty) \times (-5, \infty)$. Let $f(x, y) = \ln x + \ln(y + 5)$. For each point $(x, y) \in A$, the Hessian of $f$ is

$$H_f(x, y) = \begin{bmatrix} -1/x^2 & 0 \\ 0 & -1/(y + 5)^2 \end{bmatrix}.$$ 

Then for each $(u, v) \in \mathbb{R}^2$, we have

$$\begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} -1/x^2 & 0 \\ 0 & -1/(y + 5)^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \frac{-u^2}{x^2} - \frac{v^2}{(y + 5)^2} \leq 0.$$

That is, $H_f(x, y)$ is negative semidefinite. Hence, $f(x, y)$ is concave. See Figure 3.5.

### Jensen’s Inequality

**Theorem 3.27** (Jensen’s Inequality) Let $f : C \rightarrow \mathbb{R}$, where $C \subset \mathbb{R}^n$ is convex. Then $f$ is convex iff for every finite set of points $x_1, \ldots, x_k \in C$ and every $t = (t_1, \ldots, t_k) \in \Delta^{k-1}$ (see (3.1) for the definition of $\Delta^{k-1}$),

$$f \left( \sum_{i=1}^{k} t_i x_i \right) \leq \sum_{i=1}^{k} t_i f(x_i). \quad (3.9)$$

**Proof.** The "If" part is evident. So we only prove the "only if" part. Suppose that $f$ is convex. We shall prove (3.9) by induction on $k$. For $k = 1$ the relation is trivial (remember that $\Delta^0 = \{1\}$ when $k = 1$). For $k = 2$ the relation follows from the definition of a convex function. Suppose that $k > 2$ and that (3.9) has been established for $k - 1$. We show that (3.9) holds for $k$. 
Figure 3.5: The function \( \ln x + \ln(y + 5) \) is concave.

If \( t_k = 1 \), then there is nothing to prove. If \( t_k < 1 \), set \( T = \sum_{i=1}^{k-1} t_i \). Then \( T + t_k = 1 \), \( T = 1 - t_k > 0 \) and

\[
\sum_{i=1}^{k-1} \frac{t_i}{T} = \frac{T}{T} = 1.
\]
Hence,

\[
f \left( \sum_{i=1}^{k} t_i x_i \right) = f \left( \sum_{i=1}^{k-1} t_i x_i + t_k x_k \right)
\]

\[
= f \left( T \left[ \sum_{i=1}^{k-1} \left( \frac{t_i}{T} \right) x_i \right] + t_k x_k \right)
\]

\[
\leq T f \left( \sum_{i=1}^{k-1} \left( \frac{t_i}{T} \right) x_i \right) + t_k f(x_k)
\]

\[
\leq T \left[ \sum_{i=1}^{k-1} \left( \frac{t_i}{T} \right) f(x_i) \right] + t_k f(x_k)
\]

\[
= \sum_{i=1}^{k} t_i f(x_i).
\]

where the first inequality follows from the convexity of \( f \) and the second inequality follows from the inductive hypothesis.

\[ \square \]

### 3.5 Convexity and Optimization

**Concave Programming**

We first present two results which indicates the importance of convexity for optimization theory.

- In convex optimization problems, all *local optima* must also be *global optima*.
- If a *strictly* convex optimization problem admits a solution, the solution must be *unique*.

\[ \blacksquare \textbf{Theorem 3.28} \quad \text{Let} \ X \subset \mathbb{R}^n \text{ be convex and } f : X \to \mathbb{R} \text{ be concave. Then}
\]

(a) Any local maximizer of \( f \) is a global maximizer of \( f \).

(b) The set \( \text{argmax}\{ f(x) : x \in X \} \) of maximizers of \( f \) on \( X \) is either empty or convex.
Proof. (a) Suppose $x$ is a local maximizer but not a global maximizer of $f$. Then there exists $\varepsilon > 0$ such that

$$f(x) \geq f(y), \quad \text{for all } y \in X \cap B(x; \varepsilon),$$

and there exists $z \in X$ such that

$$f(z) > f(x). \quad (3.10)$$

Since $X$ is convex, $[tx + (1-t)z] \in X$ for all $t \in (0, 1)$. Take $\tilde{t}$ sufficiently close to 1 so that $\tilde{t}x + (1-\tilde{t})z \in B(x; \varepsilon)$. By the concavity of $f$ and (3.10), we have

$$f \left[ \tilde{t}x + (1-\tilde{t})z \right] \geq \tilde{t} f(x) + (1-\tilde{t}) f(z) > f(x).$$

A contradiction.

(b) Suppose that $x_1$ and $x_2$ are both maximizers of $f$ on $X$. Then $f(x_1) = f(x_2)$. For any $t \in (0, 1)$, we have

$$f(x_1) \geq f \left[ tx_1 + (1-t)x_2 \right] \geq tf(x_1) + (1-t)f(x_2) = f(x_1).$$

That is, $f[tx + (1-t)x_2] = f(x_1)$. Thus, $tx_1 + (1-t)x_2$ is a maximizer of $f$ on $X$. \qed

**Theorem 3.29** Let $X \subset \mathbb{R}^n$ be convex and $f : X \to \mathbb{R}$ is strictly concave. Then $\operatorname{argmax}\{f(x) : x \in X\}$ either is empty or contains a single point.

