Mathematics for Economists

Spring 2019

Kang-Oh Yi

Department of Economics

Sogang University
# Contents

1 Euclidean Space ................................................................. 1

1.1 Sets .................................................................................. 1

1.2 Number Systems ............................................................... 2

1.2.1 Real Numbers ................................................................... 2

1.2.2 Complex Numbers .......................................................... 5

1.3 Normed Space ................................................................. 8

1.4 Convex Set ......................................................................... 15

2 Sequences and Limits ............................................................ 18

2.1 Sequences ........................................................................ 18

2.2 Subsequences .................................................................... 23

2.3 Compact Set ...................................................................... 24

2.4 Convergence ..................................................................... 26

2.5 Series ............................................................................... 27

3 Univariate Function .............................................................. 29

3.1 Function ........................................................................... 29

3.2 Basic Functions ............................................................... 33

3.2.1 Polynomial Functions ..................................................... 33

3.2.2 Power Function ............................................................. 35

3.2.3 Exponential Function ..................................................... 35

3.2.4 Trigonometric Function ................................................ 36

3.3 Composite Functions ....................................................... 38

3.3.1 Composite Functions ..................................................... 38

3.3.2 Local Inverse ............................................................... 39
## 10.3 Diagonalization

- 10.3.1 Eigendecomposition................................................................. 129
- 10.3.2 Sufficient Conditions ............................................................ 130
- 10.3.3 Matrix Norm ........................................................................... 134
- 10.3.4 Matrix Power........................................................................... 136
- 10.4 Matrix Decomposition ............................................................... 139

## 10.5 Orthogonal Projection

- 10.5.1 Orthogonal Projection .............................................................. 144
- 10.5.2 Projection and Mapping ........................................................... 147
- 10.5.3 Projection Matrix ..................................................................... 148
- 10.5.4 Pseudoinverse ......................................................................... 151

## 11 Multivariable Calculus

- 11.1 Sequences, Limits, and Continuity ............................................... 153
- 11.2 Linear Functions .......................................................................... 155
- 11.3 Representing Functions ............................................................... 155
- 11.4 Partial and Directional Derivatives ............................................... 156
- 11.5 Tangent Hyperplane .................................................................... 157
- 11.6 Derivative .................................................................................... 158
- 11.7 Chain Rule .................................................................................. 161
- 11.8 Matrix Differentiation ................................................................. 163
- 11.9 Total differential .......................................................................... 164
- 11.10 Gradient .................................................................................... 164
- 11.10.1 Geometric Interpretation ......................................................... 165
- 11.10.2 Level Sets ............................................................................... 165
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.11</td>
<td>Higher order derivatives</td>
<td>167</td>
</tr>
<tr>
<td>11.12</td>
<td>Taylor’s Theorem</td>
<td>169</td>
</tr>
<tr>
<td>11.13</td>
<td>Implicit Function Theorem</td>
<td>171</td>
</tr>
<tr>
<td>11.14</td>
<td>Homogeneous functions</td>
<td>173</td>
</tr>
<tr>
<td>12</td>
<td>Quadratic Form</td>
<td>176</td>
</tr>
<tr>
<td>12.1</td>
<td>Quadratic Forms</td>
<td>176</td>
</tr>
<tr>
<td>12.2</td>
<td>Convex Sets</td>
<td>179</td>
</tr>
<tr>
<td>12.3</td>
<td>Shape of Functions</td>
<td>181</td>
</tr>
<tr>
<td>12.4</td>
<td>Transformation</td>
<td>184</td>
</tr>
<tr>
<td>13</td>
<td>Unconstrained Optimization</td>
<td>186</td>
</tr>
<tr>
<td>13.1</td>
<td>Conditions for Optimality</td>
<td>186</td>
</tr>
<tr>
<td>13.2</td>
<td>Comparative Statics</td>
<td>188</td>
</tr>
<tr>
<td>13.2.1</td>
<td>Envelope Theorem</td>
<td>188</td>
</tr>
<tr>
<td>13.2.2</td>
<td>Implicit Function Theorem</td>
<td>190</td>
</tr>
<tr>
<td>14</td>
<td>Optimization with Equality Constraints</td>
<td>191</td>
</tr>
<tr>
<td>14.1</td>
<td>Necessary Conditions</td>
<td>191</td>
</tr>
<tr>
<td>14.2</td>
<td>Lagrangian Method</td>
<td>195</td>
</tr>
<tr>
<td>14.3</td>
<td>Sufficient Conditions</td>
<td>198</td>
</tr>
<tr>
<td>14.4</td>
<td>Comparative Statics</td>
<td>201</td>
</tr>
<tr>
<td>15</td>
<td>Optimization with Inequality Constraints</td>
<td>203</td>
</tr>
<tr>
<td>15.1</td>
<td>Necessary Conditions</td>
<td>208</td>
</tr>
<tr>
<td>15.2</td>
<td>Linear Programming</td>
<td>214</td>
</tr>
<tr>
<td>15.3</td>
<td>Karush-Kuhn-Tucker Theorem</td>
<td>219</td>
</tr>
</tbody>
</table>
15.4 Sufficient Conditions ................................................................. 225
15.5 Non-Negativity Condition .......................................................... 226
15.6 Comparative Statics ................................................................. 227
16 Difference Equation Systems ....................................................... 228
  16.1 Solution and Stability ............................................................. 229
  16.2 Trajectories ........................................................................... 231
  16.3 Nonlinear System .................................................................. 232
  16.4 Stochastic System .................................................................. 232
17 Integration .................................................................................... 234
  17.1 Indefinite integral ..................................................................... 234
  17.2 Definite Integral ...................................................................... 236
A1 Logical Implications ..................................................................... 240
  A1.1 Logical Implications ............................................................... 240
  A1.2 Conditional Relationship ....................................................... 240
A2 Method of Proof .......................................................................... 242
  A2.1 Direct Proof (Proof by Deduction) ........................................... 242
  A2.2 Proof by Contradiction ............................................................ 242
  A2.3 Proof by (Mathematical) Induction .......................................... 243
  A2.4 Proof by Contraposition .......................................................... 243
1 Euclidean Space

1.1 Sets

**Definition 1-1.** A set is a collection of distinct objects. The objects in a set are called the members or elements of the set.

If $X$ is a set and $x$ is the member of a set $X$, we write $x \in X$. Otherwise, $x \notin X$. If a set $X$ has no members, we call it an empty set and denote $X = \emptyset$ or $X = \{\}$. If a set has at least one element, it is called non-empty.

**Definition 1-2.** A set is called a finite set if it has a finite number of elements. Otherwise, it is an infinite set.

**Example 1-1.** Hilbert’s paradox of the Grand Hotel.

**Definition 1-3.** If every member of $X$ is a member of $Y$ (if $x \in X$ then $x \in Y$), we say that $X$ is a subset of $Y$, and we write $X \subseteq Y$ or $Y \supseteq X$. If, in addition, there is a member of $Y$ which is not in $X$, then $X$ is said to be a proper subset of $Y$. This is written as $X \subset Y$ or $X \supset Y$. Note that $X \subseteq X$ for every set $X$.

If $X \subseteq Y$ and $Y \subseteq X$, we write $X = Y$, otherwise $X \neq Y$. Many people do not distinguish $\subseteq$ and $\subset$, and use $X \subset Y$ to indicate subset as well as a proper subset. In this case, $X \subset X$ is true.

**Definition 1-4.** Algebra of Sets

(i) Union: $X \cup Y = \{x|x \in X \text{ or } x \in Y\}$

(ii) Intersection: $X \cap Y = \{x|x \in X \text{ and } x \in Y\}$

(iii) Relative complement, set difference: $X - Y \equiv X \setminus Y = \{x|x \in A \text{ and } x \notin Y\}$

$X - Y$ is the relative complements of $Y$ in $X$ or the set difference of $X$ and $Y$.

(iv) Complement: $X^c = U \setminus X$ (complement relative to a universe $U$)

(v) Commutative laws: $X \cup Y = Y \cup X$, $X \cap Y = Y \cap X$
(vi) Associative laws: \((X \cup Y) \cup Z = X \cup (Y \cup Z), \quad X \cap (Y \cap Z) = X \cap (Y \cap Z)\)

(vii) Distributive laws: \(X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z), \quad X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)\)

**Theorem 1.1.** Let \(\{X_n\}\) be a finite or infinite collection of sets \(X_n\). **De Morgan’s Law** is:

\[
\left(\bigcup_{n=1}^{N} X_n\right)^C = \bigcap_{n=1}^{N} (X_n)^C
\]

**Proof.** Let \(A = (\bigcup_{n=1}^{N} X_n)^C\) and \(B = \bigcap_{n=1}^{N}(X_n)^C\). If \(x \in A\), then \(x \notin \bigcup_{n=1}^{N} X_n\), and \(x \notin X_n\) for any \(n\). Hence \(x \in (X_n)^C\) for every \(n\) so that \(x \in \bigcap_{n=1}^{N}(X_n)^C = B\). \(A \subset B\). Conversely, If \(x \in B\), then \(x \in (X_n)^C\) for any \(n\). Hence \(x \notin X_n\) for every \(n\) and \(x \notin \bigcup_{n=1}^{N} X_n\). Therefore, \(x \in (\bigcup_{n=1}^{N} X_n)^C = A\) and \(B \subset A\). It follows that \(A = B\).

**Example 1.2.** De Morgan’s Law with two sets

\[(X \cup Y)^C = X^C \cap Y^C, \quad (X \cap Y)^C = X^C \cup Y^C\]

**Definition 1.5.** For sets \(X\) and \(Y\), the **Cartesian product** \(X \times Y\) is the set of all ordered pairs, \(X \times Y = \{(a, b) | a \in X \text{ and } b \in Y\}\).

**Example 1.3.** Cartesian product over \(n\) sets \(X_1, X_2, \ldots, X_n\) is the set of \(n\)-tuples, \(X_1 \times X_2 \times \ldots \times X_n = \{(x_1, \ldots, x_n) | x_i \in X_i, i = 1, \ldots, n\}\). Cartesian power of a set \(X\) is the set of all points \((x_1, \ldots, x_n)\) where \(x_i\) is a member of \(X\). The plane \(\mathbb{R}^2\) is a Cartesian square of the set of real numbers, and \(\mathbb{R}^n\) is \(n\)-ary Cartesian power.

# 1.2 Number Systems

## 1.2.1 Real Numbers

**Axiom 1-6 (Field Axiom)** A **field** is a nonempty set on which addition and multiplication are defined and is satisfy 10 associated field axiom: associativity, commutative, distributive, identity, and inverses axioms.

**Axiom 1-7 (Order Axiom)** There is a complete **order**, denoted by \(\le\), on a set \(X\) is a relation with the following properties:

(i) Reflexion: For every \(x \in X\), \(x \ge x\).
(ii) Completeness: For every \( x, y \in X \), then one and only one of the statements, either \( x \leq y \) or \( y \leq x \) is true.

(iii) Transitivity: For every \( x, y, z \in S \), and if \( x \leq y \) and \( y \leq z \), then \( x \leq z \).

(iv) Antisymmetry: If \( x \leq y \) and \( y \leq x \), then \( x = y \).

Weak ordering can define the same relation with strict inequalities, but inequalities are easy to understand.

**Axiom 1-8 (Completeness Axiom)** Suppose \( X, Y \subseteq \mathbb{R} \) and both are not empty. If \( x \leq y \) for all \( x \in X \) and \( y \in Y \), then there exists a \( z \in \mathbb{R} \) such that \( x \leq z \leq y \).

These three properties characterize the real number. That is, a field that satisfies them is what we call real numbers. Here are other number systems and the symbols.

(i) natural numbers (counting numbers): \( \mathbb{N} = \{1, 2, \ldots\} \)

(ii) whole numbers: \( \mathbb{W} = \{0, \mathbb{N}\} \)

(iii) integers: \( \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\} \)

(iv) rational numbers: \( \mathbb{Q} = \{m/n | m, n \in \mathbb{Z}, n \neq 0\} \)

(v) irrational numbers: \( \mathbb{R} - \mathbb{Q} \) or \( \mathbb{R}\setminus\mathbb{Q} \)

Subscripts of + and − are used to denote the sign of elements. For instance, \( \mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\} \), \( \mathbb{Q}_{++} = \{x \in \mathbb{Q} | x > 0\} \)

The set of rational numbers and irrational numbers do not satisfy completeness axiom, but it is dense.

**Definition 1-9.** A set \( X \) is **dense** in \( \mathbb{R} \) if, for any \( x, y \in \mathbb{R} \) with \( x < y \), there exists \( z \in X \) such that \( x < z < y \).

**Definition 1-10.** An **ordered set** is a set in which an order is defined.

In the class, we use the usual ordering of real numbers.
**Definition 1-11.** For \( X \subseteq \mathbb{R} \), if there is an \( u, l \in \mathbb{R} \) such that \( x \leq u \) (\( l \leq x \)) for every \( x \in X \), we say that \( X \) is **bounded above** (below), and call \( u \) an **upper bound** (\( l \) an **lower bound**) of \( X \). A set is **bounded** if there are upper and lower bounds.

**Definition 1-12.** For \( X \subseteq \mathbb{R} \), suppose there exists an \( u, l \in \mathbb{R} \) such that

(i) \( x \leq u \) for every \( x \in X \) and \( x < u' \) implies \( u < u' \), then \( u \) is the **least upper bound** of \( X \) or the **supremum** of \( X \), and denotes \( u = \sup X \).

(ii) \( l \leq x \) for every \( x \in X \) and \( l' < x \) implies \( l' < l \), then \( l \) is called the **greatest lower bound** of \( X \) or the **infimum** of \( X \) and denotes \( l = \inf X \).

Any real number is an upper bound and a lower bound for the empty set is and thus the supremum is positive infinity and the infimum is negative infinity.

**Definition 1-13 (The Supremum Property)** If a set is a non-empty set of real members that is bounded above, there exists a supremum of the set.

**Theorem 1-2.** Completeness axiom and the supremum property are equivalent.

**Proof.** (\( \Rightarrow \)) Let \( X \subseteq \mathbb{R} \) be a nonempty set which is bounded above. Let \( U \) be the set of upper bounds for \( X \). Since \( X \) is bounded above \( U \neq \emptyset \). If \( x \in X \) and \( u \in U \), \( x \leq u \). Then by completeness axiom, there exists a \( \bar{u} \) such that \( x \leq \bar{u} \leq u \) for all \( x \in X \) and \( u \in U \). \( \bar{u} \) is the smallest upper bound. The existence of a lower bound can be shown similarly. The supremum property holds.

(\( \Leftarrow \)) Consider non-empty and bounded sets, \( X, Y \subseteq \mathbb{R} \) with \( x \leq y \) for all \( x \in X \) and \( y \in Y \). Since \( X \) is not empty and bounded above, by the supremum property, \( \bar{x} = \sup X \) exists. For any \( y \in Y \), \( y \) is a upper bound of \( X \). By the definition of supremum, \( \bar{x} \leq y \). Therefore, we have \( x \leq \bar{x} \leq y \) for all \( x \in X \) and \( y \in Y \). This is the completeness axiom.

**Definition 1-14.** If there is an \( M \in X \) (\( m \in X \)) such that \( x \leq M \) (\( m \leq x \)) for every \( x \in X \), we say that \( M \) (\( m \)) is the **maximum** of \( X \) (the **minimum** of \( X \)) and write \( M = \max X \) (\( m = \min X \)).

**Theorem 1-3.**

(i) If supremum (infimum, maximum, minimum) exist, it is unique.

(ii) If \( \max X \) (\( \min X \)) exists, then \( \sup X = \max X \) (\( \inf X = \min X \)).
Proof. (i) Assume there are two different ones, and apply the definition.

(ii) Let $M = \max X$. $a$ is an upper bound of $X$ by the definition. Consider another upper bound $M' \neq M$ of $X$. Since $M'$ is an upper bound and $M \in X$, we must have $M' > M$. ⊓⊔

1.2.2 COMPLEX NUMBERS

**Definition 1-15.** A complex number is an ordered pair $(a, b)$ of real numbers on which the operations of addition and multiplication is defined as follows. For any complex numbers $z = (a, b)$ and $w = (c, d)$,

(i) addition: $z + w = (a + c, b + d)$

(ii) multiplication: $z \cdot w = (ac - bd, ad + bc)$

The set of complex numbers is denoted by $\mathbb{C}$ and it is a field. Note that there is no natural order on the set of complex numbers and we do not compare them.

The following definition gives a popular expression of complex numbers.

**Definition 1-16.** A complex number is the quantity of the form $a + bi$ where $a, b \in \mathbb{R}$ and $i$ is an imaginary number that is the solution to the equation $x^2 = -1$.

The imaginary number of $i$ has properties that $i^2 = -1$ and $\sqrt{i} = (1 + i)/\sqrt{2}$. The real number $a$ is called the real part and denoted by $a = \text{Re}(z)$ and its imaginary part is $b = \text{Im}(z)$.

Since the set $\mathbb{C}$ is a field, when we perform the basic operations of complex numbers considering the real part, the imaginary part and the imaginary number as numbers. Also, the commutative, associative, and distributive rules of addition and multiplication can be applied.

**Theorem 1-4.** For any complex numbers $z = a + bi$ and $w = c + di$,

(i) addition: $z + w = (a + c) + (b + d)i$

(ii) multiplication: $z \cdot w = (ac - bd) + (ad + bc)i$
1.2 Real Numbers

(iii) reciprocal and division: \( \frac{1}{z} = \frac{\bar{z}}{z} \) and \( \frac{w}{z} = w \times \frac{1}{z} \), where \( \bar{z} = a - bi \)

The Cartesian form of a complex number, \( z = a + bi \), can be directly depicted by the Cartesian coordinates in the complex plane where the horizontal axis displays the real part and the vertical axis corresponds to the imaginary part.

![Complex Numbers in the Complex Plane](image)

**Figure 1-1 Complex Numbers in the Complex Plane**

The Euclidean norm of a complex number is the Euclidean distance of the point from the origin, and is called the **modulus** or the **absolute value** of the number, \( \sqrt{a^2 + b^2} \).

A complex number is often presented in polar form using the line segment connecting the point and the origin. Let \( \theta \) be the angle between the positive part of the real axis and the line segment and \( r \) be the length of the segment in the complex plane. The polar form of complex numbers is

\[
z = r(\cos \theta + i \sin \theta)
\]
Let’s consider the quadratic formula.

\[ x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

If \( b^2 - 4ac < 0 \), the quadratic form does not have a real solution. In this case, notice that the imaginary part always come with its negative value. In fact, if the coefficients of a polynomial are all real, any non-real roots have a conjugate pair, \((z, \bar{z})\).

**DEFINITION 1-17.** The **complex conjugate** of a complex number \( z = a - bi \) is the complex number with an equal real part and the negative of the imaginary part. It is denoted by \( \bar{z} = a - bi \).

Since a complex conjugate cannot be expressed by applying the basic operations, it is usually regarded as a separate operation.

**THEOREM 1-5.** For any complex numbers \( z \) and \( w \),

(i) \( \frac{z + \bar{z}}{2} = \text{Re}(z) \) and \( \frac{z - \bar{z}}{2i} = \text{Im}(z) \)

(ii) \( z + \bar{w} = \bar{z} + \bar{w} \)

(iii) \( \bar{z} \cdot \bar{w} = \bar{z} \cdot \bar{w} \)

(iv) \( \frac{1}{\bar{z}} = 1/\bar{z} \)

(v) \( |z| = z \cdot \bar{z} \)

(vi) Euler’s formula: \( e^{i\theta} = \cos \theta + i \sin \theta \) (see EXAMPLE 5-7 and EXAMPLE 5-8)

By the last property, it is easy to see that the multiplication of a complex number to another rotates the segment counterclockwise by the angle of the complex number.
1.3 NORMED SPACE

The notions we developed are based on only the binary relation between members of a set. In this section, we extend them with a metric in \( \mathbb{R}^n \)-dimensional real coordinate space denoted by \( \mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R} \) with a typical element \( \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n \). The \( n \)-tuple, an element of \( \mathbb{R}^n \), will be called a vector as a list of numbers. The definition is given in Section 6.2.

**NOTATION.** An element in \( \mathbb{R} \) is denoted by italic lower-case letters, \( x \in \mathbb{R} \), and subsets of \( \mathbb{R} \) by italic capital letters, \( X \subset \mathbb{R} \). We denote a \( n \)-component element in bold lower-case letters, \( \mathbf{x} \in \mathbb{R}^n \) and associated sets in capital letters, \( X \subset \mathbb{R}^n \). A subscript denotes an element of a set and \( x_{ij} \) is the \( j \)th component of a point \( x_i \in \mathbb{R}^n \). Vector spaces and matrices are denoted by italic capital letters such as \( V \) and \( A \).

**DEFINITION 1-18 (Inequalities)** For \( \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \),

(i) \( \mathbf{x} \geq \mathbf{y} \) if \( x_i \geq y_i \) for all \( i = 1, 2, \ldots n \)

(ii) \( \mathbf{x} > \mathbf{y} \) if \( x_i \geq y_i \) for all \( i = 1, 2, \ldots n \) and \( x_j > y_j \) for some \( j \)

(iii) \( \mathbf{x} \gg \mathbf{y} \) if \( x_i > y_i \) for all \( i = 1, 2, \ldots n \)

Note that the inequalities are partial ordering, and thus most of the definitions based on an order cannot be extended to \( \mathbb{R}^n \) space.

**DEFINITION 1-19.** A metric or distance on a set \( X \) is a function

\[
d : X \times X \to \mathbb{R}_+
\]
such that, for any $x, y, z \in X$,

(i) $d(x, y) \geq 0$ (positive definiteness)

(ii) $d(x, y) = 0$ if and only if $x = y$

(iii) $d(x, y) = d(y, x)$ (symmetry)

(iv) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

A distance function can be thought of as a rule that associates a real number to the pair of points.

**Example 1-4.** For $x, y \in \mathbb{R}^n$

$$d(x, y) = \begin{cases} 
1 & \text{if } x \neq y \\
0 & \text{if } x = y
\end{cases}, \quad d(x, y) = |x_1 - y_1|, \quad d(x, y) = \sum_{i=1}^{n}|x_i - y_i|$$

**Definition 1-20.** The distance between a point $x_0$ and a set $X$ is $\inf_{x \in X} d(x, x_0)$.

**Example 1-5.** The distance between $(2,0)$ and $\{(x, y) \in \mathbb{R}^2 | y = x\}$ is the infimum of $\sqrt{(2-x)^2 + y^2}$. Since the origin, $(2,0)$, and the line segment connecting $(2,0)$ and the line should form a right triangle, the distance is the infimum of $\sqrt{(2^2 + 2^2)} - 2^2 = 2$. \qed

**Definition 1-21.** A **metric space** $(X, d)$ is a set for which metric between all members of the set are defined.

**Definition 1-22.** A **normed (vector) space** $V \subset \mathbb{R}^n$ is a vector space equipped with a norm $\| \cdot \|: V \to \mathbb{R}_+$ such that, for all $x, y \in V$,

(i) $\|x\| \geq 0$

(ii) $\|x\| = 0$ if and only if $x = 0$

(iii) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

(iv) $\|\lambda x\| = |\lambda|\|x\|$

Roughly, a vector space can be thought of as a set of vectors. A norm is a notion of the length of a vector, and it also can be used to define a distance.
**Theorem 1-6.** Let \((V, \|\cdot\|)\) be a normed vector space with \(V \subset \mathbb{R}^n\) and \(d: V \times V \to \mathbb{R}_+\) be \(d(v,w) = \|v - w\|\), then \((V, d)\) is a metric space.

**Proof.** We show that \(\|\cdot\|\) satisfies the four properties of the metric in **Definition 1-19**. Let \(u, v, w \in V\).

(i) \(d(v, w) = \|v - w\| \geq 0\), which is obvious from **Definition 1-22**, (i).

(ii) \(d(v, w) = 0 \Leftrightarrow \|v - w\| = 0 \Leftrightarrow v - w = 0 \Leftrightarrow v = w\)

where the second relationship holds because of **Definition 1-22**, (ii).

(iii) \(d(v, w) = \|v - w\| = |-1|\|v - w\| = \|(-1)(v - w)\| = \|-v + w\| = \|w - v\| = d(w, v)\)

(iv) \(d(u, w) = \|u - w\| = \|u + (-v + v) - w\| = \|(u - v) + (v - w)\| \leq \|(u - v)\| + \|(v - w)\| = d(u, v) + d(v, w)\)

In \(\mathbb{R}^n\), we define boundedness in a different form and there is no proper definitions of least upper bound and greatest lower bound.

**Definition 1-23.** In a metric space \((X, d)\), a subset \(S \subset X\) is **bounded** if there exists \(x_0 \in X\) and \(r > 0\) such that, for all \(s \in S\), \(d(s, x_0) < r\).

Here are the most commonly used metrics.

**Definition 1-24.** Let \(p \geq 1\) be a real number. The **\(L_p\)-norm** or **\(\ell_p\)-norm** of \(x \in \mathbb{R}^n\) is

\[
\|x\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad \|x\|_\infty = \sup_{1 \leq i \leq n} |x_i|
\]

and the **\(L_p\)-distance** of \(x \in \mathbb{R}^n\) is

\[
d(x, y) = \|x - y\|_p = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \quad \|x - y\|_\infty = \sup_{1 \leq i \leq n} |x_i - y_i|
\]
**DEFINITION 1-25.** Euclidean norm and Euclidean distance of $x \in \mathbb{R}^n$ are $L_2$-norm and $L_2$-distance, respectively. A real coordinate space $\mathbb{R}^n$ with this structure is called Euclidean space and denoted by $\mathbb{E}^n$.

In following, we consider only Euclidean space (with finite dimension) and $\mathbb{E}^n$ and $\mathbb{R}^n$ are used interchangeably.

**DEFINITION 1-26.** The open ball of radius $r > 0$ around a point $x \in \mathbb{R}^n$, $B_r(x)$, is defined to be the set of all $y \in \mathbb{R}^n$ such that $\|x - y\| < r$.

An open ball centered around $x$ is often called an open neighborhood of $x$. Here are some terms related to an open ball.

**DEFINITION 1-27.** A set $X \subset \mathbb{R}^n$ is open if, for every $x \in X$, there is an $\varepsilon > 0$ such that $B_\varepsilon(x) \subset X$.

Sometimes it is said that $X$ is open in $S$. That means $X \cap S$ is open. In general, openness depends on the underlying metric space $S$.

**EXAMPLE 1-6.** $X = [0,1]$ is open in $S = [0,1]$ because $B_\varepsilon(0) = \{x \in [0,1] | |x - 0| < \varepsilon\} = [0,\varepsilon) \subset X$ and similarly $B_\varepsilon(1) \subset X$.

**EXAMPLE 1-7.** $X = \{1\}$ is open in $S = \{0,1,2\}$ because $B_1(1) = \{1\} \subset X$.

From now on, we consider only the case, $S = \mathbb{R}^n$.

**EXAMPLE 1-8.** An open ball is an open set.
Consider $B_\varepsilon(x)$. For a $y \in B_\varepsilon(x)$, let $\delta = \varepsilon - d(y, x)$. Then $B_\delta(y) \subset B_\varepsilon(x)$ for all $y \in B_\varepsilon(x)$.

**Theorem 1-7.** A set $X$ is closed if and only if its complement $\mathbb{R}^n \setminus X$ is open.

The definition of a closed set is given in Definition 2-10 and the proof is in Example 2-9. Note that “a set $X$ is not open” does not imply “$X$ is close.” The sets $\emptyset$ and $\mathbb{R}^n$ are only two sets that are both open and closed. Since the empty set contains no element, it satisfies the definition, and $\mathbb{R}^n$ contains any open ball around every point. Since they are complements to each other, they both are closed.

**Theorem 1-8.**

(i) A finite union or intersection of open (closed) sets is open (closed)

(ii) The infinite union of open sets is open.

(iii) The infinite intersection of closed sets is closed.

(iv) The intersection of infinitely many open sets might not be open.

(v) The union of infinitely many closed sets might not be closed.

**Proof.**

(i-1) Union of open sets: Let $X = \bigcup_{n \in N} X_n$ with $N \in \mathbb{N}$. For any element $x \in X$, $x \in X_n$ for at least one set $X_n$, $n \in N$. Since the set $X_n$ is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset X_n$.

(i-2) Intersection of open sets: Let $X = \bigcap_{n=1}^{N} X_n$. For any element $x \in X$, $x \in X_n$ for all $X_n$. Since all sets $X_n$ are open, there exists $\varepsilon_n > 0$ such that $B_\varepsilon(x) \subset X_n$ for all $n = 1, \ldots, N$. Let $\varepsilon = \min_n \varepsilon_n > 0$. By construction, $B_\varepsilon(x) \subset X_n$ for all $n = 1, \ldots, N$. Note that this proof works only for finite intersections.

(i-3,4) Closed set: Apply Theorem 1-1 to (i-1,2).
(ii), (iii) The proof of (i-1) applies. Note that the proof of (i-2) is only for finite interactions.

(iv), (v) Counterexamples:

\[
X_n = \left(-\frac{1}{n}, \frac{1}{n}\right), \quad i = 1, 2, \ldots, \quad \bigcap_{n=1}^{\infty} X_n = \{0\}
\]

\[
X_n = \left[\frac{1}{n}, 1\right], \quad i = 1, 2, \ldots, \quad \bigcup_{n=1}^{N} X_n = (0, 1]
\]

**Example 1-9.** Let \( X = \bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, \frac{1}{n}\right) \) and \( Y = \bigcap_{n=1}^{\infty} \left[0, \frac{1}{n}\right] \). Since only 0 is a unique member of two sets, \( X = Y \cdot \)

**Example 1-10.** If a finite set of numbers is taken out from \( \mathbb{R} \), that is open. Otherwise, it could be open or closed.

**Example 1-11.** \( \left\{ \frac{1}{n}, \frac{1}{n+1}, \ldots \right\} \) is not closed. Since 0 is an element of the complement and every open ball around 0 contains an element of the original set, the complement is not open.

**Example 1-12.** \( \left\{ \frac{1}{n}, \frac{1}{n+1}, \ldots \right\} \cup \{0\} \) is closed. The complement of the set is

\[
\bigcup_{n=1}^{\infty} \left(\frac{1}{n}, \frac{1}{n+1}\right) \cup (-\infty, 0) \cup (1, \infty),
\]

which is open as it is the union of open sets. Thus its complement is closed.

**Definition 1-28.** \( x \in X \) is an **interior point** of \( X \) if, for some \( \varepsilon > 0 \), \( B_\varepsilon(x) \subset X \).

By definition, every element of an open set is an interior point. Therefore, every \( x \in X \) is an interior point if and only if \( X \) is open.

**Definition 1-29.** The **interior** of \( X \), \( \text{int}(X) \), is the union of all open sets contained in \( X \). Simply the interior of \( X \) is the set of all interior points of \( X \).

**Definition 1-30.** A point \( x \in X \) is in the **boundary point** of \( X \) if every \( \varepsilon > 0 \), \( B_\varepsilon(x) \cap X \neq \emptyset \) and \( B_\varepsilon(x) \cap X^C \neq \emptyset \).
A boundary point is a point on the border of $X$, and it does not matter whether it is a member of $X$ or not. The only requirement is an open ball around the boundary point contains an element of $X^c$.

**EXAMPLE 1-13.** A boundary point of $X$ has zero distance to the set. If not, there exists a $\varepsilon > 0$ such that $B_\varepsilon(x_0) \cap X = \emptyset$ for the boundary point $x_0$. That is not a boundary point. \hfill \Box

**THEOREM 1-9.** A set $X$ is closed if and only if it contains all its boundary points.

**Proof.** To reach a contradiction, suppose that there is $x$ that is a boundary point of $X$ but $x_0 \not\in X$. Then $X^c$ is open and $B_\varepsilon(x_0) \subset X^c$ for some $\varepsilon > 0$. We have $B_\varepsilon(x_0) \cap X = \emptyset$, which contradicts to that $x$ is a boundary point of $X$.

Conversely, the contrapositive of the necessity is that if set $X$ is open, then it contains no boundary points or $B_\varepsilon(x) \cap X^c = \emptyset$ for some $\varepsilon > 0$. That is the definition of an open set. \hfill ■

**EXAMPLE 1-14.** The set used in **EXAMPLE 1-11** does not contain one of the boundaries 0, and thus it is not a closed set while every element in the set of **EXAMPLE 1-12** is a boundary point. \hfill \Box

**THEOREM 1-10.** A set $X \subset \mathbb{R}^n$ is **compact** if and only if it is closed and bounded.

The formal definition is in **EXAMPLE 2-12**. The compactness guarantees the existence of a maximum and a minimum.

**THEOREM 1-11.** If $X \subset \mathbb{R}$ is compact, then $\sup X \in X$ and $\inf X \in X$ as well.

**Proof.** Since $X$ is bounded, there exists a upper bound. Suppose $s = \sup X \not\in X$. Since $X$ is closed or $X^c$ is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset X^c$. Then there exists $y \in B_\varepsilon(x)$ such that $y < s$ and $y > x$ for all $x \in X$, which is a contradiction. \hfill ■

**THEOREM 1-12.** A compact set has both a maximum and a minimum.

This is equivalent to **THEOREM 1-11**, and it states that compactness is sufficient for the existence of extrema.
1.4 CONVEX SET

Some optimality conditions are closely related to convexity. Here are the related definitions, and more details will be discussed later.

**Definition 1-31.** A set $X \subset \mathbb{R}^n$ is convex if, for any $x, y \in X$ and $\lambda \in [0,1]$, $\lambda x + (1 - \lambda)y \in X$.

For any two points in a subset of Euclidean space, there is the shortest line connecting them and the length of the line is the distance. If the set contains the line segment between any two points, we call it convex. On the other hand, the point on the line segment can be thought of as a weighted arithmetic mean of two points. In statistics, a great deal of mean related inequalities is in one way or the other a reflection of convexity.

**Example 1-15.** An open ball is a convex set. For any $y, z \in B_r(x)$ and $\lambda \in [0,1]$,

$$\|\lambda y + (1 - \lambda)z - x\| = \|\lambda(y - x) + (1 - \lambda)(z - x)\| \leq \lambda \|y - x\| + (1 - \lambda)\|z - x\| \leq r$$

Therefore, $\lambda y + (1 - \lambda)z \in B_r(x)$. \hfill $\square$

In terms of the boundary of a convex set, the convexity does not rule out straight edges. In this case, the intersection of the boundaries of two convex sets could be an infinite set. With strict convexity, the intersection is either empty or singleton.

**Definition 1-32.** A set $X \subset \mathbb{R}^n$ is strictly convex if, for any $x, y \in X$ and $\lambda \in (0,1)$, $\lambda x + (1 - \lambda)y$ is an interior point of $X$.

The definition can be extended to any arbitrary number of elements.

**Definition 1-33.** A convex combination of $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ is any point $x \in \mathbb{R}^n$ of the form $x = \lambda_1 x_1 + \lambda_2 x_2 + \ldots + \lambda_k x_k$ where $\lambda_i > 0$ is a scalar with $\lambda_1 + \ldots + \lambda_k = 1$.

**Theorem 1-13.** Let $X_1, \ldots, X_f$ be convex sets. The following operations preserve convexity.

(i) The Cartesian product of convex sets, $\prod_{j=1}^f X_j$, is convex.

(ii) The intersection of convex sets $\bigcap_{j=1}^f X_j$ is convex.
1.4 Real Numbers

Proof. (i) For any \( x, y \in \prod_{j=1}^{l} X_j \), since \( \lambda x_i + (1 - \lambda) y_i \in X_{ji} \) for all \( i, j \) and \( \lambda \in [0,1] \), we have \( \lambda x + (1 - \lambda) y \in \prod_{j=1}^{l} X_j \).

(ii) For any \( x, y \in \bigcap_{j=1}^{l} X_j \), since \( \lambda x + (1 - \lambda) y \in X_j \) for all \( j \) and \( \lambda \in [0,1] \), we have \( \lambda x + (1 - \lambda) y \in \bigcap_{j=1}^{l} X_j \).

Proof. (i) For any \( x, y \in \prod_{j=1}^{l} X_j \), since \( \lambda x_i + (1 - \lambda) y_i \in X_{ji} \) for all \( i, j \) and \( \lambda \in [0,1] \), we have \( \lambda x + (1 - \lambda) y \in \prod_{j=1}^{l} X_j \).

(ii) For any \( x, y \in \bigcap_{j=1}^{l} X_j \), since \( \lambda x + (1 - \lambda) y \in X_j \) for all \( j \) and \( \lambda \in [0,1] \), we have \( \lambda x + (1 - \lambda) y \in \bigcap_{j=1}^{l} X_j \).

DEFINITION 1-34. A convex hull of a set \( X \subset \mathbb{R}^n \) is the set of all convex combinations of points in \( X \) and is denoted by \( \text{co}X \).

\[
\text{co}X = \{ \lambda x + (1 - \lambda) y | x, y \in X, \lambda \in [0,1] \}
\]

A convex hull is the smallest convex set that contains the original set and is obtained by taking all possible convex combination of the points in a set.

Here are more terms that often appear in economic analyses.

DEFINITION 1-35. Let \( x_1, x_2, \cdots, x_k \in \mathbb{R}^n \) and \( x = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k \) where \( \lambda_i \) is a scalar.

(i) \( x \) is a linear combination of \( x_1, x_2, \cdots, x_k \) for arbitrary scalars \( \lambda_i \).

(ii) \( x \) is an affine combination of \( x_1, x_2, \cdots, x_k \) if \( \lambda_1 + \cdots + \lambda_k = 1 \).

(iii) \( x \) is a conical combination of \( x_1, x_2, \cdots, x_k \) if \( \lambda_i \geq 0 \).

If a set contains the linear combination of elements in the set, it is a subspace (discussed in Section 7) Affine sets and convex cones (strictly speaking conical hull) is closed under affine and conical combinations, respectively.

DEFINITION 1-36. A set \( X \subset \mathbb{R}^n \) is affine if, for any \( x, y, z \in X \), there exists a \( \lambda \in \mathbb{R} \) such that \( z = \lambda x + (1 - \lambda) y \).

An affine set contains the line through any two distinct points in the set. The set is convex,
**EXAMPLE 1-16.** A general equation of a hyperplane is $a_1x_1 + \cdots + a_nx_n = c$ and is an affine set. A line and plane are also affine sets.

**DEFINITION 1-37.** A set $X \in \mathbb{R}^n$ is called a **cone** if for $x \in X$ and $a > 0$ the ray $ax$ is in $X$. A cone may or may not include the origin.

The definition considers only positive scalar multiplication. A convex cone requires it is closed under the positive conical combination.

**DEFINITION 1-38.** A cone $X \in \mathbb{R}^n$ is a **convex cone** if $ax + by \in X$ for all $x, y \in X$ and $a, b > 0$.

A set is a convex cone if and only if it is closed under addition and positive scalar multiplication. A subspace is affine, convex, and is a convex cone.
2 SEQUENCES AND LIMITS

2.1 SEQUENCES

**Definition 2-1.** A **sequence** is a collection of points in $\mathbb{R}^n$ indexed by positive integers, and we write $\{x_n\}_{n=1}^{\infty}$.

A sequence can be thought of as a rule that associates points $x_n \in \mathbb{R}^n$ to each positive integer $n$. The set of indices could be restricted to an interval of integers to form a finite sequence, but a sequence usually means an infinite sequence.

For the sake of simplicity, the discussion considers sequences of real numbers, but most of the definitions and theorems in this section hold for $x \in \mathbb{R}^n$ replacing $| \cdot |$ with $\| \cdot \|$ as long as the expressions make sense.

**Definition 2-2.** The **limit** of a sequence $\{x_n\}_{n=1}^{\infty}$, denoted by $\lim_{n \to \infty} x_n$, is equal to $L$ if $L$ has the property, for each $\varepsilon > 0$, there is an $N$ such that $n > N$ implies $|x_n - L| < \varepsilon$.

The definition does not say anything about the way to find a limit. Instead, we first guess a limit and check whether it satisfies the definition. It is good to know simple sequences with their limits.

**Example 2-1.** If $x_n = c$ for all $n$, then $\lim_{n \to \infty} x_n = c$. That is, for a given $\varepsilon > 0$, let $N = 1$. For any $n > N$, $|x_n - c| = 0 < \varepsilon.$

**Example 2-2.** If $x_n = 1/n$, then $\lim_{n \to \infty} x_n = 0$. For a given $\varepsilon > 0$, let $N$ be an integer greater than $\frac{1}{\varepsilon} + 1$. If $n > N$, then $|x_n - L| = \left|\frac{1}{n} - 0\right| = \frac{1}{n} < \frac{1}{N} < \varepsilon.$

**Example 2-3.** If $x_n = (-1)^n$, then $\lim_{n \to \infty} x_n$ does not exist. For any $N$, since there is an odd or an even number that is greater than $N$, $|x_n - L| = |1 + L|$ or $|1 - L|$. For any $L$, $|x_n - L|$ is always greater than or equal to $1$.

If a sequence diverges, we do not say that it “converges to infinity.”
DEFINITION 2-3. A real sequence \( \{x_n\}_{n=1}^{\infty} \) tends to infinity (negative infinity) if, for any \( K \in \mathbb{R} \), there is an \( N \) such that \( x_n > K \) (\( x_n < K \)) for all \( n > N \). We write \( \lim_{n \to \infty} x_n = \infty \) (\( \lim_{n \to \infty} x_n = -\infty \)).

If a limit exists, it is unique.

THEOREM 2-1. If \( \lim_{n \to \infty} x_n = L \) and \( \lim_{n \to \infty} x_n = M \), then \( L = M \).

Proof. To reach a contradiction, suppose \( |L - M| > 0 \) and let \( \varepsilon = |L - M|/2 \).

Then there exists \( N \) such that, for any \( n > N \), \( |x_n - L| < \varepsilon/2 \) and \( |x_n - M| < \varepsilon/2 \). However, by triangular inequality,

\[
|L - M| \leq |x_n - L| + |x_n - M| < 2\varepsilon,
\]

which contradicts to \( |L - M| = 2\varepsilon \).

THEOREM 2-2. If \( \lim_{n \to \infty} x_n = L \), then.

THEOREM 2-3. If \( \lim_{n \to \infty} x_n = L \) and \( \lim_{n \to \infty} y_n = M \), then

(i) \( \lim_{n \to \infty} \lambda x_n = \lambda L \) for all \( \lambda \in \mathbb{R} \)

(ii) \( \lim_{n \to \infty} (x_n + y_n) = L + M \)

(iii) \( \lim_{n \to \infty} (x_n \cdot y_n) = L \cdot M \)

(iv) \( \lim_{n \to \infty} (x_n/y_n) = L/M \) where \( M \neq 0 \). Note that \( \lim_{n \to \infty} (x_n/y_n) \) still could make sense if \( y_n = 0 \) for some \( n \).

Proof. For a given \( \varepsilon > 0 \), there exists \( N(\varepsilon) \) such that \( |x_n - L| < \varepsilon \) for all \( n > N(\varepsilon) \). For given \( \lambda \) and \( \varepsilon > 0 \), let \( \delta = \varepsilon/|\lambda| \) (if \( \lambda = 0 \), then obvious). Then, for a given \( \delta > 0 \), \( |\lambda x_n - \lambda L| = |\lambda| \times |x_n - L| < \delta \) for all \( n > N(\varepsilon) \).

(ii) Given \( \varepsilon > 0 \), there exist integers \( N_1, N_2 \) such that

\[
n \geq N_1 \Rightarrow |x_n - L| < \frac{\varepsilon}{2}, \quad n \geq N_2 \Rightarrow |y_n - M| < \frac{\varepsilon}{2}.
\]

Taking \( N = \max(N_1, N_2) \) gives, for all \( n > N \),
\[ |(x_n + y_n) - (L + M)| \leq |x_n - L| + |y_n - M| < \varepsilon. \]

(iii) Note that
\[ x_n y_n - L \cdot M = (x_n - L)(y_n - M) + L(y_n - M) + M(x_n - L) \]

Given \( \varepsilon > 0 \), there exist integers \( N_1, N_2 \) such that
\[ n \geq N_1 \implies |x_n - L| < \sqrt{\varepsilon}, \quad n \geq N_2 \implies |y_n - M| < \sqrt{\varepsilon}. \]

Taking \( N = \max(N_1, N_2) \) gives, for all \( n > N \),
\[ |(x_n - L)(y_n - M)| \leq \varepsilon, \]

and thus \( \lim_{n \to \infty} (x_n - L)(y_n - M) = 0. \) Combining the results shows
\[ \lim_{n \to \infty} (x_n y_n - LM) = \lim_{n \to \infty} (x_n - L)(y_n - M) + \lim_{n \to \infty} L(y_n - M) + \lim_{n \to \infty} M(x_n - L) = 0. \]

(iv) We will show \( \lim_{n \to \infty} (1/x_n) = 1/L. \) Choose \( m \) such that \( |x_n - L| < |L|/2 \) for all \( n > m. \) Then \( |x_n| > |L|/2. \) Given \( \varepsilon > 0, \) there exist integers \( N > m \) such that, for all \( n \geq N, \)
\[ |x_n - L| < \frac{1}{2} |L|^2 \varepsilon. \]

By the two inequalities,
\[ \left| \frac{1}{x_n} - \frac{1}{L} \right| = \left| \frac{x_n - L}{x_n \cdot L} \right| < \frac{2}{|L|^2} |x_n - L| < \varepsilon. \]

\[ \textbf{THEOREM 2-4 (Sandwich theorem)} \] If \( \lim_{n \to \infty} x_n = L \) and \( \lim_{n \to \infty} z_n = L, \) and there exists \( N \) such that \( x_n \leq y_n \leq z_n \) for all \( n > N, \) then \( \lim_{n \to \infty} y_n = L. \)

\[ \textbf{EXAMPLE 2-4.} \] \( \lim_{n \to \infty} (\sqrt{n + 1} - \sqrt{n}) = 0. \)
\[ 0 \leq \sqrt{n + 1} - \sqrt{n} = \frac{n + 1}{\sqrt{n + 1}} - \frac{n}{\sqrt{n}} = \frac{(n + 1)\sqrt{n} - n\sqrt{n + 1}}{\sqrt{n + 1} \times \sqrt{n}} \leq \frac{(n + 1)\sqrt{n} - n\sqrt{n}}{\sqrt{n + 1} \times \sqrt{n}} = \frac{1}{\sqrt{n + 1}} \leq \frac{1}{\sqrt{n}} \]

\[ \textbf{DEFINITION 2-4 (Bounded Sequence)} \] A sequence is \textbf{bounded} if \( \{x_n\}_{n=1}^{\infty} \) is bounded or if there exists \( K \) such that \( |x_n| < K \) for all \( n. \)
**Definition 2-5.** A sequence is **monotonic** if it is either **monotonically increasing** \((x_n \leq x_{n+1} \text{ for all } n)\) or **monotonically decreasing** \((x_n \geq x_{n+1} \text{ for all } n)\).

**Theorem 2-5.** Any bounded monotonic sequence has a limit. Especially, if \(x_n\) is bounded and monotonically increasing, then \(\lim_{n \to \infty} x_n\) is the least upper bound of \(\{x_n\}\).

**Proof.** Since \(\{x_n\}\) is bounded, by the supremum property there is a least upper bound, say \(L\). For a given \(\varepsilon > 0\), there exists \(N\) such that \(x_N > L - \varepsilon\) because otherwise, \(L\) is not a least upper bound. Since \(L\) is the least upper bound and \(\{x_n\}\) is monotonically increasing, \(L - \varepsilon < x_n \leq L\) for all \(n > N\) or \(|x_n - L| < \varepsilon\).

A sequence does not converge if \(\lim_{n \to \infty} x_n\) does not exist. Sometimes we say that a sequence diverges to \(∞\) (written \(\lim_{n \to \infty} x_n = ∞\)) if for any \(K\), there is \(N\) such that \(x_n > K\) whenever \(n > N\). By this definition, \(\{x_n\} = \{n\}\) diverges to \(∞\) and \(\{x_n\} = \{(-1)^n\}\) does not.

**Theorem 2-6.** If \(\{x_n\}\) converges to the limit \(L\), then the sequence is bounded.

**Proof.** Since \(\lim_{n \to \infty} x_n = L\), then there is \(N\) such that \(|x_n - L| \leq 1\) (\(= \varepsilon\)) for all \(n > N\). Then every element of \(\{x_n\}\) is bounded above by \(\max(x_1, \ldots, x_n, L + 1)\) and below by \(\min(x_1, \ldots, x_n, L - 1)\).

A sequence of in \(\mathbb{R}^n\) can be thought of as a collection of a sequence of real numbers. Therefore, the results can be applied to each component by component and even monotonicity can be defined in a similar way. As a matter of fact, the result of (iii) in Theorem 2-3 is the only one that cannot be extended to sequences in \(\mathbb{R}^n\).