**Exercise 3.30** Prove Theorem 3.29.

We now present a extremely important theorem, which says that the first-order conditions of the Kuhn-Tucker Theorem (Theorem 2.20) are both necessary and sufficient to identify optima of convex inequality-constrained optimization problem, provided a mild regularity condition is met.
Theorem 3.31 (Kuhn-Tucker Theorem under Convexity) Let $U \subset \mathbb{R}^n$ be open and convex. Let $f : U \to \mathbb{R}$ be a concave $C^1$ function. For $i = 1, \ldots, \ell$, let $h_i : U \to \mathbb{R}$ be convex $C^1$ functions. Suppose there is some $\bar{x} \in U$ such that

$$h_i(\bar{x}) < 0, \quad i = 1, \ldots, \ell.$$  
(This is called the Slater’s condition.) Then $x^*$ maximizes $f$ over $X := \{x \in U : h_i(x) \leq 0, i = 1, \ldots, \ell\}$ if and only if there is $\lambda^* \in \mathbb{R}^\ell$ such that the Kuhn-Tucker first-order conditions hold:

$$\lambda^* \geq 0, \quad \sum_{i=1}^{\ell} \lambda_i^* h_i(x^*) = 0. \quad \text{(KTC-1)}$$
$$\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i^* \nabla h_i(x^*). \quad \text{(KTC-2)}$$

Proof. See Sundaram (1996, Section 7.7). □

Example 3.32 Let $U = (0, \infty) \times (-5, \infty)$. Let $f(x, y) = \ln x + \ln(y + 5)$, $h_1(x, y) = x + y - 4$ and $h_2(x, y) = -y$. Consider the problem

$$\max_{(x, y) \in U} f(x, y)$$
$$\text{s.t.} \quad h_1(x, y) \leq 0, \quad h_2(x, y) \leq 0. \quad (3.11)$$

Exercise 3.33 Show that $(x^*, y^*) = (4, 0)$ is the unique point satisfying the first-order condition for a local maximizer of the problem (3.11).

Clearly, the Slater’s condition holds. Then, combining Theorem 3.31, Exercise 3.33 and Example 3.26, we conclude that $(4, 0)$ is a global maximizer. See Figure 3.6.

Slater’s Condition

For a formal demonstration of the need for Slater’s condition, let us consider the following example.
Example 3.34  Let $U = \mathbb{R}$; let $f(x) = x$ and $g(x) = x^2$. The only point in $\mathbb{R}$ satisfying $g(x) \leq 0$ is $x = 0$, so this is trivially the constrained maximizer of $f$. But

$$f'(0) = 1 \quad \text{and} \quad g'(0) = 0,$$

so there is no $\lambda \geq 0$ such that $f'(0) = \lambda g'(0)$.

Quasi-concavity and Optimization

Quasi-concave and quasi-convex functions fail to exhibit many of the sharp properties that distinguish concave and convex functions. As an example, we show that Theorem 3.31 fails for quasi-concave objective functions.

Example 3.35  Let $f : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be quasi-concave continuously differentiable functions, where

$$f(x) = \begin{cases} 
  x^3 & \text{if } x \in (-\infty, 0) \\
  0 & \text{if } x \in [0, 1] \\
  (x - 1)^2 & \text{if } x \in (1, \infty),
\end{cases}$$

and $h(x) = -x$; see Figure 3.7.

Exercise 3.36  Show that for every point $x \in [0, 1]$, there exists $\lambda \geq 0$ such that the pair $(x, \lambda)$ satisfies (KT-1) and (KT-2) (see p. 44).
Furthermore, the Slater’s condition holds. However, it is clear that no point $x \in [0, 1]$ can be a solution to the problem (Why?).

\[ f(x) \]

\[ g(x) = -x \]

**Figure 3.7:** Quasi-concavity and optimization.
REFERENCES


REFERENCES


INDEX

\( \mathcal{C}^1 \), 25
\( \mathcal{C}^\infty \), 27
\( \mathcal{C}^r \), 27
Active inequality constraint, 42
Affine manifold, 58
Chain Rule, 14
Cobb-Douglas function, 28
Component function, 5
Concave function, 65
Continuously differentiable, 25
Convex combination, 56
Convex function, 65
Convex hull, 57
Convex set, 55
Critical point, 34
Curve, 22
Directional derivative, 9
Envelope, 49
Epigraph, 67
Euclidean \( n \)-space, 2
Farkas’ Lemma, 63
First-order Taylor formula, 10
Gradient, 19
Graph, 5
Half space, 58
Helix, 22
Hessian, 27
Hyperplane, 57
Implicit function, 16
Implicit function theorem, 17
Inactive constraint, 42
Inequality constrained optimization, 42
Inferior set, 67
Inner product, 3
Interior, 4
Jacobian matrix, 16, 24
Jensen’s inequality, 68
Kuhn-Tucker Multipliers, 44
Kuhn-Tucker Theorem, 44
Lagrange multiplier, 38, 39
Lagrange multipliers, 38
Lagrange’s Method, 39
Lagrange’s Theorem, 37
Lagrangian, 39
Limit, 5
Local maximum, 33
Maximum, 33
maximum, 33
Negative definite, 28
Negative semidefinite, 28
Norm, 2
Orthogonal complement, 37
Partial derivative, 13
INDEX

Positive definite, 28
Positive semidefinite, 28
Proper Separation Theorem, 61

Quadratic form, 27
quasi-concave function, 66
quasi-convex function, 66

Regular point, 23, 43

Second-order partial derivative, 27
Separation Theorem, 61
Separation theorem, 58
Sequence, 5
Shadow price, 39
Slater’s condition, 72
Standard basis, 13
Strict local maximum, 33
Strictly concave function, 65
Strictly convex function, 65
Strictly quasi-concave function, 66
Strictly quasi-convex function, 66
Strong Separation Theorem, 62
Subgraph, 67
Superior set, 67

Tangent plane, 23

Unconstrained optimization, 33

Value function, 49, 51