**Definition 2-6.** For a sequence \(\{x_n\}\) of real numbers, let

\[
\alpha_n = \sup\{x_k | k \geq n\} = \sup\{x_n, x_{n+1}, x_{n+2}, \ldots\}, \quad \beta_n = \inf\{x_k | k \geq n\}
\]

Then

\[
\limsup_{n \to \infty} x_n = \begin{cases} +\infty & \text{if } \alpha_n = +\infty \text{ for all } n \\ \lim_{n \to \infty} \alpha_n & \text{otherwise} \end{cases}
\]

\[
\liminf_{n \to \infty} x_n = \begin{cases} -\infty & \text{if } \beta_n = -\infty \text{ for all } n \\ \lim_{n \to \infty} \beta_n & \text{otherwise} \end{cases}
\]
By definition, it must be the case that either $\alpha_n = \infty$ for all $n$ or $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \ldots$, and either $\beta_n = \infty$ for all $n$ or $\beta_1 \leq \beta_2 \leq \beta_3 \leq \ldots$.

**Example 2-5.** $\limsup_{n \to \infty} \sin x_n = 1$ and $\limsup_{n \to \infty} \cos x_n = -1$.

The big-$O$ and little-$o$ notation are standard notation when we describe the asymptotic behavior of functions. Basically, it shows the growth rate of a function or the order of a function.

**Definition 2-7.** $O(g(x))$ is a set of all functions $f(x)$ such that: for some $k > 0$, if there exists $x_0$ such that $0 \leq f(x) \leq k \cdot g(x)$ holds for all $x > x_0$.

$$\limsup_{x \to \infty} \frac{f(x)}{g(x)} < \infty$$

$o(g(x))$ is a set of all functions $f(x)$ such that: for any $k > 0$, if there exists $x_0$ such that $0 \leq f(x) \leq k \cdot g(x)$ holds for all $x > x_0$,

$$\limsup_{x \to \infty} \frac{f(x)}{g(x)} = 0$$

**Example 2-6.** $x^2 \in O(x^3)$ implies $x^2/x^3$ is bounded as $x \to \infty$. Thus $O(x^2) = O(ax^2)$ for any positive $a$. Similarly, $\ln x \in o(x)$ implies $\ln x / x$ vanishes $x \to \infty$.

The standard notion is for the case of $x \to \infty$, but they can be used for a specific range of $x$. For instance, $x \to a$ or $x \in B_r(x_0)$. Also, the definition can be extended to general functions with $0 \leq |f(x)| \leq k \cdot |g(x)|$.

**Example 2-7.**

$$f(x) = 1 + x + x^2 + x^3 + \ldots = 1 + x + x^2 + O(x^3) = 1 + x + o(x^{-1}), \quad \text{as } x \to 0$$

The second expression $f(x) = 1 + x + x^2 + O(x^3)$ means that the absolute value of the error $f(x) - (1 + x + x^2)$ is at most some constant times $|x^3|$ when $x$ is close enough to 0. The last expression means $\frac{f(x) - (1 + x)}{x}$ vanishes as $x \to 0$. 

Math for Economists, Sogang University
2.2 Subsequences

**Definition 2-8.** A subsequence of a sequence \( \{x_n\} \) is a sequence \( \{x_{n_i}\} \) where \( \{n_i\} \) is an any increasing sequence of natural numbers, \( n_1 < n_2 < \cdots \).

**Theorem 2-7.** Every sequence of real numbers has a monotone subsequence.

**Proof.**

Case 1: \( \{x_n: n > N\} \) has a maximum for every \( N \) (the maximum is decreasing as \( N \) increases). To construct a monotone decreasing subsequence, let \( x_{n_1} = \max\{x_n: n > 1\} \), \( x_{n_2} = \max\{x_n: n > n_1\} \), \( \ldots \), \( x_{n_{i+1}} = \max\{x_n: n > n_i\} \) so that the next element of the subsequence is the biggest of all remaining elements of the sequence. The result follows from that this subsequence is decreasing (\( x_{n_1} = \max\{x_n: n > n_{i-1}\} \), \( x_{n_{i+1}} = \max\{x_n: n > n_i > n_{i-1}\} \)).

Case 2: If \( \{x_n: n > N\} \) has no maximum for some \( N_0 \), then it should not have a max for all \( N > N_0 \). Pick \( n_1 = N_0 + 1 \) so that \( x_{n_1} = x_{N_0+1} \). Since \( \{x_n: n > n_1\} \) has no maximum, choose a \( x_{n_2} \) such that \( x_{n_2} > x_{n_1} \) and \( n_2 > n_1 \). This is a monotonically increasing sequence. 

Theorem 2-8 (Bolzano-Weierstrass Theorem) Every bounded sequence of real numbers has a convergent subsequence.

**Proof.** By Theorem 2-7, one can construct a monotone subsequence, and Theorem 2-5 can be applied since it is bounded.

**Theorem 2-9 (Bolzano-Weierstrass Theorem)** Every bounded sequence in \( \mathbb{R}^n \) has a convergent subsequence.
Proof. According to Theorem 2.8, the sequence of the first components of \( \{x_n\} \) denoted by \( \{x_{n1}\} \) has a convergent subsequence \( \{x_{m1}\} \). Then the sequence of the second components in \( \{x_{m1}\} \) has a convergent subsequence \( \{x_{m2}\} \). Repeating the construction of subsequences \( k \) times, we can construct a subsequence \( \{x_{m,k}\} \) of \( \{x_n\} \) in which the all sequences of components converge. Therefore, the subsequence itself converges.  

2.3 Compact Set

Definition 2.9. Let \( X \subseteq S \). \( x_0 \in S \) is a limit point of \( X \) if, for all \( \varepsilon > 0 \),

\[
B_\varepsilon(x_0) \cap X \setminus x_0 \neq \emptyset
\]

Note that \( x \) need not be a member of \( X \), but \( B_\varepsilon(x) \) should contain at least one point of \( X \) different from \( x \). This feature distinguishes it from a boundary point. If a set is a singleton, it does not have a limit point, but the unique element is a boundary point. We should be able to construct a convergent sequence \( x_n \to x \) with \( x_n \in X \setminus x \) if \( x \) is a limit point of \( X \).

Example 2.8. The set \( \left\{1, \frac{1}{2}, \frac{1}{3}, \ldots\right\} \) in Example 1-11 has one and only limit point of \( 0 \) that is the limit of the sequence \( x_n = \frac{1}{n} \).

Theorem 2.10. \( x_0 \) is a limit point of \( X \) if and only if there exists a sequence of \( x_n \in X \setminus x_0 \) converging to \( x_0 \).

Proof. Choose \( x_n \in B_{1/n}(x_0) \cap X \). Conversely, by the definition of converging sequence, for any \( \varepsilon < 0 \), there exists \( N \) such that \( x_n \in B_\varepsilon(x_0) \) whenever \( n > N \). Since \( x_n \in X \), \( B_\varepsilon(x_0) \cap X \) is not empty.

Definition 2.10. \( X \) is closed if \( X \) contains all of its limit points.

Theorem 2.11. A set is closed if and only if it contains the limit of every convergent sequence.

Let \( X \) be a closed set, and let \( \{x_n\} \subseteq X \) be a sequence converging to \( x_0 \). Suppose \( x_0 \notin X \). Since \( X^c \) is open, there is an \( \varepsilon > 0 \) for which \( B_\varepsilon(x_0) \subseteq X^c \). Since \( \{x_n\} \to x_0 \), for a sufficiently large \( n \), we have both \( x_n \in X \) and \( x_n \in B_\varepsilon(x_0) \), a contradiction.
For the converse, suppose \( X \) is not closed. Since \( X^c \) is not open, there is at least one element \( x_0 \in X^c \) such that every \( B_\varepsilon(x_0) \) contains at least one element of \((X^c)^c = X\). For every \( n \in \mathbb{N} \), let \( x_n \in B_{1/n}(x_0) \cap X \). Then the sequence converges to \( x_0 \not\in X \), i.e., \( x_0 \) is a limit point of \( X \) but is not in \( X \). \( X \) does not contain all its limit points, a contradiction. ■

According to the theorem, the empty set is closed and so is any finite set.

**EXAMPLE 2-9. Proof of Theorem 1-7 (closed set)**

If \( X \) is closed, every \( x \in X^c \) is not a limit point of \( X \). Then there exists \( \varepsilon > 0 \) such that \( B_\varepsilon(x) \cap X = \emptyset \) and \( B_\varepsilon(x) \cap X^c \neq \emptyset \). Therefore, every \( x \in X^c \) is an interior point of \( X^c \) and \( X^c \) is open.

Conversely, let \( x \) be a limit point of \( X \). Then for any \( \varepsilon > 0 \), \( B_\varepsilon(x) \cap X \neq \emptyset \) and \( x \) is not an interior point of \( X^c \). Since \( X^c \) is open, \( x \not\in X^c \) or \( x \in X \). ■

**DEFINITION 2-11.** The **closure** of a set \( X \), denoted by \( \bar{X} \), is the union of \( X \) and all of the limit points of \( X \).

Therefore, a set is closed if and only if it coincides with its closure. Note that every limit point is in a closure, but the converse is not true. The closure of a singleton set is the set itself but is not a limit point.

**EXAMPLE 2-10.** The sequence \( x_n = 1/n \) converges to 0, and thus the set \( X = (0, 1) \) is not closed.

**EXAMPLE 2-11.** Since \( \mathbb{Q} \) is dense in \( \mathbb{R} \), we can construct a sequence \( x_i \in B_{1/n}(\sqrt{2}) \subset \mathbb{Q} \), which converges to \( \sqrt{2} \). The closure of \( \mathbb{Q} \) is \( \mathbb{R} \).

**DEFINITION 2-12.** \( X \subset \mathbb{R} \) is a **compact set** if every sequence in \( X \) has a convergent subsequence converging to a point in \( X \).

**EXAMPLE 2-12. Proof of Theorem 1-10 (Compact Set)**

Note that bounded sequence has a convergent subsequence and Theorem 2-11 shows the equivalence between closed set and limit points. Therefore, to prove Theorem 1-10, it is sufficient to show that every sequence in \( X \) has a convergent subsequence, then \( X \) is bounded. The contrapositive is true because we can always construct a monotonically increasing sequence if \( X \) is not bounded as shown in Case 2 in the proof of Theorem 2-7. ■
2.4 CONVERGENCE

**THEOREM 2-12.** A sequence of real numbers converges if and only if every subsequence of the sequence converges.

**Proof.** Suppose \( \lim_{n \to \infty} x_n = L \). That is, for a given \( \varepsilon > 0 \), there exists \( N \) such that \( |x_n - L| < \varepsilon \) whenever \( n > N \). Since \( \{n_i\} \) is not bounded, we can pick \( I \) so that \( n_i > N \) whenever \( i > I \). Therefore, there exists \( I \) such that \( |x_{n_i} - L| < \varepsilon \) whenever \( i > I \). Conversely, any sequence is a subsequence by itself. \( \blacksquare \)

**COROLLARY 2-13.** If the sequence \( \{x_{n_i}\} \) has two subsequences converging to different limits, then the sequence \( \{x_n\} \) does not converge.

**DEFINITION 2-13.** A sequence \( \{x_n\} \) is a **Cauchy sequence** if, for each \( \varepsilon > 0 \), there is \( N \) such that for any \( n, m > N \), \( |x_n - x_m| < \varepsilon \).

**THEOREM 2-14.**

(i) If \( \{x_n\} \) is a Cauchy sequence, then \( \{x_n\} \) is bounded.

(ii) If \( \{x_n\} \) converges, then \( \{x_n\} \) is a Cauchy sequence.

(iii) Every Cauchy sequence converges.

**Proof.** (i) There exists a \( N \) such that \( |x_n - x_m| < 1 \) for all \( m, n > N \). The sequence is bounded above by \( \max(x_1, \ldots, x_N, x_N + 1) \) and below by \( \min(x_1, \ldots, x_N, x_N - 1) \).

(ii) For any \( \varepsilon > 0 \), there exists \( N \) such that \( |x_n - L| < \varepsilon/2 \) for \( n > N \) and \( |x_m - L| < \varepsilon/2 \) for \( m > N \). By triangle inequality, \( |x_n - x_m| \leq |x_n - L| + |x_m - L| < \varepsilon \).

(iii) Since \( \{x_n\} \) is bounded by (i), it has a convergent subsequence, \( \{x_{n_i}\} \) with \( \lim_{i \to \infty} x_{n_i} = L \). For any \( \varepsilon > 0 \), there exists a \( N \) such that, for all \( n_i > N \), \( |x_{n_i} - L| < \varepsilon/2 \). Since \( \{x_n\} \) is a Cauchy sequence, there exists \( M \) such that \( |x_n - x_m| < \varepsilon/2 \) for all \( m, n > M \). Letting \( n_i = m \) and \( R = \max(N, M) \), for any \( n > R \),

\[
|x_n - L| = |x_n - x_m + x_{n_i} - L| \leq |x_n - x_m| + |x_{n_i} - L| < \varepsilon.
\]  \( \blacksquare \)
2.5 **SERIES**

**DEFINITION 2-14.** A series is the sum of the elements of a sequence.

**EXAMPLE 2-13.** The harmonic series is not bounded.

\[
\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \cdots
\]

\[
> 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots
\]

\[
= 1 + \left(\frac{1}{2}\right) + \left(\frac{1}{4}\right) + \left(\frac{1}{8}\right) + \left(\frac{1}{16}\right) + \cdots
\]

\[
= 1 + \frac{1}{2} + \frac{1}{2} + \cdots = \infty
\]

In the case of series, since the number of terms is infinite, rearranging terms could alter the limit. For instance,

\[
\sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) = \ln 2
\]

\[
\sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) = \ln 2
\]

\[
1 + \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} + \cdots = \frac{3}{2} \ln 2
\]

In the second series, two positive terms are followed by a single negative term.

**DEFINITION 2-15.** Consider a series \( \sum_{n=1}^{\infty} x_n \) where \( x_n \) is an element of a normed space. If the series \( \sum_{n=1}^{\infty} \|x_n\| \) is bounded, it is said to converge absolutely, or that it is absolutely convergent.

**THEOREM 2-15.** If a series converges absolutely, it converges. The converse is not true.

**Proof.** Suppose \( \sum_{n=1}^{\infty} a_n \) is absolutely convergent. Let \( A_i = |a_1| + \cdots + |a_i| \) be the sequence of the partial sum of norms, and \( S_i = a_1 + \cdots + a_i \) is the sequence of the partial sum. Since the series converges absolutely, there exists an integer \( N \) such that for \( n > m > N \)

\[
|T_n - T_m| = |a_n| + \cdots + |a_{m+1}| < \varepsilon
\]

By the triangular inequality
\[ |S_n - S_m| = |a_1 + \cdots + a_i| \leq |a_n| + \cdots + |a_{m+1}| = |T_n - T_m| < \varepsilon \]

The sequence of \( S_i \) is Cauchy and it converges.

The converse is not true because \( \sum_{n=1}^{\infty} (-1)^{n+1}/n \) converges but \( \sum_{n=1}^{\infty} 1/n \) does not.
3 Univariate Function

3.1 Function

**Definition 3-1.** A function $f$ from a set $X$ to a set $Y$ is a mapping that assigns to each element of $X$ exactly one element of $Y$.

An alternative definition is that a function $f$ from $X$ to $Y$ is a subset of the Cartesian product $X \times Y$ with the property that every element of $X$ is the first component of one and only one ordered pair in the subset.

We can think of the element of $X$ as a key and the associated element in $Y$ as a value in a dictionary. Two different words could have the same meaning but a word could not have different meanings.

This information is summarized in $f: X \to Y$. When $X, Y \subset \mathbb{R}$, the function $f(x)$ is called a real-valued function of a real variable. A function of one variable is often called a univariate or univariable function.

**Definition 3-2.** Consider a function $f: X \to Y$.

(i) The set $X$ is called the **domain** of function $f$, the set of argument values for which the function is defined.

(ii) The set $Y$ is called the **codomain** (or range) of function $f$, a set that includes all the possible outcomes of a function.

(iii) The **range** of function $f$ is the set of $f(X)$, the complete set of all possible values of the function.

(iv) For any subset of $Z \subset X$, $f(Z) = \{f(x) | x \in Z\}$ is called the **image** of the set $Z$ for function $f(x)$.

(v) The **inverse image** of a set $W \subset Y$, $f^{-1}(W) = \{x \in X | f(x) \in W\}$, is the set of points in $X$ that maps to points in $W$. 
Note that the inverse image does not necessarily define a function from \( Y \) to \( X \), because it could be that there are two distinct elements of \( X \) that map to the same point in \( Y \).

**Example 3.1.** \( f^{-1}(f(Z)) \neq Z \). Try \( f(x) = x^2 \) and \( Z = [0,1] \).

The image has linear property in the sense that \( f(A \triangle B) = f(A) \triangle f(B) \) for set operators \( \triangle \).

**Theorem 3.1.** Suppose \( f: X \to Y \) and \( A, B \subset X \) and \( C, D \subset Y \). \( A_i \) is a set such that \( \bigcup_{i \in I} A_i \subset X \) with a finite set \( I \).

(i) \( f(A \cap B) \subset f(A) \cap f(B) \), and \( f(\bigcap_{i \in I} A_i) \subset \bigcap_{i \in I} f(A_i) \)

(ii) \( f(A \cup B) = f(A) \cup f(B) \), and \( f(\bigcup_{i \in I} A_i) = \bigcup_{i \in I} f(A_i) \)

(iii) \( f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D) \), and \( f^{-1}(\bigcap_{i \in I} C_i) = \bigcap_{i \in I} f^{-1}(C_i) \)

(iv) \( f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D) \), and \( f^{-1}(\bigcup_{i \in I} A_i) = \bigcup_{i \in I} f^{-1}(A_i) \)

(v) If \( C \subset D \), \( f^{-1}(C) \subset f^{-1}(D) \)

(vi) \( f^{-1}(C - D) \subset f^{-1}(C) - f^{-1}(D) \)

**Proof.**

(i) Let \( y \in f(A \cap B) \). There exists an \( x \in A \cap B \) such that \( f(x) = y \). Since \( x \in A \) and \( x \in B \), that implies \( f(x) \in f(A) \) and \( f(x) \in f(B) \), or \( y = f(x) \in f(A) \cap f(B) \). The converse is not true. An counterexample is \( f(x) = x^2 \) with \( A = [-1,0] \) and \( B = [0,1] \).

(ii) \( f(x) \in f(A \cup B) \iff f(x) \in f(A) \) or \( f(x) \in f(B) \iff f(x) \in f(A) \cup f(B) \)

(iii) \( x \in f^{-1}(C \cap D) \iff f(x) \in C \cap D \iff f(x) \in C \) and \( f(x) \in D \)

\[
\iff x \in f^{-1}(C) \text{ and } x \in f^{-1}(D) \iff f(x) \in f^{-1}(C) \cap f^{-1}(D)
\]

(iv) \( f^{-1}(C \cup D) \iff f(x) \in C \cup D \iff f(x) \in C \) or \( f(x) \in D \)

\[
\iff x \in f^{-1}(C) \text{ or } x \in f^{-1}(D) \iff f^{-1}(C) \cup f^{-1}(D)
\]

(v) Let \( y \in C \). Then there exists a \( x \in f^{-1}(C) \) such that \( f(x) = y \). Since \( C \subset D \), \( x \in f^{-1}(D) \).
(vi) Let $y \in (C - D)$. Then there exists a $x \in f^{-1}(C - D)$ such that $f(x) = y$. That is, $x \in f^{-1}(C)$ but $x \in f^{-1}(D)$.

**Definition 3-3.** If $f(X)$ is bounded above, it is said that $f(x)$ is bounded above. Similarly, we say that $f(x)$ is bounded below if $f(X)$ is bounded below. If both, $f(X)$ is bounded.

**Definition 3-4.** A function is monotonically increasing if

$$x < y \Rightarrow f(x) < f(y)$$

**Definition 3-5.** A function $f: X \to Y$ is a one-to-one or an injective function if

$$f(x) \neq f(y) \text{ for any } x, y \in X \text{ with } x \neq y$$

It is obvious that a monotonically increasing function is injective.

If a function is injective, we can refine the result of Theorem 3-1, (i).

**Theorem 3-2.** Suppose $f: X \to Y$ is injective and $A, B \subset X$. Then $f(A \cap B) = f(A) \cap f(B)$ and $f(\bigcap_{i \in I} A_i) = \bigcap_{i \in I} f(A_i)$

**Proof.** Let $y \in f(A) \cap f(B)$. There exists $y \in f(A)$ and $y \in f(B)$, and $x_1 \in A$ and $x_2 \in B$ such that $f(x_1) = y$ and $f(x_2) = y$. Since $f$ is injective, we have $x_1 = x_2$ or $x_1 \in A$ and $x_1 \in B$. That implies that $y = f(x_1) = f(A \cap B)$. The result follows from the proof of Theorem 3-1, (i).

**Definition 3-6.** A function $f: X \to Y$ is an onto or a surjective function if

$$Y = f(X)$$

If the range is known, one can set the range of a function to the codomain so that the function is surjective.

**Definition 3-7.** A function $f: X \to Y$ is a bijective function if it is both one-to-one and onto.

**Example 3-2.** $f(x) = x^2$ is not injective, but if the domain and codomain is restricted to nonnegative real numbers, it is bijective.

**Definition 3-8.** Two sets $A$ and $B$ are numerically equivalent (or have the same cardinality) if there is a bijection $f: A \to B$. 

Math for Economists, Sogang University
When sets are finite, the elements of the sets can be uniquely matched and paired. When they are infinite sets, the similar property holds without uniqueness.

A set is countable if it is numerically equivalent to the set of natural numbers.

**Example 3-3.** A random variable is a function from the set of outcomes in an experiment to the set of real numbers. Roughly, if the image of a random variable is countable, we call it a **discrete** random variable. If uncountable, it is a **continuous** random variable.

**Example 3-4.** The set of integers is countable. Define \( f : \mathbb{N} \to \mathbb{Z} \) by

\[
\begin{align*}
f(1) &= 0 \\
f(2) &= 1 \\
f(3) &= -1 \\
&\vdots \\
f(n) &= (-1)^n [n/2]
\end{align*}
\]

where \([x]\) is the greatest integer less than or equal to \(x\). Note that both sets are infinite, and the statement like “One half of the elements of \(\mathbb{Z}\) are in \(\mathbb{N}\)” is meaningless.

**Example 3-5.** The set of rational number \(\mathbb{Q}\) is countable.

\[
\mathbb{Q} = \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, n \neq 0 \right\} = \left\{ \frac{m}{n} : m \in \mathbb{Z}, n \in \mathbb{N} \right\}
\]

<table>
<thead>
<tr>
<th>(n)</th>
<th>0</th>
<th>1</th>
<th>1/2</th>
<th>1/3</th>
<th>1/4</th>
<th>1/5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>2</td>
<td>-2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/2</td>
<td>-1/2</td>
<td>1</td>
<td>-1</td>
<td>3/2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1/3</td>
<td>-1/3</td>
<td>2/3</td>
<td>-2/3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1/4</td>
<td>-1/4</td>
<td>1/2</td>
<td>-1/2</td>
<td>3/4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1/5</td>
<td>-1/5</td>
<td>2/5</td>
<td>-2/5</td>
<td>3/5</td>
</tr>
</tbody>
</table>

**Figure 3-1 Counting Rational Numbers**

By moving back and forth on off-diagonals omitting repeats, we have a bijection function:

\[
\begin{align*}
f(1) &= 0 \\
f(2) &= 1 \\
f(3) &= 1/2
\end{align*}
\]
Therefore, \( \mathbb{Q} \) is countable.

**Definition 3-9.** A function \( f: X \to Y \) is **invertible** if, for each \( y \in Y \), there is a unique \( x \in X \) that solves the equation \( y = f(x) \).

A function is invertible if the inverse image mapping is a function from \( Y \) to \( X \). Then, for every \( y \in Y \), an inverse function \( f^{-1}(y) \) is defined to be the point \( x \in X \) for which \( f(x) = y \). By definitions, an invertible function is a bijection. If a function is not a constant function, a proper restriction of the domain could make most functions at least locally invertible.

**Theorem 3-3.** A function \( f: X \to Y \) is invertible if and only if it is a bijective.

**Proof.** Suppose \( f(x) \) is not injective. Then there exists some \( a \neq b \) such that \( f(a) = f(b) \), but the solution of \( x \) to \( f(a) = f(x) \) is not unique. So the function is injective. And every \( y \in Y \) is associated with a (unique) \( x \in X \), it is surjective. Converse is the definition of an invertible function: there is a unique \( x \in X \) that solves the equation \( y = f(x) \) (injection) for each \( y \in Y \) (surjection).

Section 3.3.2 discusses properties of an inverse function.

### 3.2 Basic Functions

#### 3.2.1 Polynomial Functions

**Definition 3-10.** Consider a function \( f: X \to Y \).

(i) \( f(x) \) is a **constant** function if \( f(x) = c \) for all \( x \in X \).

(ii) \( f(x) \) is an **identity** function if \( f(x) = x \) for all \( x \in X \), and often denoted by \( \text{id}_X(x) \).

(iii) \( f(x) \) is a **linear** function if \( f(ax + by) = af(x) + bf(y) \) for all \( x, y \in X \).

(iv) \( f(x) \) is an **affine** function if \( f(x) - f(y) = L(x - y) \) where \( L(\cdot) \) is a linear function.

In many economics applications, \( f(x) = a + bx \) is often called “linear function.” However, strictly speaking, an affine function is not linear. Both linear and affine function have a
constant slope at any $x$, but a linear function is restricted by $f(0) = 0$. An affine transformation of a function $f(x)$ has a form of $af(x) + b$.

**Function Arithmetic and Notations**

Given two functions of $f(x)$ and $g(x)$, we can define new functions.

(i) Scalar multiplication: $\lambda f$ denotes $(\lambda f)(x) = \lambda f(x)$

(ii) Addition: $(f + g)$ denotes $(f + g)(x) = f(x) + g(x)$

(iii) Product: $(f \cdot g)$ denotes $(f \cdot g)(x) = f(x) \cdot g(x)$

(iv) Quotient: $(f/g)$ denotes $(f/g)(x) = f(x)/g(x)$ for $g(x) \neq 0$ for all $x$.

**Definition 3-11.** A **polynomial** is an expression that can be built with scalar multiplication, addition, product, and a non-negative integer power of variables. A **polynomial function** of one variable is of the form,

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_r x^r$$

The largest degree of any term with non-zero coefficient is the degree of a polynomial. If $a_r \neq 0$, then $r$ is called the degree of the polynomial.

(i) A polynomial function of degree 0: constant function

(ii) A polynomial function of degree 1: affine (linear) function\(^1\)

(iii) A polynomial function of degree 2, 3: quadratic function, cubic function

(iv) A polynomial function can be written in the form of $(x - c_1) \times \cdots \times (x - c_r)$.

**Definition 3-12.** A function $f(x)$ is called a **rational function** if it can be expresses as a quotient of polynomials.

\(^1\) An affine function might not be a polynomial of degree 1 as a constant function is affine.
3.2.2 POWER FUNCTION

To define power functions with a real exponent, we use exponential and logarithm functions such as \( x^r = \exp(r \ln x) \).

\[
f(x) = x^r, \quad r \in \mathbb{R}, \quad x \in \mathbb{R}_{++}
\]

The proof below is limited to integer exponents.

**THEOREM 3-4 (Properties of Power Function)**

(i) \( x^r \cdot x^s = x^{r+s} \)  
(ii) \( (x^r)^s = x^{rs} \)  
(iii) \( (x \cdot y)^r = x^r y^r \)  
(iv) \( x^r / x^s = x^{r-s} \)

(v) \( x^0 = 1 \)  
(vi) \( x^{-1} = 1/x \)  
(vii) \( y = x^{1/r}, \ y^r = x \)

**Proof.**

(i) Obviously, \( x^r = x^r \) for all \( r \geq 1 \). Since \( x^{r+1} = x^r \cdot x \), for any \( r, s > 1 \), \( x^{r+s} = x^r \cdot x^s \).

(ii) By (i), \( x^r \times \cdots \times x^r = x^{r+r} = x^{r-s} \).

(iii) \( (x \cdot y)^r = (x \cdot y) \times \cdots \times (x \cdot y) = (x \times \cdots \times x)(y \times \cdots \times y) = x^r y^r \)

(vi) The result follows from \( x^r = x^{r-s} \cdot x^s \).

(v) (v) Since \( x^{r+1} = x^r \cdot x \), \( x^r = \frac{x^{r+1}}{x} \) for all \( r \geq 1 \). By defining the relation for all integers,

\[
x^0 = \frac{x^1}{x} = 1 \quad \text{and} \quad x^{-1} = \frac{x^0}{x} = \frac{1}{x}
\]

(vii) This is the \( n \)th root of a number \( x \), or the power with a rational exponent. \( y = x^{1/r} \) is the solution to \( y^r = x \) because \( y^r = (x^{1/r}) \times \cdots \times (x^{1/r}) = x \).

3.2.3 EXPONENTIAL FUNCTION

\[
f(x) = a^x, \quad a > 0 \text{ and } a \neq 1
\]

**THEOREM 3-5.** For a \( a \geq 0 \) and \( b > 1 \),

Math for Economists, Sogang University
(i) \(a^x \cdot a^y = a^{x+y}\)  (ii) \((a^x)^y = a^{xy}\)  (iii) \(a^x/a^y = a^{x-y}\)  (iv) \(a^0 = 1\)  (v) \(a^{-x} = 1/a^x\)

The base of an exponential function can be changed. Consider \(a^x\) with \(a > 0\). As long as \(a > 0\), it is always possible to write \(a = b^k\) so that \(a^x = b^{kx}\). Since changing the base of the exponential function merely results in a scalar multiplication of power, mathematical analyses are usually based on a standardized base. The term exponential function is almost exclusively used for the natural exponential function, \(e^x\), where \(e\) is Euler’s number, approximately 2.7182. The exponential function \(e^x\) is defined as the solution \(y\) to

\[
x = \int_1^y \frac{dt}{t}, \quad \int_1^e \frac{dt}{t} = 1
\]

**EXAMPLE 3-6.** Compound interest

The interest compounded annually over \(t\) years at an annual interest rate of \(r\) is

\[(1 + r)^t\]

When the interest is compounded \(n\) times per year, it becomes

\[(1 + \frac{r}{n})^{nt}\]

As \(n\), the number of compounding per year, increases without limit, the continuous compound interest converges to

\[
\lim_{n \to \infty} \left(1 + \frac{r}{n}\right)^{nt} = e^{rt}
\]

and the effective annual interest rate approaches \(e^r - 1\). The proof is in EXAMPLE 5-2.

### 3.2.4 TRIGONOMETRIC FUNCTION

\[f(x) = \sin x, \quad f(x) = \cos x\]

Trigonometric functions can be defined as coordinate values of points on the unit circle in the Euclidean space. The input of trigonometric function is the value for the angle formed by the positive half of \(x\) axis and the ray from the origin through the point on the circle. The value
of an angle is measured by the angle itself by degrees or by radians, the length of an arc of a unite circle.

\[ \text{Figure 3-2 Trigonometric Functions: Sine and Cosine Functions} \]

The cosine function is defined as the \( x \)-coordinate value of the point, and the sine function is the \( y \)-coordinate value.

In the calculation, it is recommended to use a radian as an argument.

\[ \text{Figure 3-3 Graph of Sine and Cosine Functions} \]
3.3 COMPOSITE FUNCTIONS

3.3.1 COMPOSITE FUNCTIONS

**Definition 3.13.** Consider two functions, \( g: X \to Y \) and \( f: Y \to Z \). The **composite function** \( f \circ g \) is a mapping from \( X \) to \( Z \) for every \( x \in X \).

The composition of two functions \( f \) and \( g \) is defined \((f \circ g)\) where \( f \) is a function whose domain is the range of \( g \), and is written as \((f \circ g)(x) = f(g(x))\).

\[
\begin{array}{ccc}
X & \longrightarrow & Y \\
\text{Ant} & \longrightarrow & \text{Herbi} \\
\text{Bird} & \longrightarrow & \text{Omni} \\
\text{Cat} & \longrightarrow & \text{Carni} \\
\text{Dog} & \longrightarrow & \text{@} \\
\end{array}
\]

**Theorem 3.6 (Properties of Composite Function)**

(i) Associative rule: \( f \circ (g \circ h)(x) = (f \circ g) \circ h(x) \)

(ii) The composition of one-to-one (onto) functions is always one-to-one (onto).

**Proof.** (i) \( f \circ (g \circ h)(x) = f(g(h(x))) = f(g(h(x))) = (f \circ g)(h(x)) = (f \circ g) \circ h(x) \)

(ii) Let \( h(x) = f \circ g(x) \).

\[
h(x) = h(x') \iff f(g(x)) = f(g(x')) \iff g(x) = g(x') \iff x = x'
\]

This shows that \( h(x) = h(x') \) implies \( x = x' \). That is the definition of an injective function.

To show \( h(x) \) surjective, suppose \( f: Y \to Z \) and \( g: X \to Y \), and \( h = f \circ g: X \to Z \). Since \( f(X) = Z \) and \( g(Y) = Z \),

\[
(f \circ g)(X) \equiv \{ z \in Z | (f \circ g)(x) = z \text{ for some } x \in X \}
\]

\[
= \{ z \in Z | f(g(x)) = z \text{ for some } x \in X \}
\]

\[
= \{ z \in Z | f(y) = z \text{ for some } y \in g(X) \}
\]
\[
\equiv f(g(Y)) = f(Y) = Z
\]
Therefore, \( h(X) = (f \circ g)(X) = Z \) and thus \( h(x) \) is surjective.

**EXAMPLE 3-7.** A function \( h(x) = (x - 2)^2 \) can be expressed as a composite function.

\[
h(x) = (x - 2)^2, \quad x \rightarrow x - 2, \quad y \rightarrow y \times y
\]

\[
h(x) = (f \circ g)(x) \text{ with } f(y) = y^2 \text{ and } g(x) = x - 2
\]

### 3.3.2 Local Inverse

As shown in **THEOREM 3-3**, since an invertible function is injective, \( X \) and \( Y \) have the same cardinality. Recall that the definition of a function as a subset of the Cartesian product \( X \times Y \). The inverse is simply the Cartesian product \( Y \times X \), and their graphs are symmetric about the line \( y = x \). Therefore, a typical way to find an inverse function is to find a function \( g(y) = x \) such that \( y = f(x) \).

A non-invertible function could be invertible for some subsets of domain and codomain,

**DEFINITION 3-14.** \( f^{-1}: J \rightarrow I \) is a **local inverse** for \( f: X \rightarrow Y \) if, for all \( x \in I \) and \( y \in J \),

\[
x = f^{-1}(y) \text{ if and only if } y = f(x)
\]

**EXAMPLE 3-8.** Consider \( f(x) = x^2 \) and \( g(x) = x - 2 \). \( g: \mathbb{R} \rightarrow \mathbb{R} \) is bijective and thus invertible while \( f: \mathbb{R} \rightarrow \mathbb{R} \) is neither injective nor surjective. However, \( f: \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) and \( f: \mathbb{R}_- \rightarrow \mathbb{R}_+ \) are invertible.

The implicit functional form of the inverse of \( g \) is \( x = g^{-1}(x) - 2 \) and thus \( g^{-1}(x) = x + 2 \).

By the same token, the inverse of \( f \) satisfies \( x = (f^{-1}(x))^2 \). The inverse of \( f: \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) is \( f^{-1}(x) = \sqrt{x} \) and the inverse of \( f: \mathbb{R}_- \rightarrow \mathbb{R}_+ \) is \( f^{-1}(x) = -\sqrt{-x} \). \( \square \)

The inverse of a function satisfies \( (f^{-1} \circ f)(x) = x \) and \( (f \circ f^{-1})(y) = y \) for all \( x \in X \) and \( y \in Y \), and it can be written as \( (f^{-1} \circ f)(x) = \text{id}_X(x) = x \). Note that \( f^{-1}(x) \) is different from \( 1/f(x) = [f(x)]^{-1} \), which is a reciprocal.

**THEOREM 3-7.** \( f^{-1} \circ f = \text{id}_X \) and \( f \circ f^{-1} = \text{id}_Y \).
3.3 Real Numbers

**Proof.** For every $x \in X$, there is a unique $y \in Y$ such that $y = f(x)$. Also, for each $y \in Y$, $f^{-1}(y)$ has a unique value $a \in X$ that solves $f(a) = y$. Therefore, for any pair of $(x, y)$ such that $y = f(x)$, we must have $a = x$. That is, $f^{-1}(f(x)) = x$ or $f^{-1} \circ f = \text{id}_Y$. To prove the converse, set $g = f^{-1}$ and $g^{-1} = f$ and apply the same argument. ■

**Example 3.9.** Function $f(x) = (x - 2)^2$ is not invertible but is invertible locally. For instance, $f : [2, \infty) \rightarrow \mathbb{R}_+$ has an inverse, $f^{-1}(x) = \sqrt{x} + 2$. $(f^{-1} \circ f)(x) = \sqrt{(x - 2)^2} + 2 = x - 2 + 2 = x$ and $(f \circ f^{-1})(y) = \left(\sqrt{y + 2} - 2\right)^2 = y$ as expected. □

**Theorem 3.8.** $(f \circ g)^{-1} = (g^{-1} \circ f^{-1})$, where $f^{-1}$ and $g^{-1}$ are the inverse of $f$ and $g$.

**Proof.** By the associative rule, \((g^{-1} \circ f^{-1}) \circ (f \circ g)\)(x) = $(g^{-1} \circ f^{-1} \circ f \circ g)(x) = x$ and \((f \circ g) \circ (g^{-1} \circ f^{-1})\)(x) = $(f \circ g \circ g^{-1} \circ f^{-1})(x) = x$. That is the definition of inverse.

**Example 3.10.** $h(x) = (x - 2)^2$ is a composite function such that $h(x) = f(g(x))$ with $f(x) = x^2$ and $g(x) = x - 2$. Since $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, $f : \mathbb{R}_- \rightarrow \mathbb{R}_+$, and $g : \mathbb{R} \rightarrow \mathbb{R}$ are invertible, we have to restrict the domain of $g$ such that $g : X \rightarrow \mathbb{R}_+$ or $X = [2, \infty)$ or $(-\infty, 2]$ to get the composite function invertible. Let’s take $(-\infty, 2]$. Since $g^{-1}(x) = x + 2$ and $f^{-1}(x) = \sqrt{-x}$, $h^{-1}(x) = (f \circ g)^{-1}(x) = (g^{-1} \circ f^{-1})(x) = \sqrt{-x} + 2$. □

### 3.3.3 Logarithmic Function

A logarithmic function can be defined as an inverse function of an exponential function. Note that an exponential function $a^x$ with base $a > 1$ is strictly increasing and is strictly decreasing for $0 < a < 1$ and its range is $(0, \infty)$. Therefore, it has an inverse $g : (0, \infty) \rightarrow \mathbb{R}$ such that $g(a^x) = x$ and $a^{g(x)} = y$ and is denoted by $g(y) = \log_a y$. By definition,

\[
a^{\log_a x} = x \quad \text{and} \quad \log_a a^x = x
\]

**Theorem 3.9.**

(i) $\log_a (x \cdot y) = \log_a x + \log_a y$

(ii) $\log_a (x/y) = \log_a x - \log_a y$

(iii) $\log_a 1/x = -\log_a x$

(iv) $\log_a x^y = y \log_a x$

Math for Economists, Sogang University
(v) \( \log_a 1 = 0 \)  
(vi) \( \log_b x = \log_a x / \log_a b \)

**Proof.** Let \( w = \log_a x \), \( z = \log_a y \) so that \( x = a^w \) and \( y = a^z \)

(i) \( \log_a (x \cdot y) = \log_a (a^w \cdot a^z) = \log_a (a^{w+z}) = w + z = \log_a x + \log_a y \)

(ii) \( \log_a x^{-1} = \log_a a^{-w} = -w = -\log_a x \)

(iii) \( \log_a x^y = \log_a a^{yw} = yw = y \log_a x \)

(iv) \( \log_a (x \times 1/x) = \log_a x + \log_a 1/x, \ 0 = \log_a x + \log_a 1/x \)

(v) \( \log_a (x \times 1/x) = \log_a x + \log_a 1/x, \ 0 = \log_a x + \log_a 1/x \)

Let \( u = \log_b x \). Then \( x = b^u \). \( \log_a x = \log_a b^u = u \log_a b = \log_b x \times \log_a b \)

\[ \blacksquare \]

The logarithm of a number to the base of \( e \) is called natural logarithm and is written as \( \ln x \).

### 3.4 LIMIT OF A FUNCTION

Consider a function \( f: (a, b) \to \mathbb{R} \), which is defined on an open set except possibly at some \( x_0 \in (a, b) \).

**Definition 3-15. Right-hand limit** \( L^+ \) is the limit of \( f(x) \) from the right at \( x_0 \in (a, b) \) if, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[ |f(x) - L^+| < \varepsilon \quad \text{whenever} \quad 0 < x - x_0 < \delta. \]

And the limit is denoted as

\[ L^+ = \lim_{x \to x_0^+} f(x) = \lim_{x \to x_0^+} f(x) \]

Note that we consider only \( x > x_0 \) and \( x = x_0 \) is not restricted in the definition. Also, it is not necessary that the function is defined at \( x_0 \) nor \( f(x_0) = L^+ \).

**Definition 3-16. Left-hand limit** \( L^- \) is the limit of \( f(x) \) from the left at \( x_0 \in (a, b) \) if, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[ |f(x) - L^-| < \varepsilon \quad \text{whenever} \quad 0 < x_0 - x < \delta. \]

When we apply the definition, it is convenient to define \( \delta \) as a function of \( \varepsilon \).
**Example 3-11.** Let’s find the right- and left-limit of \( f(x) = \frac{2x(x-1)}{x-1} \). The function is not defined at 1, but those limits could exist. For a given \( \varepsilon > 0 \), we have to find the range of \( x \) with which the distance of \( f(x) \) to the right-hand limit \( L^+ \) is less than \( \varepsilon \). Since the definition provides only the conditions of a limit, we need to start from a guess, \( L^+ = 2 \) because \( \frac{2x(x-1)}{x-1} = 2x \) for all \( x \neq 1 \). Since \( |f(x) - L^+| = |2x - 2| = 2|x - 1| < \varepsilon \), if \( 0 < |x - 1| < \varepsilon/2 \), \( |f(x) - L^+| < \varepsilon \). Taking \( \delta = \varepsilon/2 \) shows that \( \lim_{x \to 1^+} f(x) = 2 \). Similarly, the left-hand limit is 2.

**Definition 3-17.** The limit of a function \( f(x) \) is \( \lim_{x \to x_0} f(x) = L \), if, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that \( |f(x) - L| < \varepsilon \) whenever \( 0 < |x - x_0| < \delta \).

If \( \lim_{x \to x_0} f(x) = L \), then, for any "horizontal band" (measured by \( \varepsilon \)), there is a "vertical band" (measured by \( \delta \)) such that if \( x \) is in the vertical band then \( f(x) \) is in the horizontal band.

From the definitions, clearly the limit of a function, \( \lim_{x \to x_0} f(x) = L \), exists if and only if \( \lim_{x \uparrow x_0} f(x) = \lim_{x \downarrow x_0} f(x) = L \).

**Theorem 3-10 (Uniqueness)** If \( \lim_{x \to x_0} f(x) = L \) and \( \lim_{x \to x_0} f(x) = M \), then \( L = M \).

**Proof.** Suppose \( L \neq M \) to reach a contradiction. Let \( \varepsilon = |L - M|/2 \). Since \( L \) and \( M \) are the limits, there exists \( \delta > 0 \) such that \( |f(x) - L| < \varepsilon \) and \( |f(x) - M| < \varepsilon \) for \( 0 < |x - x_0| < \delta \). But this contradicts to

\[
|f(x) - L| + |f(x) - M| \geq |L - M| = 2\varepsilon.
\]


Consider $|x - x_0| < \delta$. Then for any $\delta > 0$, there exists $N$ such that $n > N$ implies $\frac{1}{n} < \delta$. A similar argument holds for $|f(x) - L| < \varepsilon$. Therefore, if the same condition is met for every convergent sequence, it should be the limit of the function.

**THEOREM 3-11.** $\lim_{x \to x_0} f(x) = L$ if and only if, for all sequences $\{x_n\}$ such that $\lim_{n \to \infty} x_n = x_0$, $\lim_{n \to \infty} f(x_n) = L$.

**Proof.** Suppose that $\lim_{x \to x_0} f(x) = L$ and consider sequences, $\{x_n\}$ with $\lim_{n \to \infty} x_n = x_0$. By the definition, for any $\varepsilon > 0$, there is $\delta > 0$ such that $|f(x) - L| < \varepsilon$ whenever $0 < |x - x_0| < \delta$. And for any $\delta > 0$, there exists $N$ such that $0 < |x_n - x_0| < \delta$ for every $n > N$. Pick $N$ such that $n > N$ implies $0 < |x_n - x_0| < \delta$ so that $|f(x_n) - L| < \varepsilon$ whenever $n > N$.

Conversely, suppose $\lim_{x \to x_0} f(x) \neq L$. Then, for some $\varepsilon > 0$ and for every $\delta > 0$, there is $x$ such that $0 < |x - x_0| < \delta$ but $|f(x) - L| \geq \varepsilon$. However, for any $x_n$ such that $0 < |x_n - x_0| < 1/n$ ($\lim_{n \to \infty} x_n = x_0$), $|f(x_n) - L|$ vanishes as $n$ increases. ■

**THEOREM 3-12 (Limits of Elementary Functions)** If $\lim_{x \to x_0} f(x) = L$ and $\lim_{x \to x_0} g(x) = M$, then

(i) $\lim_{x \to x_0} \lambda \cdot f(x) = \lambda \cdot L$

(ii) $\lim_{x \to x_0} (f(x) + g(x)) = L + M$

(iii) $\lim_{x \to x_0} (f(x) \cdot g(x)) = L \cdot M$

(iv) $\lim_{x \to x_0} \left(\frac{f(x)}{g(x)}\right) = \frac{L}{M}$ when $M \neq 0$

**Proof.** (i) If $\lambda = 0$, the proof is trivial. Suppose $\lambda \neq 0$. Given $\varepsilon > 0$, by definition, there exist $\delta$ such that

$$0 < |x - x_0| < \delta \implies |f(x) - L| < \frac{\varepsilon}{|\lambda|}.$$

Therefore, $0 < |x - x_0| < \delta$ implies $|\lambda| \cdot |f(x) - L| = |\lambda \cdot f(x) - \lambda \cdot L| < \varepsilon$.

(iii) Given $\varepsilon > 0$, by definition, there exist $\delta_1$ and $\delta_2$ such that
\[
0 < |x - x_0| < \delta_1 \Rightarrow |f(x) - L| < \frac{\varepsilon}{2} \quad \text{and} \quad 0 < |x - x_0| < \delta_2 \Rightarrow |f(x) - M| < \frac{\varepsilon}{2}
\]

Letting \( \delta = \min(\delta_1, \delta_2) \), \( 0 < |x - x_0| < \delta \) implies, with triangle inequality,

\[
|f(x) + g(x) - L - M| \leq |f(x) - L| + |g(x) - M| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\]

(iii) Given \( \varepsilon > 0 \), by definition, there exist \( \delta_1 \) and \( \delta_2 \) such that

\[
0 < |x - x_0| < \delta_1 \Rightarrow |f(x) - L| < \frac{\varepsilon}{4A}
\]

and

\[
0 < |x - x_0| < \delta_2 \Rightarrow |g(x) - M| < \min \left( A, \frac{\varepsilon}{4A} \right)
\]

where \( A = \max(|L|, |M|, 1) \). Letting \( \delta = \min(\delta_2, \delta_2) \), \( 0 < |x - x_0| < \delta \) implies

\[
|f(x) \cdot g(x) - L \cdot M| = |f(x) \cdot g(x) - L \cdot g(x) + L \cdot g(x) - L \cdot M| \\
\leq | f(x) \cdot g(x) - L \cdot g(x) | + | L \cdot g(x) - L \cdot M | \\
= |g(x)| \cdot | f(x) - L | + | L | \cdot | g(x) - M |
\]

Since \( |g(x)| = |g(x) - M + M| \leq |g(x) - M| + |M| < A + A, \)

\[
|f(x) \cdot g(x) - L \cdot M| < 2A \times \frac{\varepsilon}{4A} + A \times \frac{\varepsilon}{4A} = \frac{3\varepsilon}{4}
\]

(iv) I will prove \( \lim_{x->x_0} \frac{1}{g(x)} = \frac{1}{M} \) when \( M \neq 0 \). Since \( \lim_{x->x_0} g(x) = M \), for a given \( \varepsilon > 0 \), there exists \( \delta_1 \) such that

\[
0 < |x - x_0| < \delta_1 \Rightarrow |g(x) - M| < \frac{|M|}{2},
\]

\[
|M| = |M - g(x) + g(x)| \leq |g(x) - M| + |g(x)| < \frac{|M|}{2} + |g(x)|
\]

Rearranging terms gives

\[
|M| < \frac{|M|}{2} + |g(x)|, \quad \frac{|M|}{2} < |g(x)|, \quad \frac{1}{|g(x)|} < \frac{2}{|M|}
\]
\[
\left| \frac{1}{g(x)} - \frac{1}{M} \right| = \left| \frac{g(x) - M}{M \cdot g(x)} \right| = \frac{|g(x) - M|}{|M| \cdot |g(x)|} < \frac{2}{|M|^2} |g(x) - M|
\]

Since there is \( \delta_2 > 0 \) such that
\[
0 < |x - x_0| < \delta_2 \Rightarrow |g(x) - M| < \frac{|M|^2 \varepsilon}{2},
\]
Taking \( \delta = \min(\delta_1, \delta_2) \) completes the proof. \( \blacklozenge \)

**EXAMPLE 3.12.** Find \( \lim_{x \to -\infty} x/2^x \) and \( \lim_{n \to \infty} \left[ 2^{-n} \times \frac{n^2 + 1}{n + 1} \right] \) if exists.

\[
\lim_{x \to -\infty} x/2^x = \lim_{y \to \infty} -y/2^y = \lim_{y \to \infty} -y \cdot 2^y = -\infty
\]

\[
\lim_{n \to \infty} \left[ 2^{-n} \times \frac{n^2 + 1}{n + 1} \right] = \lim_{n \to \infty} \left[ \frac{n^2 + 1}{2^n \cdot (n + 1)} \right] \leq \lim_{n \to \infty} \frac{n}{2^n} = 0
\]

### 3.5 Continuity

Continuity is indispensable in most economic analyses and is assumed almost everywhere.

#### 3.5.1 CONTINUOUS FUNCTIONS

In general, for a given convergent sequence \( x_n \to x_0 \), \( f(x_0) \) is different from \( \lim_{n \to \infty} f(x_n) \). The limit condition does not put any restriction on the behavior of \( f(x) \) at \( x_0 \). If the same requirement holds in the limit, we have continuity.

**DEFINITION 3.18.** A function \( f \) is **continuous** at \( x_0 \),

\[
f(x_0) = \lim_{x \to x_0} f(x).
\]

That is, \( \lim_{x \to x_0} f(x) \) exists and should be equal to \( f(x_0) \). The definition can be rephrased such as:

**THEOREM 3.13.** A function \( f \) is continuous at \( x_0 \) if and only if, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
|x - x_0| < \delta \quad \Rightarrow \quad |f(x) - f(x_0)| < \varepsilon.
\]
Notice that the condition is almost the same as the definition of function limit except that the condition $0 < |x - x_0| < \delta$ is replaced with $|x - x_0| < \delta$. That is, a function should also satisfy the property at the limit as well.

**Theorem 3-14.** $f(x)$ is continuous at $x_0$ if and only if $\lim_{n \to \infty} f(x_n) = f(x_0)$ for all sequences $\{x_n\}$ with $\lim_{n \to \infty} x_n = x_0$. That is,

$$f \left( \lim_{n \to \infty} x_n \right) = \lim_{n \to \infty} f(x_n)$$

Note that the latter condition is equivalent to those in Theorem 3-11 with $L = f(x_0)$.

**Proof.** We prove the contrapositive: for a sequence whose limit is $x_0$, if $\lim_{n \to \infty} f(x_n) \neq f(x_0)$, the function is not continuous.

(⇒) The condition on the sequence $\{x_n\}$ implies for any $\delta > 0$, there exists $N$ such that for all $n > N$,

$$|x_n - x_0| < \delta$$

But $|f(x_n) - f(x_0)| \geq \varepsilon$, which violates the continuity of $f$.

(⇐) Suppose $f$ is not continuous at $x_0$. Then for some $\varepsilon > 0$ such that for any $\delta > 0$ there exists $x'$ such that

$$|x' - x_0| < \delta \quad \Rightarrow \quad |f(x') - f(x_0)| > \varepsilon.$$

Choosing $x_n = x'$ for $\delta = 1/n$, $x_n$ converges to $x_0$ but $\lim_{n \to \infty} f(x_n) \neq f(x_0)$. ■

**Example 3-13.** Note that continuity is defined on a point. The following function is continuous only at 0.

$$f(x) = \begin{cases} x & x \in \mathbb{Q} \\ 0 & x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

For any $x \neq 0$, there is no such $\delta > 0$ that satisfies the definition for $\varepsilon = x/2$. ■

**Theorem 3-15.** The scalar multiplication, sum, product, and quotient of functions that continuous at $x_0$ are also continuous at the point.
**Proof.** By **Theorem 3-12**, \( \lim_{x \to x_0} \lambda \cdot f(x) = \lambda \cdot \lim_{x \to x_0} f(x) \). Since \( f(x) \) is continuous,

\[
 f(x_0) = \lim_{x \to x_0} f(x) \quad \text{or} \quad \lambda \cdot f(x_0) = \lambda \cdot \lim_{x \to x_0} f(x)
\]

Therefore, \( \lim_{x \to x_0} \lambda \cdot f(x) = \lambda \cdot f(x_0) \), which is the definition of continuity. We can apply the same argument to the sum, product, and quotient of functions. ■

**Theorem 3-16.** If \( g: X \to Y \) and \( f: Y \to Z \) are continuous, so is \( f \circ g \).

**Proof.** Given \( \varepsilon > 0 \), since \( f(y) \) is continuous at \( g(x_0) \), there exists \( \delta_2 > 0 \) such that

\[
|y - g(x_0)| < \delta_2 \quad \Rightarrow \quad |f(y) - f(g(x_0))| < \varepsilon.
\]

Since \( g(x) \) is continuous at \( x_0 \), there exists \( \delta_1 > 0 \) such that

\[
|x - x_0| < \delta_1 \quad \Rightarrow \quad |g(x) - g(x_0)| < \delta_2.
\]

Therefore,

\[
|x - x_0| < \delta_1 \quad \Rightarrow \quad |f(y) - f(g(x_0))| < \varepsilon. \quad \blacksquare
\]

**Definition 3-19.** A function \( f(x) \) is **continuous** on \( X \) if \( f \) is continuous at every point in \( X \).

### 3.5.2 Properties

**Theorem 3-17.** The following properties are equivalent.

(i) \( f:X \to \mathbb{R} \) is continuous on \( X \).

(ii) For any open set \( C \subset f(X) \), \( f^{-1}(C) \) is open in \( X \).

(iii) For any closed set \( D \subset f(X) \), \( f^{-1}(D) \) is closed in \( X \).

**Proof.** (i)⇒(ii) Let \( C \subset f(X) \) be open. Then for any \( f(x) \in C \) there exists \( r > 0 \) such that \( B_r(f(x)) \subset C \). By continuity, for any \( 0 < \varepsilon \leq r \), there exists \( \delta > 0 \) such that \( f(B_\delta(x)) \subset B_{\lambda r}(f(x)) \subset C \). Therefore, \( B_\delta(x) \subset f^{-1}(C) \) for every \( x \in f^{-1}(C) \subset X \) with a properly chosen \( \delta \) depending on \( x \) as well as \( \varepsilon \). This is the definition of an open set.

Math for Economists, Sogang University
Here are some useful properties of a continuous function. Let \( D \subset f(X) \) be a closed set. Consider a sequence \( x_n \to x_0 \) such that \( x_n \in f^{-1}(D) \). Since \( f \) is continuous, \( \lim_{n \to \infty} f(x_n) = f(x_0) \) and \( f(x_0) \in D \). Therefore, \( f^{-1}(D) \) is closed.

(ii) \( \Rightarrow \) (i) By the definition of a continuous function, it is sufficient to show that for any \( B_\epsilon(f(x)) \) with \( \epsilon > 0 \), there exists a \( \delta > 0 \) such that \( f(B_\delta(x)) \subset B_\epsilon(f(x)) \). Since \( B_\epsilon(f(x)) \) is open, \( f^{-1}(B_\epsilon(f(x))) \) is open. Therefore, \( x \) is an interior point for \( f^{-1}(B_\epsilon(f(x))) \), or \( B_\delta(x) \subset f^{-1}(B_\epsilon(f(x))) \). This implies that \( f(B_\delta(x)) \subset B_\epsilon(f(x)) \).

(ii) \( \iff \) (iii) Let \( D \subset f(X) \) be a closed set in \( f(X) \). Then \( f(X) \setminus D \) is open. Then \( f^{-1}(f(X) \setminus D) \) is open by (i) \( \Rightarrow \) (ii). Since \( f^{-1}(f(X) \setminus D) = X \setminus f^{-1}(D) \), \( f^{-1}(D) \) is closed. The similar argument can be applied to the converse with a open set \( C \subset f(X) \).

The converse of (ii) and (iii) are not true.

**Example 3.14** (An Alternative Proof of Continuity of Composite function) Let \( C \subset \mathbb{R} \) is open. Then \( h^{-1}(C) = g^{-1}(f^{-1}(C)) \) and \( f^{-1}(C) = D \cap Y \) for some open set \( D \subset \mathbb{R} \). Therefore, \( h^{-1}(C) = g^{-1}(D \cap Y) = g^{-1}(D) \) is open in \( D \), and \( h(x) \) is continuous.

**Theorem 3.18.** A real-valued function \( f: X \to \mathbb{R} \) is **continuous** if and only if, for every \( c \in f(X) \), all its upper contour set of \( \{ x \in X | f(x) \geq c \} \) and lower contour set of \( \{ x \in X | f(x) \leq c \} \) are closed in \( X \).

(\( \Rightarrow \)) Since \( f \) is continuous, the inverse image of a closed set, \( [f(x), \infty) \), is also closed. Hence the upper contour set is closed. The same argument applies to a lower contour set.

(\( \Leftarrow \)) Since \( \{ x \in X | f(x) > a \} \) and \( \{ x \in X | f(x) < b \} \) are open, their intersection is open.

\[
f^{-1}((a, b)) = \{ x \in X | f(x) > a \} \cap \{ x \in X | f(x) < b \}
\]

Any open set \( C \in X \) can be represented by the union of open sets (at least the union of open balls around each element of \( C \)), \( C = \bigcup_{i \in I} (a_i, b_i) \). By **Theorem 3.1.**(v), since

\[
f^{-1}(C) = f^{-1}\left( \bigcup_{i \in I} (a_i, b_i) \right) = \bigcup_{i \in I} f^{-1}((a_i, b_i))
\]

and the union of open sets is open, \( f^{-1}(C) \) is open. The result follows from **Theorem 3.17.**

Here are some useful properties of a continuous function.
There is no sudden jump in the graph of a continuous function. If \( f(x) < f(y) \), then a sufficiently small change in \( x \) or \( y \) does not change the inequality.

**Theorem 3.19.** Suppose that a function \( f(x) \) is continuous on \([a, b]\). For any \( a \leq x, y \leq b \), if \( f(x) < f(y) \), then there exists \( \epsilon > 0 \) such that \( f(x) < f(z) \) for all \( z \in B_\epsilon(y) \cap [a, b] \).

**Proof.** Suppose not. Then there exists a sequence \( y_n \) converging to \( y \) such that \( f(y_n) < f(x) \) for all \( n \). That contradicts to that \( f(x) \) is continuous. \( \blacksquare \)

We also have the sandwich theorem for the function values.

**Theorem 3.20.** For any sequences \( x_n \) and \( y_n \) such that \( \lim_{n \to \infty} x_n = x_0 \) and \( \lim_{n \to \infty} y_n = y_0 \), if \( f(x) \) is continuous and \( f(x_n) \geq f(y_n) \) for all \( n \), then \( f(x_0) \geq f(y_0) \).

**Proof.** Since \( f \) is continuous, for any \( \epsilon > 0 \), there exists a \( \delta > 0 \) such that \( f(x_n) \in B_\epsilon(f(x_0)) \) and \( f(y_n) \in B_\epsilon(f(y_0)) \) whenever \( x_n \in B_\delta(x_0) \) and \( y_n \in B_\delta(y_0) \). Moreover, since \( f(x_0) \in B_\epsilon(f(x_0)) \) and \( f(y_0) \in B_\epsilon(f(y_0)) \), we have \( f(x_0) \geq f(y_0) \). \( \blacksquare \)

A monotone function is continuous almost everywhere.

**Theorem 3.21.** If a function is monotonic on \([a, b]\), then one-sided limit always exist at for any \( c \in (a, b) \).

**Proof.** Let \( c \in (a, b) \) and \( L = \{x \in [a, b] | x \leq c\} \). \( L \) is not empty and bounded. Moreover, \( f(L) \) is bounded (otherwise \( f(b) \) is not defined.) Take a sequence \( \{l_n\} \in L \) such that \( l_n \to c^- \) and \( l_n < l_{n+1} \). Then, for any such \( l_n \), \( f(l_n) \) is bounded and monotone, it has a limit. \( \blacksquare \)

**Theorem 3.22.** If a function is monotonic on \([a, b]\), then

\[
D = \{d \in [a, b] | f \text{ is discontinuous at } d\}
\]

is finite or countable.

**Proof.** If \( d \in D \), then \( f(d^-) < f(d^+) \). Since \( \mathbb{Q} \) is dense in \( \mathbb{R} \), for every \( d \in D \), we can choose \( g(d) \) such that

\[
f(d^-) < g(d) < f(d^+)
\]

This defines a function \( g: D \to \mathbb{Q} \). Since \( c > d \) implies \( f(c^-) \geq f(d^+) \),
\[ c > d \Rightarrow g(c) > f(x^-) \geq f(d^+) > g(d) \]

\( g(x) \) is increasing, and thus 1-1. Therefore, \( g(x) \) is a bijection from \( D \) to a subset of \( \mathbb{Q} \). □

With continuity, some properties of the domain are preserved for its image. In particular, the compactness of the image of a function is what we usually assume to guarantee the existence of extrema.

**Theorem 3-23.** If \( f(x) \) is continuous on \( X \), the image of a compact subset of \( X \) is compact.

**Proof.** We show that every sequence in \( f(X) \) has a convergent subsequence. Let \( \{y_n\} \) be a sequence in \( f(X) \). We will show that \( \{y_n\} \) has a convergent subsequence. For each \( n \in \mathbb{N} \), let \( x_n \in X \) be such that \( f(x_n) = y_n \). Since \( X \) is compact, \( \{x_n\} \) has a subsequence \( \{x_{n_k}\} \) that converges to some \( x \in X \). Since \( f(x) \) is continuous, \( \{f(x_{n_k})\} \) converges to \( f(x) \in Y \). Since \( x \in X \), we have \( f(x) \in f(X) \). □

An immediate consequence of the result is that if a function defined on an interval is continuous, it attains its maximum and minimum.

**Theorem 3-24 (Extreme Value Theorem)** A continuous real-valued function on a compact subset \( X \) attains a maximum and a minimum on \( X \).

Also, the properties of closedness and boundedness are preserved independently.

**Theorem 3-25.** If the domain of a continuous function is bounded, then the image is bounded.

**Proof.** Suppose that it is not bounded. Then there must exist \( \{x_n\} \) with \( x_n \in [a, b] \) for all \( n \) and \( |f(x_n)| \geq n \) for all \( n \). Since \( \{x_n\} \) is bounded, it has a subsequence \( \{x_{n_i}\} \) converging to a limit \( x_0 \in [a, b] \). By continuity, \( \lim_{n \to \infty} f(x_{n_i}) = f(x_0) \) which contradicts the hypothesis that

\[ \lim_{i \to \infty} f(x_{n_i}) = \infty. \]

**Theorem 3-26.** If the domain of a continuous function is closed, then the image is also closed.

**Proof.** Consider a convergent sequence \( y_n \in f(X) \). Take a sequence \( x_n \in X \) such that \( y_n = f(x_n) \). Since \( y_n \) converges, \( \lim_{n \to \infty} f(x_n) \) exists. Since the domain is closed and \( f(x) \) is continuous, \( \lim_{n \to \infty} x_n = x_0 \in X \) and \( \lim_{n \to \infty} f(x_{n}) = f(x_0) \in f(X) \). □
The intermediate value theorem is used to check the existence of a solution to an equation system. Once we know \( f(x_1) > 0 \) and \( f(x_2) < 0 \) and the function is continuous, the theorem guarantees that there exists a point such that \( f(x) = 0 \). \text{THEOREM 3-32} is the identical result.

\textbf{THEOREM 3-27 (Intermediate Value Theorem)} If \( f(x) \) is continuous on \([a, b] \) and \( z \in [f(a), f(b)] \) then there exists \( c \in (a, b) \) such that \( f(c) = z \).

\textbf{Proof.} We assume \( f(a) \leq z \leq f(b) \). If \( f(a) \geq f(b) \), you can flip them accordingly.

Let \( L = \{ t \in [a, b] | f(t) \leq z \} \). Since \( f(a) \leq z \), \( a \in L \) and \( L \) is not empty. By the supremum property, \( \sup L \in \mathbb{R} \) exists. Let \( c = \sup L \). Since \( a \in L \) and \( a \leq c \) and \( L \subset [a, b] \). We claim that \( f(c) = z \).

Let

\[ t_n = \min \left( c + \frac{1}{n}, b \right) \geq c \]

If \( t_n > c \), \( t_n \notin L \). If \( t_n = c \), \( t_n = b \) and \( f(t_n) > z \) and \( t_n \notin L \). In either case \( f(t_n) \geq z \).

Since \( f(x) \) is continuous at \( c \), \( f(c) = \lim_{n \to \infty} f(t_n) \geq z \).

Since \( c = \sup L \), there exists \( s_n \in L \) such that

\[ c \geq s_n \geq c - \frac{1}{n} \]

Since \( s_n \in L \), \( f(s_n) < z \). Again, by continuity, \( f(c) = \lim_{n \to \infty} f(s_n) \leq z \).

By putting them together, we have \( z \leq f(c) \leq z \) or \( f(x) = z \). Since \( f(a) < d \) and \( f(b) > d \), and \( a \neq c \) and \( c \neq b \), \( c \in (a, b) \).

\textbf{EXAMPLE 3-15.} There exists \( x \in \mathbb{R} \) such that \( x^2 = 2 \).

Let \( f(x) = x^2 \) for \( x \in [0, 2] \). \( f(x) \) is continuous and \( 0 = f(0) < 2 < f(2) \). By the intermediate value theorem, there exists \( c \in [0,1] \) such that \( f(c) = 2 \) or \( c^2 = 2 \).

An interesting application of the intermediate value theorem is the existence of a fixed point. If \( f: [a, b] \to [a, b] \) is a continuous function, then \( f \) has a fixed point \( f(x) = x \) for \( x \in [a, b] \).
3.5 Real Numbers

**THEOREM 3-28 (Brouwer’s fixed point theorem)** Every continuous function from a compact set to itself has a fixed point.

**Proof.** Consider \( g(x) = f(x) - x \), which is continuous on \([a, b]\). Since \( g(a) = f(a) - a \geq 0 \) and \( g(b) = f(b) - b \leq 0 \), by the intermediate value theorem, there exists \( x \in [a, b] \) with \( g(x) = 0 \) or \( f(x) = x \). \( \square \)

**DEFINITION 3-20.** A function \( f \colon (X, d) \to (Y, \rho) \) is **uniformly continuous** on \( X \) if, for any \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that \( d(x, x_0) < \delta \) implies \( \rho(f(x), f(x_0)) < \varepsilon \) for all \( x_0 \in X \).

The uniform continuity requires that the definition of the continuity at a point should hold for every point in the domain with common \( \varepsilon \) and \( \delta \).

**EXAMPLE 3-16.** Consider \( f(x) = 1/x \) for \( x \in (0,1] \). This is not uniformly continuous.

For some \( \varepsilon > 0 \) and \( x_0 \in (0,1] \), if

\[
x = \frac{x_0}{1 + \varepsilon x_0} \Rightarrow x < x_0 \Rightarrow \frac{1}{x} - \frac{1}{x_0} > 0
\]

\[
|f(x) - f(x_0)| = \frac{1}{x} - \frac{1}{x_0} = \frac{1 + \varepsilon x_0}{x_0} - \frac{1}{x_0} = \varepsilon
\]

Then there must exist \( \delta > 0 \) such that

\[
|x - x_0| = \left| \frac{x_0}{1 + \varepsilon x_0} - x_0 \right| \geq \delta > 0
\]

But, since

\[
\delta \leq x_0 - \frac{x_0}{1 + \varepsilon x_0} = \frac{\varepsilon x_0^2}{1 + \varepsilon x_0} < \varepsilon x_0^2
\]

\( \delta \to 0 \) as \( x_0 \to 0 \), and there is no such \( \delta > 0 \) that works for all \( x \in (0,1] \). \( \square \)

**DEFINITION 3-21.** Let \((X, d)\) and \((Y, \rho)\) be metri spaces. A function \( f \colon (X, d) \to (Y, \rho) \) is called a **homeomorphism** if it is bijective, continuous, and its inverse is continuous.

Now suppose that \( f(x) \) is a homeomorphism and \( U \subset X \). Let \( g \colon Y \to X \) be the inverse of \( f(x) \). Then by **THEOREM 3-17**, \( U \) is open in \( X \) \( \iff g^{-1}(U) \) is open in \( (f(X), \rho) \) \( \iff f(U) \) is open in \( (f(X), \rho) \)
3 Univariate Function

This states that \((X, d)\) and \((f(X), \rho)\) are identical in terms of properties that can be characterized solely in terms of open sets. Such properties are called “topological properties.”

3.5.3 CONNECTED SET

No new results are presented in this section. Some of the previous results are reproduced using different notions.

**Definition 3-22.** A real interval \(X \subset \mathbb{R}\) is a set of real numbers such that for any \(x, y \in X\) with \(x < y\), if \(x \leq z \leq y\), then \(z \in X\).

We use the term interval quite often, and we are usually interested in its nature of “connectedness.” Since an interval is defined for real numbers, we need a notion to extend the idea to a multi-dimensional space.

**Definition 3-23.** Two sets \(A\) and \(B\) in a metric space are separated if

\[
\overline{A} \cap B = A \cap \overline{B} = \emptyset
\]

A set in a metric space is connected if it cannot be written as the union of two nonempty separated sets.

**Example 3-17.** \((0,1)\) and \((1,2)\) are separated, and \((0,1)\) and \([1,2]\) are not separated.

\[
\overline{(0,1)} \cap (1,2) = (0,1) \cap (1,2) = \emptyset \quad \text{and} \quad \overline{(0,1)} \cap [1,2] \neq \emptyset
\]

In \(\mathbb{R}\), a connected set is an interval. \(\square\)

**Theorem 3-29.** A set \(X \subset \mathbb{R}\) is connected if and only if it is an interval.

(\(\Rightarrow\)) We prove the contrapositive. If \(X\) is not an interval, then it is not connected. Since it is not an interval, for some \(x, y \in X\), there exists a \(x \leq z \leq y\) but \(z \notin X\). Let

\[
A = X \cap (-\infty, z) \quad \text{and} \quad B = X \cap (z, \infty)
\]

Then

\[
\overline{A} \cap B \subset (-\infty, z) \cap (z, \infty) = (-\infty, z] \cap (z, \infty) = \emptyset
\]
\[ A \cap \bar{B} = \cap (\neg \infty, z) \cap (z, \infty) = (\neg \infty, z) \cap [z, \infty) = \emptyset \]  

Since \( A \cup B = (X \cap (\neg \infty, z)) \cup (X \cap (z, \infty)) = X \setminus z = X \) and \( x \in A \) and \( y \in B \), \( X \) is not connected.

(\(=\)) Suppose not. Then there exists non-empty open subsets \( A, B \in X \) such that \( X = A \cup B \) and \( A \cap B = \emptyset \). Let \( a \in A \) and \( b \in B \) with \( a < b \). Let

\[ \alpha = \sup\{ x \in \mathbb{R} | (a, x) \cap X \subseteq A \} \]

Note that \( a \in \bar{A} \) and \( a \leq b \). Otherwise, \( \alpha \) is not a supremum of the set. Moreover, \( X \) is not connected and \( \alpha \in X \), \( b \in X \setminus A \) and thus \( \alpha < b \). Therefore, since \( A \) is open, there exists \( \varepsilon > 0 \) such that \( (a, \alpha + \varepsilon) \cap X \in A \), which is a contradiction.

**Theorem 3-30.** If a real interval \( X \) is a union of two nonempty subsets such that \( X = A \cup B \), one of the sets contains an element at zero distance from the other.

**Proof.** Every element of \( X \) belongs either to \( A \) or to \( B \). Assume that \( A \) and \( B \) have no elements in common. Otherwise, the result is trivial. Let \( a \in A \) and \( b \in B \) with \( a < b \), and \( B = \{ x \in B | x > a \} \). Clearly, \( B \) is non-empty and bounded below.

If \( \inf B \notin B \), then \( \inf B \in A \). However, \( \inf B \) has zero distance to \( B \). Therefore, there is a point in \( A \) which is zero distance from \( B \).

If \( \inf B \in B \), then \( \inf B > a \). And the open interval \( (a, \inf B) \) is a non-empty subset of \( A \).

Hence, \( \inf B \) is a point in \( B \) which is zero distance from \( A \).

**Theorem 3-31.** Let \( X \) be a non-empty subset of \( \mathbb{R} \). For a point \( x_0 \) whose distance to \( X \) is zero, there exists a sequence \( \{ x_n \} \) in \( X \) such that \( \lim_{n \to \infty} x_n = x_0 \).

**Proof.** We need the result that for any \( n > N \), there exists \( x_n \) such that \( |x_0 - x_n| < \frac{1}{n} \).

Suppose not so that for any \( n > N \), there is no \( x \) such that \( |x - x_0| < \frac{1}{n} \). Then \( \frac{1}{n} \) is a lower bound of the set \( \{ |x - x_0| \mid x \in X \} \). This contradicts to \( x_0 \) has zero distance to \( X \). Since \( \lim_{n \to \infty} 1/n = 0 \), by Sandwich theorem, \( |x - x_0| < 1/n \) implies \( \lim_{n \to \infty} x_n = x_0 \).
**Theorem 3.32.** If $f$ is continuous on an interval, then its image is also an interval.

**Proof.** For a $y \in f(X)$, let $L = \{x | f(x) \leq y\}$ and $U = \{x | f(x) \geq y\}$ with $l \in L$ and $u \in U$. Then $L$ and $U$ are nonempty and $f(X) = L \cup U$. By Theorem 3.30, a point in one subset is at zero distance from the other. Without loss of generality, $l \in L$ is at zero distance from $U$.

By Theorem 3.31, there is a sequence $(u_n)$ in $U$ such that $\lim_{n \to \infty} u_n = l$. Since $f(x)$ is continuous on $X$, $\lim_{n \to \infty} f(u_n) = f(l)$. By Theorem 3.20, $\lim_{n \to \infty} f(u_n) \geq y$ and thus $f(l) \geq y$.

By construction, $f(l) \leq y$ and therefore $f(l) = y$. 

∎
4 DERIVATIVES

The slope of a linear line is defined as the ratio of the function value change between two points, and it is the same at all points. The slope contains two pieces of information, i.e., the direction to which function value increases and steepness of the line, which play a key role in the optimization problem. If we apply the same definition to a non-linear function, we have a near approximation of a function on an interval such as, for some \( \delta \),

\[
\frac{f(x + \delta/2) - f(x - \delta/2)}{\delta} \approx \frac{f(x + \delta) - f(x)}{\delta} \approx \frac{f(x) - f(x - \delta)}{\delta}
\]

However, it depends not only on \( x \) but also on the choice of \( \delta \). If only the local behavior is of interest, then it makes sense to consider the slope for a small \( \delta \) or even in the limit of \( \delta \to 0 \).

\[
\lim_{\delta \to 0} \frac{f(x + \delta/2) - f(x - \delta/2)}{\delta} = \lim_{\delta \to 0} \left[ \frac{f(x + \delta) - f(x)}{2\delta} + \frac{f(x) - f(x - \delta)}{2\delta} \right]
\]

which is the slope of the tangent line at each point on a curve.

4.1 DERIVATIVES

**Definition 4-1.** \( f: \mathbb{R} \to \mathbb{R} \) is **differentiable** at \( x_0 \) if the following limit exists.

\[
\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}
\]

The value of limit is called the (first) derivative of \( f \) at \( x_0 \). Otherwise, the function is not differentiable at \( x_0 \).

<table>
<thead>
<tr>
<th>Leibniz</th>
<th>Lagrange</th>
<th>Newton</th>
<th>Euler</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{df(x_0)}{dx} ), ( \frac{d}{dx} f(x_0) ), ( \frac{dy}{dx} ) ( x = x_0 )</td>
<td>( \frac{dy}{dx} (x_0) ) ( f'(x_0) )</td>
<td>( \dot{y} )</td>
<td>( Df(x_0), D_x y, D_x f(x_0) )</td>
</tr>
</tbody>
</table>

**Example 4-1.** Consider \( f(x) = ax \).
\[
\frac{d}{dx} ax_0 = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0} \frac{ax - ax_0}{x - x_0} = \lim_{x \to x_0} a(x - x_0) = a
\]

The derivative is constant for all \( x_0 \in \mathbb{R} \).

**Theorem 4-1.** If \( f(x) \) is differentiable at \( x_0 \), \( f(x) \) is continuous at \( x_0 \).

**Proof.**

\[
\lim_{x \to x_0} f(x) = f(x_0) = f'(x_0) \times \lim_{x \to x_0} (x - x_0) = 0
\]

**Alternative Proof.** Suppose that \( f \) is not continuous at \( x_0 \). Then there exists \( \{x_n\} \) converging to \( x_0 \), and \( x_n \neq x_0 \) for each \( n \). Since \( f \) is differentiable at \( x_0 \), for \( \varepsilon > 0 \) given, there exists \( N \) and \( L \) such that for \( n > N \),

\[
\left| \frac{f(x_n) - f(x_0)}{x_n - x_0} - L \right| < \varepsilon
\]

However, this contradicts to \( \lim_{n \to \infty} [f(x_n) - f(x_0)] \neq 0 \) and \( \left| \frac{f(x_n) - f(x_0)}{x_n - x_0} \right| \) cannot approach a finite limit.

The inverse is not true. \( f(x) = |x| \) is not differentiable at \( x = 0 \).

The derivative of \( f(x) \) at \( x_0 \) is a number as it is the slope of the tangent line to \((x_0, f(x_0))\) while the derivative of \( f(x) \) refers to the expression of the derivative at an arbitrary point. If \( f(x) \) is differentiable over an interval, the collection of derivatives at each point can be treated as a function, which is called derivative function of \( f(x) \) or derivative of \( f(x) \).

**Example 4-2.** \( f(x) = x^2 \)

\[
\begin{align*}
 f'(2) &= \lim_{h \to 0} \frac{f(2 + h) - f(2)}{h} = \lim_{h \to 0} \frac{(2 + h)^2 - 2^2}{h} = \lim_{h \to 0} \frac{4h + h^2}{h} = \lim_{h \to 0} (4h + h) = 4 \\
 f'(x) &= \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = \lim_{h \to 0} \frac{(x + h)^2 - x^2}{h} = \lim_{h \to 0} \frac{2xh + h^2}{h} = \lim_{h \to 0} (2xh + h) = 2x
\end{align*}
\]

**Theorem 4-2.** If \( f(x) \) and \( g(x) \) are differentiable, then

(i) \( (\lambda f)'(x) = \lambda f'(x) \)

(ii) \( (f + g)'(x) = f'(x) + g'(x) \)
(iii) \((f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)\)

(iv) \(\left( \frac{f}{g} \right)'(x) = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g^2(x)}\) when \(g \neq 0\)

**Proof.** For (iii), evaluate

\[
\frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0} = \frac{f(x)g(x) - f(x)g(x_0) + f(x)g(x_0) - f(x_0)g(x_0)}{x - x_0}
\]

\[
= f(x) \frac{g(x) - g(x_0)}{x - x_0} + g(x_0) \frac{f(x) - f(x_0)}{x - x_0}
\]

(iv)

\[
\frac{f(x)/g(x) - f(x_0)/g(x_0)}{x - x_0} = \frac{1}{g(x)g(x_0)} \left[ g(x_0) \frac{f(x) - f(x_0)}{x - x_0} - f(x_0) \frac{g(x) - g(x_0)}{x - x_0} \right]
\]

By (ii) and (iii), polynomials are differentiable.

**Example 4-3.** Let’s find the derivative of a power function, \(f(x) = x^r, \ r \in \mathbb{Z}_{++}\). By binomial theorem,

\[
f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = \lim_{h \to 0} \frac{(x + h)^r - x^r}{h}
\]

\[
= \lim_{h \to 0} \frac{\binom{r}{0} x^r h^0 + \binom{r}{1} x^{r-1} h + \cdots + \binom{r}{k} x^{r-k} h^k + \cdots + \binom{r}{0} x^{0} h^n}{h}
\]

\[
= \lim_{h \to 0} \frac{\binom{r}{1} x^{r-1} h + \cdots + \binom{r}{k} x^{r-k} h^k + \cdots + \binom{r}{0} x^{0} h^n}{h}
\]

\[
= \lim_{h \to 0} \left( \binom{r}{1} x^{r-1} + \binom{r}{2} x^{r-2} h + \cdots + \binom{r}{k} x^{r-k} h^{k-1} + \cdots + \binom{r}{0} x^{0} h^{n-1} \right)
\]

\[
= \binom{r}{1} x^{r-1} = r x^{r-1}
\]

This formula holds even for a non-positive real number \(r\). See **Theorem 4-6**.

\(\Box\)

### 4.2 Composite Function

**Theorem 4-3 (Chain Rule)** \((f \circ g)'(x) = f'(g(x)) \cdot g'(x)\) if \(f\) and \(g\) are differentiable and \(f \circ g\) is defined.
\[ Df \circ g(x) = f'(y) \cdot g'(x), \quad y = g(x) \]

**Proof.**

\[
\lim_{x \to x_0} \frac{f \circ g(x) - f \circ g(x_0)}{x - x_0} = \lim_{x \to x_0} \frac{f \circ g(x) - f \circ g(x_0)}{g(x) - g(x_0)} \cdot \frac{g(x) - g(x_0)}{x - x_0}
\]

This argument is informal because it is possible to have \( g(x) = g(x_0) \) even if \( x \neq x_0 \). To avoid the problem, use

\[
\lim_{x \to x_0} \left( F(x) \times \frac{g(x) - g(x_0)}{x - x_0} \right), \quad F(x) = \begin{cases} \frac{f(g(x)) - f(g(x_0))}{g(x) - g(x_0)} & g(x) \neq g(x_0) \\ f'(g(x_0)) & g(x) = g(x_0) \end{cases}
\]

\[
\lim_{x \to x_0} \frac{f \circ g(x) - f \circ g(x_0)}{x - x_0} = \lim_{x \to x_0} \left[ F(x) \times \frac{g(x) - g(x_0)}{x - x_0} \right] = \lim_{x \to x_0} F(x) \times \lim_{x \to x_0} \frac{g(x) - g(x_0)}{x - x_0}
\]

Therefore, \( \lim_{x \to x_0} F(x) = f'(g(x_0)) \) and that completes the proof. \( \square \)

**Example 4-4.** Consider \( f(x) = (x^2 + 2x)^3 \). Since \( f(x) = h \circ g(x) \), where \( h(y) = y^3 \) and \( g(x) = x^2 + 2x \), and \( h'(y) = 3y^2 \) and \( g'(x) = 2(x + 1) \),

\[ f'(x) = h'(y)g'(x) \Big|_{y=x^2+2x} = 3(x^2 + 2x)^2 \times 2(x + 1) = 6(x^2 + 2x)^2(x + 1) \]

A tangent line serves as a linear approximation of a nonlinear function and the derivative is its slope. Therefore, if the derivative of a function is positive over an interval, the function is increasing.

**Theorem 4-4.** For a differentiable function \( f(x) \) on \((a, b)\), if \( f'(x) > 0 \) for all \( x \in (a, b) \), then \( f(x) \) is strictly increasing.

**Proof.** Since \( f'(x) > 0 \),

\[
\lim_{h \to 0} \frac{f(x + h) - f(x)}{h} > 0
\]

At every \( x \in (a, b) \), there exists a \( h_x > 0 \) such that \( f(y) - f(x) > 0 \) for all \( x < y < x + h_x \), and \( f(y) - f(x) < 0 \) for all \( x - h_x < y < x \). Therefore, it is strictly increasing. \( \square \)

The converse is not true because \( f(x) = x^3 \) is strictly increasing but \( f'(0) = 0 \).
These properties are very useful when we draw graphs of functions. By checking the signs of derivatives, we can identify whether the function is increasing or decreasing and whether there are local maxima and minima.

**Theorem 4-5 (Inverse Function Theorem)** If \( f: \mathbb{R} \to \mathbb{R} \) is continuously differentiable and \( f'(x_0) \neq 0 \), \( f(x) \) is locally invertible in a neighborhood of \( x_0 \), and

\[
Df^{-1}(y_0) = \frac{1}{Df(x_0)}, \quad y_0 = f(x_0)
\]

**Proof.** If \( f'(x) > 0 \), then \( f(x) \) is strictly increasing "near" \( x_0 \). Since \( f:X \to f(X) \) is invertible if and only if it is a bijection on \( f(X) \), either \( f'(x) \geq 0 \) or \( f'(x) \leq 0 \) for all \( x \in X \) implies that it is invertible. The second part holds because \( f^{-1} \circ f(x) = x \) and \( D(f^{-1} \circ f(x)) = 1 \).

The theorem asserts that if the linear approximation of a function is invertible, then the function is locally invertible. For one variable functions, it requires only differentiability and nonzero derivative at a point.

**Example 4-5.** Let’s find \( Df^{-1}(y) \) at \( x = 1 \).

\[
f(x) = \frac{x - 1}{x + 1}, \quad f'(x) = \frac{2}{(x + 1)^2} \bigg|_{x=1} = \frac{1}{2}
\]

\( Df^{-1}(0) = 2 \) because of \( f(1) = 0 \). A common mistake is to write \( Df^{-1}(1) = 2 \). Since the point we consider is \((1,0)\), the point that we use to evaluate \( Df^{-1}(y) \) is not 1 but 0.

### 4.3 Basic Functions

**Theorem 4-6.**

(i) \( D(ax + b) = a, \quad x \in \mathbb{R} \)

(ii) \( Dx^n = nx^{n-1}, \quad x \in \mathbb{R}_{++}, n \in \mathbb{R} \)

(iii) \( D \ln x = 1/x, \quad x \in \mathbb{R}_{++} \)

(iv) \( De^x = e^x, \quad x \in \mathbb{R} \)
(v) \( D \sin \theta = \cos \theta \) and \( D \cos \theta = -\sin \theta \), \( \theta \in \mathbb{R} \)

**Proof.** (i) Apply the definition and find the limit.

(ii) We need the result of (iii) and (iv) for the proof.

\[
f(x) = x^n = e^{n \ln x} \quad \text{and thus} \quad f'(x) = e^{n \ln x} \times D_x e^{n \ln x} = e^{n \ln x} \times \frac{n}{x} = nx^{n-1}
\]

(iii) Since logarithmic and exponential functions are inverse to each other, once one of the derivatives is characterized, it is easy to find the other. We take \( e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n \) as the definition of \( e^x \) (to prove the existence of the limit, we have to use the derivative of the logarithmic function. In this sense, this proof is not complete. See Example 5-2).

\[
\frac{d \ln x}{dx} = \lim_{h \to 0} \frac{\ln(x+h) - \ln x}{h} = \lim_{h \to 0} \frac{1}{h} \ln \left(\frac{x+h}{x}\right) = \lim_{h \to 0} \frac{1}{h} \ln \left(1 + \frac{h}{x}\right) = \lim_{n \to 0} \ln \left(1 + \frac{h}{x}\right)^{x/h}
\]

Letting \( n = \frac{x}{h} \) or \( h = \frac{x}{n} \),

\[
\frac{1}{x} \times \lim_{h \to 0} \ln \left(1 + \frac{h}{x}\right)^{x/h} = \frac{1}{x} \times \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = \frac{1}{x} \times \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n
\]

The last inequality follows from the continuity of \( \ln x \). Therefore, by the definition of \( e^x \),

\[
\frac{d \ln x}{dx} = \frac{1}{x} \times \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = \frac{1}{x} \times \ln e = \frac{1}{x}
\]

Alternatively, we can apply the second fundamental theorem of calculus of Theorem 17-6 to another definition of logarithm, \( \ln x = \int_1^x \frac{1}{t} \, dt \) so that \( D \ln x = 1/x \).

(iv) Given the derivative of the logarithm, by differentiating the identity \( \ln(e^x) = x \), we have \( De^x/e^x = 1 \) or \( De^x = e^x \).

(v) \( D \sin \theta = \cos \theta \) and \( D \cos \theta = -\sin \theta \), \( \theta \in \mathbb{R} \)

To find the derivative of the triangular function, we need the limits of two expressions.

\[
\frac{\sin \theta}{\theta} \quad \text{and} \quad \frac{1 - \cos \theta}{\theta}
\]

Let’s first find the limit of \( \frac{\sin \theta}{\theta} \).
In the figure above, we have the following inequality in the neighborhood of 0,
\[ \sin \theta < \theta < \tan \theta \]
Dividing it by \( \sin \theta \) gives
\[ 1 < \frac{\theta}{\sin \theta} < \frac{\tan \theta}{\sin \theta} = \frac{1}{\cos \theta}, \quad \cos \theta < \frac{\sin \theta}{\theta} < 1 \]
Since \( \lim \theta \to 0 \cos \theta = 1 \), by Sandwich theorem,
\[ \lim_{\theta \to 0} \frac{\sin \theta}{\theta} = 1 \]
Secondly,
\[
\frac{1 - \cos x}{x} = \frac{1 - \cos x}{x} \times \frac{1 + \cos x}{1 + \cos x} = \frac{1 - \cos^2 x}{x(1 + \cos x)} = \frac{\sin^2 x}{x(1 + \cos x)} = \frac{\sin x}{x} \times \frac{\sin x}{1 + \cos x}
\]
\[ \lim_{\theta \to 0} \left[ \frac{\sin x}{x} \times \frac{\sin x}{1 + \cos x} \right] = \lim_{\theta \to 0} \frac{\sin x}{x} \times \lim_{\theta \to 0} \frac{\sin x}{1 + \cos x} = \lim_{\theta \to 0} \frac{\sin x}{1 + \cos x} = 0 \]
Now we are ready to find the derivative of the sine function.\(^2\)
\[
\frac{d}{d\theta} \sin \theta = \lim_{h \to 0} \frac{\sin(\theta + h) - \sin \theta}{h} = \lim_{h \to 0} \frac{\cos \theta \sin h + \sin \theta \cos h - \sin \theta}{h}
\]
\[^2\text{compound-angle formulae}\]
\[
\cos(A + B) = \cos A \cos B - \sin A \sin B, \quad \cos(A - B) = \cos A \cos B + \sin A \sin B \\
\sin(A + B) = \sin A \cos B + \cos A \sin B, \quad \sin(A - B) = \sin A \cos B - \cos A \sin B
\]
= \cos \theta \lim_{h \to 0} \frac{\sin h}{h} - \sin \theta \lim_{h \to 0} \frac{1 - \cos h}{h} = \cos \theta

Since \( \sin^2 \theta + \cos^2 \theta = 1 \),

\[ 2 \sin \theta \frac{d}{d\theta} \sin \theta + 2 \cos \theta \frac{d}{d\theta} \cos \theta = 0 \Rightarrow 2 \sin \theta \cos \theta + 2 \cos \theta \frac{d}{d\theta} \cos \theta = 0 \]

Therefore, \( \frac{d}{d\theta} \cos \theta = -\sin \theta \).

The following figure demonstrates the geometric intuition as the two shaded right triangles are congruent.

![Geometric Intuition Diagram]

\[ \frac{d}{d\theta} \sin \theta = \cos \theta, \quad \frac{d}{d\theta} \cos \theta = -\sin \theta \]

**Example 4.6.**

\[ D \log_a x = \frac{\ln x}{\ln a} = \frac{1}{x \ln a}, \quad D \log_a a = \frac{\ln a}{\ln x} = -\frac{\ln a}{x(\ln x)^2} \]

Consider \( D_x f(x) \). The following trick is useful when it is difficult to isolate the expression involving the variable.

Let \( g(x) = x f(x) \).

\[ \ln g(x) = f(x) \ln x, \quad \frac{g'(x)}{g(x)} = f'(x) \ln x + \frac{f(x)}{x}, \quad g'(x) = g(x) \left( f'(x) \ln x + \frac{f(x)}{x} \right) \]

or
\[ D x^f(x) = x^f(x) \left( f'(x) \ln x + \frac{f(x)}{x} \right) \]

4.4 Higher Order Derivatives

If \( f(x) \) is differentiable, then you can construct a new function, called the second derivative of \( f'(x) \), \( f''(x) \), provided that \( f'(x) \) is differentiable:

\[ f'''(x) = \lim_{x \to x_0} \frac{f'(x) - f'(x_0)}{x - x_0} \]

The second derivative is the derivative of the first derivative. If \( f^{(n-1)}(x) \) exists and is differentiable, then we denote its derivative by \( f^{(n)}(x) \). We call it the \( n^{th} \) derivative.

The first derivative has a nice interpretation of slope, the change in function value for a unit change of the argument. The second derivative can also be interpreted in the same way: the change in the first derivative for a unit change of the argument. If \( f''(x) < 0 \), the slope of the tangent line gets smaller as \( x \) increases.

**Definition 4-2.** If \( f^{(n-1)}(x) \) is differentiable and \( f^{(n)}(x) \) is continuous, then \( f(x) \) is said to be \( n \) times **continuously differentiable**. The set of \( n \) times continuously differentiable functions is denoted by \( C^n \).

4.5 Differential

In many applications, it is sometimes convenient to use a linear approximation of a function.

**Definition 4-3.** The differential of a function \( f(x) \) is the function \( df \) of two independent real variables \( x \) and \( dx \) given by \( df(x, dx) = f'(x)dx \). Since \( dx(x, dx) = dx \), we usually write \( df(x) = f'(x)dx \).

Note that \( dx \in \mathbb{R} \) is an additional real variable that is independent of \( x \in X \). Since the points on the tangent line of \( f(x) \) at \( x_0 \) have a relationship of \( y = f'(x_0)x \), the differential can be regarded as the equation of the tangent line or a linear approximation of a function. For this reason, a derivative is sometimes referred to as a differential coefficient.
Let $\Delta y = f(x + \Delta x) - f(x)$. (I use $\Delta$ notation to distinguish the variable $dx$ in differential from the $dx$ in the operator $df(x)/dx$.)

$$\Delta y = \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad \text{and} \quad \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = f'(x)$$

This implies that the limit of the ratio of differentials, $dy$ and $dx$, is $f'(x)$. That is, the derivative is the same as the limit of a quotient of differentials.

The approximation error of differential as a first-order approximation vanishes quite fast such as $\varepsilon/\Delta x \to 0$ as $\Delta x \to 0$. Since $\Delta y = f'(x)\Delta x + \varepsilon$ where $\varepsilon$ is the error in approximation,

$$\Delta y = \frac{f(x + \Delta x) - f(x)}{\Delta x} + \frac{\varepsilon}{\Delta x}$$

$\Delta x \to 0$, $\varepsilon/\Delta x \to 0$.

**Example 4-7.** The result sometimes helps to simplify calculations.

$$\frac{d \ln x(p)}{d \ln p} = \frac{d \ln x(e^q)}{dq} = \frac{1}{x(p)} \frac{dx(p)}{dp} \times e^q = \frac{p}{x(p)} \frac{dx(p)}{dp}$$

With differentials, we do not have to find the inverse of $\ln p$.

$$d \ln x(p) = \frac{1}{x(p)} \frac{dx(p)}{dp} \quad \text{and} \quad d \ln p = \frac{1}{p} \frac{dp}{dp}, \quad d \ln x(p) = \frac{p}{x(p)} \frac{dx(p)}{dp} \quad \square$$

5 Applications of Derivatives

5.1 Local Extrema

The maximum and minimum of a function, collectively called extrema, always exists if the range is a compact set. The definition is identical to the maximum of a subset of real numbers.

**Definition 5-1.** The point $x^*$ is a **global maximum** (or **absolute Maximum**) point of $f: X \to Y$ if $f(x) \leq f(x^*)$ for all $x \in X$. A **global minimum** is defined similarly.
However, there is no simple way to characterize the absolute maximum in general. In particular, if an extremum could occur at the boundary of the domain, we have to check every boundary point.

On the other hand, at a global maximum, the function value must get smaller by a change of \( x \) no matter how large or small the change is. But, the converse is not true because of the presence of local extrema.

**Definition 5-2.** The point \( x^* \) is a **local maximum** (or **relative maximum**) point of \( f(x) \) if there exists a \( \delta > 0 \) such that \( f(x) \leq f(x^*) \) for all \( x \) with \( |x - x^*| < \delta \).

**Definition 5-3.** The point \( x^* \) is a **local minimum** (or **relative minimum**) point of \( f(x) \) if there exists a \( \delta > 0 \) such that \( f(x^*) \leq f(x) \) for all \( x \) with \( |x - x^*| < \delta \).

It is now true that a point is a local maximum if and only if the function value gets smaller by an infinitesimally small change of the point. Note that to be a local maximum point \( f(x) \) must be defined over the interval \( B_\delta(x^*) \). The first figure below shows that the point is a global maximum but not a local maximum.

![Figure 5-1 Global Maximum and Local Maxima](image)

If a slope is well defined, the sign of slope completely captures the relationship between the changes in \( x \) and \( f(x) \). To utilize the property, we restrict the analysis to the set of continuous functions. In all three cases in Figure 5-1, the slope is not defined at the solid dots.

If a function is continuous, then the condition can be restated that if a function attains its local maximum at \( x^* \), then the slope must be positive over \( (x^* - \varepsilon, x^*) \) and negative over \( (x^*, x^* + \varepsilon) \) for some \( \varepsilon > 0 \). Moreover, if a function is differentiable, the sign of the slope is changed from positive to negative as the graph crosses the local maximum point and we must have
We have \( f'(x^*) = 0 \). Since we have exactly the opposite at a local minimum, we have the same requirement for the derivative.

**Theorem 5.1 (First Order Necessary Condition)** If \( f(x) \) is differentiable at \( x_0 \) and \( x_0 \) is a local maximum or a local minimum point, then \( f(x^*) = 0 \).

**Proof.** This proof considers the only maximum. For \( x > x^* \), \( \frac{f(x) - f(x^*)}{x - x^*} \leq 0 \) and \( \frac{f(x) - f(x^*)}{x - x^*} \geq 0 \) for \( x < x^* \). Therefore, \( \lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*} \leq 0 \) and \( \lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*} \geq 0 \) if existed. By differentiability, both limits exist and are equal.

\( f'(x) = 0 \) could happen in a flat region of a graph or at an inflection point. The first order condition is not sufficient to capture the change in the sign of slope. The condition of the local maximum requires that the slope changes from positive to negative. That is, it requires a negative second derivative.

**Theorem 5.2 (Second Order Sufficient Condition)** Suppose \( f(x) \) is twice continuously differentiable on an open interval \((a, b)\) and \( x^* \in (a, b) \) is a critical point of \( f(x) \).

(i) If \( x^* \) is a local maximum, then \( f''(x^*) \leq 0 \).

(ii) If \( x^* \) is a local minimum, then \( f''(x^*) \geq 0 \).

(iii) If \( f''(x^*) < 0 \), then \( x^* \) is a local maximum.

(iv) If \( f''(x^*) > 0 \), \( x^* \) is a local minimum.

**Proof.** (i) Note that \( f'(x) \geq 0 \) for \( x \in (a, x^*) \) and \( f'(x) \leq 0 \) for \( x \in (x^*, b) \). Therefore,

\[
\lim_{x \to x^*} \frac{f'(x) - f'(x^*)}{x - x^*} = \lim_{x \to x^*} \frac{f'(x)}{x - x^*} \leq 0 \quad \text{and similarly} \quad \lim_{x \to x^*} \frac{f'(x)}{x - x^*} \leq 0
\]

Since \( f(x) \) is twice continuously differentiable at \( x^* \), \( f''(x^*) \leq 0 \).

(iii) Since \( f''(x^*) < 0 \), by the continuity of \( f''(x) \), there exists a \( \epsilon > 0 \) such that \( f''(x) < 0 \) for all \( |x - x^*| < \epsilon \). \( f'(x) \) is decreasing on \((x^* - \epsilon < x < x^* + \epsilon)\) and \( f'(x^*) = 0 \). Therefore, we have \( f'(x) > 0 \) for \( x \in (x^* - \epsilon, x^*) \) and \( f'(x) < 0 \) for \( x \in (x^*, x^* + \epsilon) \). Therefore, \( f(x) \) is increasing on \((x^* - \epsilon, x^*)\) and decreasing on \((x^*, x^* + \epsilon)\).
The proofs of (ii) and (iv) are identical with inversing inequalities. 

In the proof of (iii) and (iv), strict inequality is essential. Note that (i) and (iii) are not converse to each other as well as (ii) and (iv). Check the shape of $x^4$.

**THEOREM 5-1** shows that local extrema must be critical points, and we can find the critical points by solving the first order condition. The second derivatives tell whether a critical point is maxima or minima.

**THEOREM 5-3 (Necessary Conditions)** Suppose that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable and $f(x^*)$ attains its local maximum at $x^*$. Then $f'(x^*) = 0$ and $D^2f(x^*) \leq 0$.

**THEOREM 5-4 (Sufficient Conditions)** Suppose that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable. If $Df(x^*) = 0$ and $D^2f(x^*) < 0$, $f(x^*)$ attains its local maximum at $x^*$.

Again if $Df(x) = D^2f(x) = 0$, we cannot tell whether $x$ is an extremum point or not. See **THEOREM 5-20** for a treatment.

### 5.2 MEAN VALUE THEOREM

**THEOREM 5-5 (Rolle’s Theorem)** Suppose that $f$ defined on $[a, b]$ is continuous on $[a, b]$ and differentiable on $(a, b)$. If $f(a) = f(b)$ then, for some $c \in (a, b)$, $f'(c) = 0$.

**Proof.** If an extremum is different from $f(a) = f(b)$, then the result follow from **THEOREM 5-1**. Otherwise $f(x)$ is constant over $[a, b]$ and $f'(x) = 0$ for all $x$. 

**THEOREM 5-6 (Mean Value Theorem)** Suppose that $f$ defined on $[a, b]$ is continuous on $[a, b]$ and differentiable on $(a, b)$. Then, for some $c \in (a, b)$,

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

The mean value theorem implies that $f(b) = f(a) + (b - a)f'(c)$ for some value of $c \in (a, b)$ and when $f(a)$ and $f'(a)$ are known, $f(b) = f(a) + (b - a)f'(a)$ is a good linear approximation of $f(b)$, especially when $b$ is close to $a$. 

---

The mean value theorem implies that $f(b) = f(a) + (b - a)f'(c)$ for some value of $c \in (a, b)$ and when $f(a)$ and $f'(a)$ are known, $f(b) = f(a) + (b - a)f'(a)$ is a good linear approximation of $f(b)$, especially when $b$ is close to $a$. 

---

Math for Economists, Sogang University
5 Applications of Derivatives

**Proof.** Let \( F(x) = f(x) - \left[ \frac{f(b) - f(a)}{b - a} \right] x \), (from \( F(x) = f(x) + hx \), \( h \) is chosen so that \( F(a) = F(b) \)). Since \( F(a) = F(b) \) and \( F'(x) = f'(x) - \left[ \frac{f(b) - f(a)}{b - a} \right] \), by Rolle's theorem, there exists \( c \in (a, b) \) such that \( F'(x) = 0 \) or \( f'(c) = \frac{f(b) - f(a)}{b - a} \).

**Example 5-1.** Theorem 4-4 is direct from Theorem 5-6.

**Theorem 5-7.** If \( f(x) \) is continuous on \([a, b]\) and differentiable on \((a, b)\), and \( f'(x) = 0 \) for all \( x \in (a, b) \), then \( f(x) \) is constant for all \( x \in [a, b] \).

**Proof.** By the mean value theorem, \( 0 = f'(c) = \frac{f(y) - f(a)}{y - a} \) for all \( a < y \leq b \) where \( c \in (a, y) \). Therefore, \( f(y) = f(a) \).

5.3 L'Hôpital's Rule

The limit of an indeterminate form is sometimes difficult to find. If a function is the quotient of differentiable functions, there is a simple formula to handle the case. Let's start with a theorem that is required to derive the formula.

**Theorem 5-8 (The Cauchy's Mean Value Theorem)** If \( f \) and \( g \) are continuous on \([a, b]\) and differentiable on \((a, b)\) with \( g' \neq 0 \) on \((a, b)\), then there exists \( c \in (a, b) \) such that

\[
\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.
\]

When \( g(x) = x \), this is MVT. This is not trivial because MVT shows only \( f'(c) = \frac{f(b) - f(a)}{b - a} \) and \( g'(d) = \frac{g(b) - g(a)}{b - a} \).

**Proof.** Apply Rolle's theorem to \( F(x) = f(x) - \left[ \frac{f(b) - f(a)}{g(b) - g(a)} \right] g(x) \). Since

\[
f(a) - \left[ \frac{f(b) - f(a)}{g(b) - g(a)} \right] g(a) = f(b) - \left[ \frac{f(b) - f(a)}{g(b) - g(a)} \right] g(b),
\]

there exists \( c \in (a, b) \) such that \( F'(c) = 0 \) or \( f'(c) = \frac{f(b) - f(a)}{g(b) - g(a)} g'(c) \).
**Theorem 5.9 (L'Hôpital's Rule)** Suppose $f$ and $g$ are continuous on $[a, b]$ and differentiable on $(a, b)$. If $\lim_{x \to a} \frac{f'(x)}{g'(x)}$ exists and $f(a) = g(a) = 0$,

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

**Proof.** Let $\lim_{x \to a} \frac{f'(x)}{g'(x)} = L$. Since the limit exists, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left| \frac{f'(x)}{g'(x)} - L \right| < \frac{\varepsilon}{2} \text{ whenever } a < x < a + \delta.$$

Since $g'(x) \neq 0$ on $(a, a + \delta)$ and $f$ and $g$ are continuous on $[a, b]$, by the Cauchy's mean value theorem, there exists $a < c < x$, such that

$$\frac{f'(c)}{g'(c)} = \frac{f(x) - f(a)}{g(x) - g(a)}.$$

Since $f(a) = g(a) = 0$,

$$\frac{f'(c)}{g'(c)} = \frac{f(x)}{g(x)}.$$

Since the (right) limit exists, as $\delta$ tends to 0, $x$ converges to $a$ and so does $c$. Therefore,

$$\left| \frac{f'(x)}{g'(x)} - L \right| < \frac{\varepsilon}{2} \iff \left| \frac{f'(c)}{g'(c)} - L \right| = \left| \frac{f(x)}{g(x)} - L \right| < \frac{\varepsilon}{2}.$$

**Corollary 5.10.** Suppose $f$ and $g$ are differentiable on $(a, b)$. If $\lim_{x \to a} \frac{f'(x)}{g'(x)}$ exists and $\lim_{x \to a} |f(x)| = \lim_{x \to a} |g(x)| = \infty$,

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)}.$$

**Proof.** Let $\lim_{x \to a} \frac{f'(x)}{g'(x)} = L$. Since the limit exists, for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left| \frac{f'(x)}{g'(x)} - L \right| < \frac{\varepsilon}{2} \text{ whenever } a < x < a + \delta.$$

By the Cauchy's mean value theorem, there exists a $c \in (x, a + \delta)$ with $a < x < a + \delta$ such that
\[ \frac{f'(c)}{g'(c)} = \frac{f(x) - f(a + \delta)}{g(x) - g(a + \delta)} = \frac{\frac{f(x)}{g(x)} - \frac{f(a + \delta)}{g(x)}}{1 - \frac{g(a + \delta)}{g(x)}} \]

\[ \frac{f'(c)}{g'(c)} \left( 1 - \frac{g(a + \delta)}{g(x)} \right) = \frac{f(x) - f(a + \delta)}{g(x)} \]

\[ \frac{f'(c)}{g'(c)} = \frac{f(x)}{g(x)} - \left( \frac{f(a + \delta)}{g(x)} \right) \times \frac{g(a + \delta)}{g(x)} \]

Since \( \left| \frac{f'(c)}{g'(c)} - L \right| < \frac{\epsilon}{2} \) for all \( a < z < a + \delta \), it should hold for \( z = c \). We have

\[ \left| \frac{f'(c)}{g'(c)} - L \right| < \frac{\epsilon}{2} \quad \text{whenever} \quad a < x < c < a + \delta \]

Combining the two expressions above,

\[ \left| \frac{f(x)}{g(x)} - L - \left( \frac{f(a + \delta)}{g(x)} \right) \times \frac{g(a + \delta)}{g(x)} \right| < \frac{\epsilon}{2} \quad \text{for all} \quad a < x < c < a + \delta \]

As \( x \to a^+ \), note that \( \frac{f(a + \delta)}{g(x)} \) and \( \frac{g(a + \delta)}{g(x)} \) vanish as \( |g(x)| \to \infty \), and \( \frac{f'(c)}{g'(c)} \) is bounded. Since the terms in the parenthesis approaches to zero, there exists \( \gamma > 0 \) such that

\[ \left| \frac{f(a + \gamma)}{g(x)} - \frac{f'(c)}{g'(c)} \times \frac{g(a + \gamma)}{g(x)} \right| < \frac{\epsilon}{2} \quad \text{for all} \quad a < x \leq c < a + \gamma. \]

Therefore,

\[ \left| \frac{f(x)}{g(x)} - L \right| < \epsilon \quad \text{for all} \quad a < x \leq c < a + \min(\delta, \gamma). \]

**COROLLARY 5-11.** L'Hôpital's rule holds for the limits as \( x \to \pm \infty \).

**Proof.** When \( x \to \infty \), we can change variables with \( t = x^{-1} \) to convert limits as \( x \to \infty \) to one side limits as \( t \to 0^+ \).

\[ \lim_{x \to \infty} \frac{f'(x)}{g'(x)} = \lim_{t \to 0^+} \frac{f'(t^{-1})}{g'(t^{-1})} = L \]

By Chain Rule,
\[
\frac{D_t f(t^{-1})}{D_t g(t^{-1})} = \frac{f'(t^{-1}) \times (-t^2)}{g'(t^{-1}) \times (-t^2)} = \frac{f'(t^{-1})}{g'(t^{-1})}
\]

Therefore, the result follows from L'Hôpital's Rule and COROLLARY 5-10.

\[
\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{t \to 0^+} \frac{f(t^{-1})}{g(t^{-1})} = \lim_{t \to 0^+} D_t f(t^{-1}) = D_t g(t^{-1}) = L. \tag*{\blacksquare}
\]

**COROLLARY 5-12.** L'Hôpital's Rule holds when the limit is infinity,

\[
\lim_{x \to a^+} \frac{f'(x)}{g'(x)} = \pm \infty.
\]

**Proof.** When \( f(a) = g(a) = 0 \),

\[
\frac{f'(c)}{g'(c)} = \frac{f(c) - f(a)}{g(c) - g(a)} = \frac{f(c) - f(a)}{g(c) - g(a)}
\]

Since \( x \) and \( c \) are chosen such that \( a < x < c < a + \delta \), \( c \) and \( x \) converges to \( a \) as \( \delta \to 0 \) and \( \frac{f'(c)}{g'(c)} \) tends to infinity.

When \( \lim_{x \to a^+} |f(x)| = \lim_{x \to a^+} |g(x)| = \infty \), the previous proof can be modified easily. For any \( M > 0 \), there exists \( \delta > 0 \) such that

\[
\frac{f'(x)}{g'(x)} > M \quad \text{whenever} \quad a < x < a + \delta.
\]

Since \( \frac{f'(x)}{g'(x)} > \frac{M}{2} \) for all \( a < x < a + \delta \), we have

\[
\frac{f'(c)}{g'(c)} > M \quad \text{whenever} \quad a < x < c < a + \delta.
\]

As in the proof of COROLLARY 5-10, combining it with

\[
\frac{f'(c)}{g'(c)} = \frac{f(c)}{g(c)} - \left( \frac{f(a + \delta)}{g(a + \delta)} - \frac{f'(c)}{g'(c)} \times \frac{g(a + \delta)}{g(x)} \right)
\]

gives

\[
\left| \frac{f(x)}{g(x)} - \left( \frac{f(a + \delta)}{g(a + \delta)} - \frac{f'(c)}{g'(c)} \times \frac{g(a + \delta)}{g(x)} \right) \right| > M \quad \text{for all} \quad a < x < c < a + \delta.
\]

Math for Economists, Sogang University
As \( x \to a^+ \), note that \( \frac{f(a+\delta)}{g(x)} \) and \( \frac{g(a+\delta)}{g(x)} \) vanish as \( |g(x)| \to \infty \), and \( \frac{f'(c)}{g'(c)} \) is bounded. Since the terms in the parenthesis approaches to zero, for any \( \varepsilon > 0 \), there exists \( \gamma > 0 \) such that

\[
\left| \frac{f(a+\gamma)}{g(x)} - \frac{f'(c)}{g'(c)} \times \frac{g(a+\gamma)}{g(x)} \right| < \varepsilon \quad \text{for all} \quad a < x < c < a + \gamma.
\]

Therefore,

\[
\left| \frac{f(x)}{g(x)} - L \right| > \frac{M}{2} \quad \text{for all} \quad a < x < c < a + \min(\delta, \gamma).
\]

The results of THEOREM 5-9 through COROLLARY 5-12 show that L'Hôpital's rule can be applied to any indeterminant quotient forms.

**EXAMPLE 5-2. A proof of EXAMPLE 3-6**

Let \( f(n) = \left(1 + \frac{x}{n}\right)^n \) and \( g(n) = n \ln \left(1 + \frac{x}{n}\right) = \frac{\ln(1 + \frac{x}{n})}{1/n} \) for \( n \neq 0 \). Apply L'Hôpital's rule to \( \lim_{n \to \infty} g(n) \) to get

\[
\lim_{n \to \infty} g(n) = \lim_{n \to \infty} \frac{\ln \left(1 + \frac{x}{n}\right)}{1/n} = \lim_{n \to \infty} -\frac{x}{n^2 \left(1 + \frac{x}{n}\right)} = r
\]

Since \( f(n) = \exp(g(n)) \) and \( e^x \) is continuous,

\[
\lim_{n \to \infty} f(n) = \lim_{n \to \infty} \exp(g(n)) = \exp \left( \lim_{n \to \infty} g(n) \right) = e^x.
\]

**EXAMPLE 5-3.** If limit (including infinity) exists, we can apply the rule recursively. For any \( n > 0 \),

\[
\lim_{x \to \infty} \frac{e^x}{x^n} = \lim_{x \to \infty} \frac{e^x}{nx^{n-1}} = \cdots = \lim_{x \to \infty} \frac{e^x}{n \times (n-1) \times \cdots \times 2 \times x} = \lim_{x \to \infty} \frac{e^x}{n!} = \infty
\]

**THEOREM 5-13.** If a function is known to be differentiable over \( X \) except for \( a \in X \) and continuous at \( a \). Then

\[
f''(a) = \lim_{x \to a^+} f''(x)
\]
**Proof.** Consider the function $h(x) = f(x) - f(a)$ and $g(x) = x - a$. By the continuity of $f$ at $a$, $\lim_{x \to a} h(x) = 0$, and $\lim_{x \to a} g(x) = 0$ because a polynomial is continuous everywhere. Then by L'Hôpital's Rule,

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a} \frac{h(x)}{g(a)} = \lim_{x \to a} f'(x).$$

This can be shown using Mean Value Theorem with $\frac{f(x) - f(a)}{x - a} = f'(c)$, and right hand limit and left hand limit.

### 5.4 **Taylor's Theorem**

For differentiable functions, the derivatives give information about tangents to graphs. For $f: \mathbb{R} \to \mathbb{R}$, Graph$(f) = \{(x, y)|f(x) = y\}$. The line tangent to the graph at $(x_0, f(x_0))$ has the slope of $f'(x_0)$. In the equation, $y - f(x_0) = (x - x_0)f'(x_0)$.

The derivative at $x_0$ can be used to an approximation of the function value at points near $x_0$. The slope of the line connecting $(x_0, f(x_0))$ to $(x, f(x))$ is

$$\frac{f(x) - f(x_0)}{x - x_0}$$

Thus the limit gives a slope that is valid "near" $x_0$ as a good approximation. The linear or the first-order approximation of $f(x)$ at $x_0$ is

$$A_1(x) = f(x_0) + f'(x_0)(x - x_0)$$

Observe that the difference $f(x) - A_1(x)$ converges to zero and thus it is quite accurate in the close neighborhood of $x_0$.

$$\lim_{x \to x_0} \frac{f(x) - A_1(x)}{x - x_0} = \lim_{x \to x_0} \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0)$$

If we need a better approximation, a quadratic form could be used.

$$A_2(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

so that
\[ \lim_{x \to x_0} \frac{f(x) - A_2(x)}{(x - x_0)^2} = 0. \]

The second order approximation is better in the sense that the approximation error is vanishing at the order of \( O((x - x_0)^2) \) as \( x \to x_0 \) while the convergence rate is \( O(x - x_0) \) in the linear approximation. This result can be generalized.

**Theorem 5-14 (Taylor’s Theorem)** Suppose that \( f \) is defined on an interval \([a, b]\) and \( f^{(n+1)} \) is defined on \((a, b)\). Then, for each \( x \in [a, b] \), there exists \( c \in [a, x] \) such that

\[
 f(x) = f(a) + f'(a)(x - a) + \cdots + \frac{1}{n!}f^{(n)}(a)(x - a)^n + \frac{1}{(n + 1)!}f^{(n+1)}(c)(x - a)^{n+1}
\]

**Proof.** Let

\[ F(z) = f(a) - \sum_{k=0}^{n} \frac{f^{(k)}(z)}{k!} (x - z)^k \]

\[ G(z) = F(z) - \left( \frac{x - z}{x - a} \right)^{n+1} F(a) \]

\[ F(x) = 0 \quad \text{and} \quad F'(z) = -\frac{f^{(n+1)}(z)}{n!} (x - z)^n, \quad G(x) = G(a) = 0. \]

By Mean Value Theorem, there exists a \( c \in [a, x] \) such that \( G'(c) = 0 \).

\[
0 = -\frac{f^{(n+1)}(z)}{n!} (x - z)^n + \left( \frac{x - c}{x - a} \right)^n F(a)
\]

\[ F(a) = \frac{f^{(n+1)}(c)}{(n + 1)!} (x - a)^{n+1} \]

Therefore,

\[
f(a) - \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x - a)^k = \frac{f^{(n+1)}(c)}{(n + 1)!} (x - a)^{n+1}.
\]

Rearranging terms completes the proof. \(\square\)

**Definition 5-4.** The \( n \)th order Taylor Polynomial, or the \( n \)th order Taylor Approximation is:
\[ A_n(x) = f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^n}{n!} f^{(n)}(a) \]

**Theorem 5-15 (Convergence Rate)**

\[ \lim_{x \to a} \frac{f(x) - A_n(x)}{(x-a)^n} = 0 \]

**Proof.** Let

\[ h_n(x) = \begin{cases} 
  f(x) - A_n(x) & x \neq a \\
  0 & x = a 
\end{cases} \]

Since the numerator and the denominator are zero at \( x = a \), we can apply L'Hôpital's Rule repeatedly.

\[
\lim_{x \to a} \frac{f(x) - A_n(x)}{(x-a)^n} = \lim_{x \to a} \frac{D[f(x) - A_n(x)]}{D(x-a)^n} = \cdots = \lim_{x \to a} \frac{D^{n-1}[f(x) - A_n(x)]}{D^{n-1}n(x-a)^{n-1}} \\
= \lim_{x \to a} \frac{f^{(n-1)}(x) - A_n^{(n-1)}(x)}{(n-1)! (x-a)} = \frac{1}{n!} [f^{(n)}(x) - f^{(n)}(a)] = 0. \quad \blacksquare
\]

Note that the exact value of \( c \) in the theorem does not matter to determine the convergence rate.

**Example 5-4.** If \( a \) satisfies \( f'(a) = 0 \), then by Taylor theorem, \( f(x) = f(a) + 0 + \frac{(x-a)^2}{2} f''(c) \). Suppose \( f''(a) \geq 0 \) and \( f'' \) is continuous at \( a \). Then \( f(x) = f(a) + \frac{(x-a)^2}{2} f''(c) \) implies that \( f(x) \geq f(a) \) for all \( x \) "close" to \( x_0 \). That is, if \( f'(a) = 0 \) and \( f''(a) \geq 0 \), then \( a \) is a local minimum.

**Example 5-5.** Let’s apply Taylor’s theorem to exponential and logarithm functions.

Since \( f(x) = f^{(n)}(x) = e^x \), and \( f(0) = f^{(n)}(0) = 1 \), by applying Taylor theorem,

\[ f(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^{n-1}}{(n-1)!} + \frac{x^n}{n!} + h_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!} + h_n(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} \]

In case of a logarithm function, \( f(x) = \ln x \), since

\[ f^{(k)}(x) = (k-1)! (-1)^{k-1} x^{-k}, \]

\[ f^{(k)}(1) = (k-1)! (-1)^{k-1} \]

and

Math for Economists, Sogang University
$$f(x) = \sum_{k=1}^{n} \frac{1}{k} (-1)^{k-1} (x - 1)^k + h_n(x) = \sum_{k=0}^{\infty} \frac{1}{k} (-1)^{k-1} (x - 1)^k$$

**Example 5-6.** Find the fourth order approximations of $\sqrt{2}(= 1.414213)$ and $\sqrt{3}(= 1.7320508)$ at 1.

$$f(x) = x^{\frac{1}{2}}, \quad f'(x) = \frac{1}{2} x^{-\frac{1}{2}}, \quad f''(x) = -\frac{1}{4} x^{-\frac{3}{2}}, \quad f'''(x) = \frac{3}{8} x^{-\frac{5}{2}}, \quad f^{(4)} = -\frac{15}{16} x^{-\frac{7}{2}}$$

$$\sqrt{2} = 1 + \frac{1}{2} + \frac{1}{2} \times \left(-\frac{1}{4}\right) + \frac{3}{8} + \frac{1}{4!} \times \left(-\frac{15}{16}\right) + h_4(3)$$

$$= 1 + \frac{1}{2} - \frac{1}{8} + \frac{1}{16} - \frac{15}{384} + h_4(3) = 1.3984$$

$$\sqrt{3} = 1 + \frac{1}{2} \times 2 + \frac{1}{2} \times \left(-\frac{1}{4}\right) \times 2^2 + \frac{3}{8} \times 2^3 + \frac{1}{4!} \times \left(-\frac{15}{16}\right) \times 2^4 + h_4(3)$$

$$= 1 + 1 - \frac{1}{2} + \frac{1}{2} \times \frac{15}{24} + h_4(3) = 1.375$$

**Example 5-7.** Taylor expansions of triangular functions

$$\cos \theta = \cos 0 + (-\sin 0)\theta + \frac{1}{2!} (-\cos 0)\theta^2 + \frac{1}{3!} (\sin 0)\theta^3 + \frac{1}{4!} (\cos 0)\theta^4 + \cdots$$

$$= 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots$$

$$\sin \theta = \sin 0 + (\cos 0)\theta + \frac{1}{2!} (-\sin 0)\theta^2 + \frac{1}{3!} (\cos 0)\theta^3 + \frac{1}{4!} (\sin 0)\theta^4 + \cdots$$

$$= \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots$$

**Example 5-8.** Euler’s formula is the fundamental relationship between the trigonometric functions and the complex exponential function.
When \( \theta = \pi \), we have Euler’s identity, \( e^{i\pi} + 1 = 0 \).

\[
\frac{\partial e^{i\theta}}{\partial \theta} = e^{i\theta} \times i, \quad \frac{\partial^2 e^{i\theta}}{\partial \theta^2} = e^{i\theta} \times i^2 = -e^{i\theta}, \quad \frac{\partial^3 e^{i\theta}}{\partial \theta^3} = -e^{i\theta} \times i, \quad \frac{\partial^4 e^{i\theta}}{\partial \theta^4} = e^{i\theta}
\]

### 5.5 Graph

Sketching the graph of a function is based mostly on the first derivative (the slope) and the second derivative (the curvature) with some information on several critical points.

**Definition 5.5.**

(i) \((x, f(x))\) is a **critical point** if \( f’(x) = 0 \) or \( f’(x) \) is not defined. If \( f’(x) = 0 \), it is called a **stationary point**.
(ii) \((x, f(x))\) is an **inflection point** if \(f''(x) = 0\) and \(f'''(x) \neq 0\).

(iii) \(\{(x, y): y = b\}\) is a **horizontal asymptote** if \(\lim_{x \to \infty} f(x) = b\).

(iv) \(\{(x, y): x = c\}\) is a **vertical asymptote** if \(\lim_{x \to c} f(x) = \pm\infty\).

If \(f'(x) > 0\) on an interval, the function is increasing in the interval.

If \(f''(x) > 0\) (\(f''(x) < 0\)) on an interval, the slope of the function is increasing (decreasing), and is said that the function is convex (concave) in the interval.

To determine the shape of a function:

Step 1. Find the intervals where \(f(x) > 0\), \(f(x) < 0\), or \(f(x) = 0\).

Step 2. Find the critical points and the vertical and horizontal asymptotes. Rational functions often involve asymptotes.

Step 3. Plot the critical points and asymptotes.

Step 4. Find intervals with the same signs of \(f'\) and \(f''\) to determine the slopes and curvatures.

**EXAMPLE 5-9.** \(f(x) = x^2 - 2x + a\) with \(a = -3\)

Step 1. \(x^2 - 2x - 3 = (x + 1)(x - 3)\)

<table>
<thead>
<tr>
<th></th>
<th>(x &lt; -1)</th>
<th>(x = -1)</th>
<th>(-1 &lt; x &lt; 3)</th>
<th>(x = 3)</th>
<th>(3 &lt; x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f(x))</td>
<td>(+)</td>
<td>0</td>
<td>(−)</td>
<td>0</td>
<td>(+)</td>
</tr>
</tbody>
</table>

Step 2. \(f'(x) = 2x - 2 = 2(x - 1)\), \(f'(1) = 0\)

Step 3. \(f'(x) = 2(x - 1)\) and \(f''(x) = 2\)

---

3 This is the condition under which the first derivative has a strict extremum at the point. Formally, an inflection point is where the curve changes from being concave to convex or the other way when a tangent exists.
\[ f(x) \begin{array}{ccccccc} x < -1 & x = -1 & -1 < x < 1 & x = 1 & 1 < x < 3 & x = 3 & 3 < x \\ (+) & 0 & (-) & (-) & (0) & (+) & \\
\end{array} \]

\[ f'(x) \begin{array}{ccccccc} (-) & (-) & (-) & 0 & (+) & (+) & (+) \\
\end{array} \]

\[ f''(x) \begin{array}{ccccccc} (+) & (+) & (+) & (+) & (+) & (+) \\
\end{array} \]

5.6 Concave Functions

**Definition 5.6.** A function is **concave** on an interval \([a, b]\) if, for all \(x, y \in [a, b]\) and \(\lambda \in [0,1]\),

\[ f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) \]

**Theorem 5.16.** The sum of concave functions is concave.

And the definition can be extended to the case of many points, that is known as Jensen’s inequality.

**Theorem 5.17 (Jensen’s Inequality)** \(f\) is concave if and only if

\[ f(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \geq \lambda_1 f(x_1) + \lambda_2 f(x_2) + \cdots + \lambda_n f(x_n) \]

for all \(x_1, \cdots, x_n \in [a, b]\) and \(\lambda_1, \cdots, \lambda_n \in [0,1]\) with \(\sum_{i=1}^n \lambda_i = 1\).

**Proof.** Apply the definition repeatedly.

\[ f \left( \lambda_1 x_1 + (1-\lambda_1) \left( \frac{\lambda_2}{1-\lambda_1} x_2 + \cdots + \frac{\lambda_n}{1-\lambda_1} x_n \right) \right) \]

\[ \geq \lambda_1 f(x_1) + (1-\lambda_1) f \left( \frac{\lambda_2}{1-\lambda_1} x_2 + \cdots + \frac{\lambda_n}{1-\lambda_1} x_n \right) \]

\[ \geq \lambda_1 f(x_1) + (1-\lambda_1) \frac{\lambda_2}{1-\lambda_1} f(x_2) + (1-\lambda_1-\lambda_2) f \left( \frac{\lambda_3}{1-\lambda_1-\lambda_2} x_3 + \cdots + \frac{\lambda_n}{1-\lambda_1-\lambda_2} x_n \right) \]

\[ : \]
For the converse, take $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$. If $f(x)$ is twice continuously differentiable, the following proof is more intuitive. Since $f$ concave, $f''(x) \leq 0$ for all $x \in X$. By Taylor’s Theorem, for some $C_i \leq 0$,

$$f(x_i) = f(a) + f'(a)(x_i - a) + C_i$$

Multiply $\lambda_i$ and the sum for all $i$ gives

$$\lambda_1 f(x_1) + \cdots + \lambda_n f(x_n) = f(a) + f'(a)\left((\lambda_1 x_1 + \cdots + \lambda_n x_n) - a\right) + (\lambda_1 C_1 + \cdots + \lambda_n C_n)$$

Substituting $a = (\lambda_1 x_1 + \cdots + \lambda_n x_n)$ gives the result. □

Its continuous version is presented in THEOREM 17-8.

**DEFINITION 5-7.** A function is **strictly concave** (strictly convex) in an interval $[a, b]$ if, for all $x, y \in [a, b]$ with $x \neq y$ and $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda) f(y)$$

Geometrically, the definition says that the graph of the function always lies above the segments connecting two points on the graph. This has several immediate consequences. If there are two distinct local maxima, then so are all the convex combinations of the two.

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y) \geq \min(f(x), f(y))$$

Also, the relationship shows that strict concavity guarantees a unique maximum if exists.

Moreover, if $x^*$ is a local maximum, it is also a global maximum. Otherwise, letting $x^{**}$ be a global maximum, for all $\lambda \in (0,1)$,

$$f(\lambda x^* + (1-\lambda)x^{**}) > f(x^*)$$

and $x^*$ is not a local maximum.

Another way to describe the graph of concave functions is that the graph always lies below its tangents, or the graph is bounded above by the first order approximation.

**THEOREM 5-18.** A function is concave over an interval $[a, b]$ if and only if, for all $c \in [a, b]$,

$$f(x) \leq f(c) + f'(c)(x - c)$$
Proof.

(⇒) From the definition,

\[ f(\lambda x + (1 - \lambda)c) \geq \lambda f(x) + (1 - \lambda)f(c) \]

\[ f(\lambda x + (1 - \lambda)c) \geq f(c) + \lambda(f(x) - f(c)) \]

\[ f(x) - f(c) \leq \frac{f(\lambda x + (1 - \lambda)c) - f(c)}{\lambda} \]

Taking limit of \( \lambda \to 0 \) on both sides gives

\[ f(x) - f(c) \leq \frac{df(\lambda x + (1 - \lambda)c)}{d\lambda} \bigg|_{\lambda=0} \]

\[ f(x) - f(c) \leq f'(\lambda x + (1 - \lambda)c)(x - c) \bigg|_{\lambda=0} = f'(c)(x - c) \]

(⇐) Let \( y = \lambda c + (1 - \lambda)x \).

\[ f(x) - f(y) \leq f'(y)(x - y) \]

\[ f(c) - f(y) \leq f'(y)(c - y) \]

Multiply the first inequality by \((1 - \lambda)\) and the second by \(\lambda\), and add them to get

\[ (1 - \lambda)(f(x) - f(y)) + \lambda(f(c) - f(y)) \leq (1 - \lambda)f'(y)(x - y) + \lambda f'(y)(c - y) \]

\[ \lambda f(c) + (1 - \lambda) f(x) - f(y) \leq (1 - \lambda)f'(y)(x - y) + \lambda f'(y)(c - y) \]

Since \( y = \lambda c + (1 - \lambda)x, (1 - \lambda)(x - y) + \lambda(c - y) = 0 \).

\[ \lambda f(c) + (1 - \lambda) f(x) \leq f(y) = f(\lambda c + (1 - \lambda)x). \]

If a maximum exists, there must be a unique maximum or infinitely many maxima. This property can be viewed as a direct consequence of the following result.

**Theorem 5-19.** Suppose function \( f \) is concave on an interval \([a, b]\). Then all the upper contour sets are either empty or convex.

**Proof.** For \( x, y \in \{ x \mid f(x) \geq c \} \),

Math for Economists, Sogang University
\[ f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) \geq \lambda c + (1 - \lambda)c = c. \]

Therefore, \( \lambda x + (1 - \lambda)y \in \{ x \mid f(x) \geq c \} \).

\[ \square \]

Note that concavity is only sufficient. An example is \( f(x) = e^x \). As stated before, this property is related to the property is known as quasi-concavity.

**Definition 5-8.** A function is **quasi-concave** on an interval \([a, b]\) if, for all \( x, y \in [a, b] \) and \( \lambda \in [0,1] \),

\[ f(\lambda x + (1 - \lambda)y) \geq \min(f(x), f(y)) \]

In words, the upper contour set is convex.

Finally, the following theorem complements Theorem 5-2, the case of zero second derivative.

**Theorem 5-20.** Suppose function \( f \) is twice continuously differentiable on \((a, b)\) . \( f''(x) \leq 0 \) for all \( x \in (a, b) \) if and only if \( f \) is concave on \((a, b)\).

**Proof.** Let \( z = \lambda x + (1 - \lambda)y \) with \( x < y \) from some \( \lambda \in [0,1] \). By the mean value theorem, for some \( c \in (x, z) \),

\[ \frac{f(z) - f(x)}{(z - x)} = f'(c) \]

\[ f(z) - f(x) = (1 - \lambda)f'(c)(y - x) \]

Since \( f''(x) \leq 0 \) or \( f'(x) \) is decreasing,

\[ f(z) - f(x) \geq (1 - \lambda)f'(z)(y - x). \]

And similarly, for some \( d \in (z, x) \)

\[ f(z) - f(y) = -\lambda f'(d)(y - x). \]

Again since \( f'(x) \) is decreasing,

\[ f(z) - f(y) \geq -\lambda f'(z)(y - x). \]

Multiply \( f(z) - f(x) \geq (1 - \lambda)f'(z)(y - x) \) by \( \lambda \) and \( f(z) - f(x') \geq -\lambda f'(z)(y - x) \) by \((1 - \lambda)\), and the sum of the two inequalities gives the result.
Conversely, if \( f \) is concave, then for all \( \lambda \in [0,1] \)

\[
\frac{f(z) - f(x)}{1 - \lambda} \geq \frac{f(y) - f(x)}{\lambda}
\]

As \( \lambda \to 1 \),

\[(y - x)f'(x) \geq (y - x)f'(y),\]

which implies the \( f'(x) \) is not increasing, or \( f''(x) \leq 0 \).

**Theorem 5-21.** Suppose function \( f : \mathcal{X} \to \mathbb{R} \) is twicely continuously differentiable and \( f''(x) \leq 0 \) for all \( x \in \mathcal{X} \). Then a critical point is a global maximum.

**Proof.** The result is immediate from Taylor’s theorem.

\[
f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{1}{2}f''(c)(x - \bar{x})^2 = f(\bar{x}) + \frac{1}{2}f''(c)(x - \bar{x})^2 \leq f(\bar{x})
\]

Note that does not depend on the second derivative at the maximum. Since the condition guarantees that the first derivative or the slope diminishes, the first order condition is sufficient to characterize a unique local maximum or a global maximum.

**Example 5-10.** \( f(x) = x^4 \) is globally concave, and the critical point is a global maximum.

All the argument for the maximum can be applied to the minimum with replacing “concave” with “convex.”

**Definition 5-9.** A function is **convex** in an interval \([a, b]\) if, for all \( x, y \in [a, b] \) and \( \lambda \in [0,1] \),

\[
f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)
\]

**Definition 5-10.** A function is **quasi-convex** on an interval \([a, b]\) if, for all \( x, y \in [a, b] \) and \( \lambda \in [0,1] \),

\[
f(\lambda x + (1 - \lambda)y) \leq \max(f(x), f(y))
\]

\(^4\) If confused, use \( g(\lambda) = f(\lambda x + (1 - \lambda)y) \) so that \( g(0) = f(y) \) and \( g(1) = f(x) \).
The lower contour set is convex.
6 LINEAR ALGEBRA

6.1 EUCLIDEAN SPACE

Roughly, Euclidean space $\mathbb{E}^n$ is the space of all $n$-tuples of real numbers, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. (See DEFINITION 1-25.) Such $n$-tuples are sometimes called points or vectors, and $x_1, x_2, \ldots, x_n$ are called the coordinates of $\mathbf{x}$. Point and vector may seem the same but have distinctly different definitions and interpretations.

There are several types of quantities often used in economics. Scalars are fully described by a magnitude such as a length, amount, and time while vectors have both a magnitude (length) and a direction. For instance, distance and displacement have different information. In economics, we are mostly interested in the direction that is related to the geometric properties of vectors and vector spaces.

6.2 VECTOR

DEFINITION 6-1. A (free) vector is an object that has a magnitude (length) and direction, and a $n$-component (real) vector is denoted by a list of $n$ real numbers,

$$\mathbf{x} = \begin{pmatrix} \vdots \\ x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{where } x_i \in \mathbb{R} \text{ for all } i$$

Matrix notion is often used to denote it by a $n$-component column (or row) matrix,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{x} = (x_1 \ldots x_n) = [x_1 \ldots x_n]$$

As a vector, it does not matter whether we write it vertically or horizontally as long as it is ordered. The distinction is relevant only when they are treated as a matrix for matrix operations.

DEFINITION 6-2. A bound vector is a vector with a definite initial point and terminal point.
A particular initial point is of no importance in free vectors as long as they have the same magnitude and direction. Since most economic analyses are about changes or slopes, the main discussion is limited to free vectors that can be thought of as a position vector that is a bounded vector with a common initial point, the origin. In the following, vector denotes free vector.

**Definition 6-3.** A $n \times 1$ **standard (or canonical) basis vector** $e_i$ is the $n \times 1$ vector with all components zero except the $i^{th}$, which is 1.

When $n = 3$, $e_1 = (1,0,0)$ , $e_2 = (0,1,0)$, and $e_3 = (0,0,1)$. A vector $x = (x_1,x_2,x_3)$ can be written as

$$x = x_1e_1 + x_2e_2 + x_3e_3$$

Two vectors are said to be equal if they have the same magnitude and direction. That is, $x = y$ if $x_1 = y_1$, $x_2 = y_2$, and $x_n = y_n$ where $x = (x_1,\cdots,x_3)$ and $y = (y_1,\cdots,y_n)$.

**Definition 6-4.** The **magnitude** or **length** or **norm** of a vector is denoted by $\|x\|$ and defined as Euclidean norm,

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

**Definition 6-5.** A **unit vector** is a vector with a magnitude of one and often denoted by with a hat, $\hat{x}$. A **zero vector** is the vector with the length of zero, $0 = (0,\cdots,0)$.

Any non-zero vectors can be normalized to a unit vector.

$$\hat{x} = \frac{x}{\|x\|} = \frac{x_1}{\|x\|}e_1 + \frac{x_2}{\|x\|}e_2 + \cdots + \frac{x_n}{\|x\|}e_n$$

A zero vector has an arbitrary or indeterminate direction.

**Definition 6-6 (Vector Algebra using Coordinate)** For $x, y, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$,

(i) **scalar multiplication**: $\lambda x = (\lambda x_1, \lambda x_2, \cdots, \lambda x_n)$

(ii) **addition**: $x + y = (x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n)$
6.3 Dot Product

**Definition 6-7 (Dot Product, Scalar Product)**

\[ \mathbf{x} \cdot \mathbf{y} = x_1y_1 + \cdots + x_ny_n \]

**Theorem 6-1 (Properties)**

(i) commutative rule: \( \mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x} \)

(ii) distributive rule: \( \mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z} \), \( (\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + 2(\mathbf{x} \cdot \mathbf{y}) + \mathbf{y} \cdot \mathbf{y} \)

(iii) bilinear \( \mathbf{x} \cdot (\lambda \mathbf{y}) = \lambda (\mathbf{x} \cdot \mathbf{y}) = (\lambda \mathbf{x}) \cdot \mathbf{y} \)

Note that dot product is not associative and a dot product gives a real number.

The length of a vector is calculated by the Euclidean norm and it is the same as the square root of the dot product of a vector.

\[ \|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}, \quad \|\mathbf{x} - \mathbf{y}\| = \sqrt{(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})} \]

The dot product has an interesting geometric interpretation. The Law of Cosines states that

\[ c^2 = a^2 + b^2 - 2ab \cos \theta \]

where \( a, b, c \) are the lengths of each side of a triangle, and \( \theta \) is the angle between the sides corresponding to \( a, b \).

\[ (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) = \mathbf{x} \cdot \mathbf{x} + \mathbf{y} \cdot \mathbf{y} - 2\|\mathbf{x}\| \cdot \|\mathbf{y}\| \cos \theta \]

\[ \mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cos \theta \]

Because of this property, we can use to test the orthogonality of vectors. Since \( \cos(\pi/2) = 0 \),

---

5 Some people do not distinguish dot product and inner product, but they are not the same. Dot product, \( \mathbf{x} \cdot \mathbf{y} \), is defined only for Euclidean space while inner product, \( \langle \mathbf{x}, \mathbf{y} \rangle \), is defined even for infinite dimensional spaces and it needs not be a positive definite in general spaces. In Euclidean space, they are identical.
\[ \mathbf{x} \perp \mathbf{y} \iff \mathbf{x} \cdot \mathbf{y} = 0 \]

**Theorem 6-2.** The angle between two vectors \( \mathbf{x} \) and \( \mathbf{y} \) in \( \mathbb{R}^n \) is

\[
\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|}
\]

**Theorem 6-3.** Let \( \mathbf{x} \) and \( \mathbf{y} \) be \( n \)-component real vectors. Then

(i) \( \|\mathbf{x}\| + \|\mathbf{y}\| \geq \|\mathbf{x} - \mathbf{y}\| \) (triangle inequality)

(ii) \( \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 \) if and only if \( \mathbf{x} \cdot \mathbf{y} = 0 \) (Pythagoras theorem)

(iii) \( \|\mathbf{x}\| \cdot \|\mathbf{y}\| \geq \|\mathbf{x} \cdot \mathbf{y}\| \) (Cauchy-Schwartz inequality)

**Proof.**

(i),(ii) \( \|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\| \cdot \|\mathbf{y}\| \cos \theta \leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\| \cdot \|\mathbf{y}\| = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2 \)

Note that \( \|\mathbf{x}\| + \|\mathbf{y}\| \geq \|\mathbf{x} + \mathbf{y}\| \) holds too.

(iii) \( \mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \times \cos \theta \), \( |\mathbf{x} \cdot \mathbf{y}| = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \times |\cos \theta| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\| \)  ■

## 7 Vector Space

### 7.1 Vector Space

**Definition 7-1.** A non-empty set \( V \) of \( n \)-component vectors is a **vector space** if it is closed under addition and scalar multiplication (that is, \( \mathbf{u} + \mathbf{v} \) and \( \lambda \mathbf{u} \) are in \( V \) when \( \mathbf{u} \) and \( \mathbf{v} \) are in \( V \).) \( \mathbb{R}^n \) denotes the set of all \( n \)-component vectors.

**Example 7-1.** A line and a plane passing the origin are vector space. By definition, a vector space should contain the origin, \( \mathbf{0} \). For instance, \( X = \{(x_1, x_2) \in \mathbb{R}^2 | x_1 + 2x_2 = 0 \} \) is a vector space because, for any \( \mathbf{x}, \mathbf{y} \in X \), \( a\mathbf{x} + b\mathbf{y} \in X \),

\[ a\mathbf{x} = (ax_1, ax_2), \quad (ax_1) + 2(ax_2) = a(x_1 + 2x_2) = 0 \]

and
\[
x + y = (x_1 + y_1, x_2 + y_2), \quad (x_1 + y_1) + 2(x_2 + y_2) = (x_1 + 2x_2) + (y_1 + 2y_2) = 0
\]

**Definition 7-2.** A **subspace** of a vector space \( V \) is a subset of \( V \) which is itself a vector space.

**Definition 7-3.** A set of vectors \( x_1, \ldots, x_m \) in a vector space, \( x \) is said to **span** \( V \) if and only if every vector in \( V \) can be written as a linear combination of \( x_1, \ldots, x_m \).

**Example 7-2.** Let \( X = \{(1,1)\} \subset \mathbb{R}^2 \). The span of \( X \) is the collection of all vectors that has a form \( \alpha(1,1) = (\alpha, \alpha) \) with \( \alpha \in \mathbb{R} \). That is a 45-degree line passing the origin.

### 7.2 Linear Independence

**Definition 7-4.** A set of \( n \)-component vectors \( X = \{x_1, \ldots, x_m\} \) is **linearly dependent** if and only if there exists scalars \( \lambda_1, \ldots, \lambda_m \) not all 0 such that \( \lambda_1 x_1 + \cdots + \lambda_m x_m = 0 \) (that is, such that one of the \( x_i \) can be expressed as a linear combination of the rest.) Otherwise, the \( X \) is **linearly independent**. (So, if any \( x = 0 \), the set is linearly dependent.)

**Theorem 7-1.** If \( X = (x_1, \ldots, x_m) \) is a collection of linearly independent vectors and \( z \in Z \) where \( Z \) is the set of all linear combinations of \( (x_1, \ldots, x_m) \), then there are unique \( (\lambda_1, \ldots, \lambda_m) \) such that \( z = \lambda_1 x_1 + \cdots + \lambda_m x_m \).

**Proof.** Suppose there are two so that
\[
z = \lambda_1 x_1 + \cdots + \lambda_m x_m \quad \text{and} \quad z = \eta_1 x_1 + \cdots + \eta_m x_m
\]
Subtract one from the other to get
\[
0 = (\lambda_1 - \eta_1)x_1 + \cdots + (\lambda_m - \eta_m)x_m \tag*{\blacksquare}
\]

**Definition 7-5.** A **basis** for a vector space \( V \) is a subset of vectors in \( V \) which is linearly independent and which spans \( V \). The trivial space \( \{0\} \) has no basis.

**Example 7-3.** Consider a vector space \( V \) spanned by \( X = \{(1,2,1), (-1,1,1)\} \). Since those two vectors are linearly independent, \( X \) itself is a basis of \( V \).

**Definition 7-6.** The maximum number of linearly independent vectors in a vector space is called the **dimension** of \( V \), written \( \dim(V) \).
**Theorem 7-2.** If $V$ is an $n$-dimensional vector space, then any set of $n + 1$ vectors cannot be linearly independent.

**Proof.** Let $X = \{x_1, \ldots, x_n\}$ be a basis of $V$. To reach a contradiction, suppose that $\{v_1, \ldots, v_{n+1}\}$ is linearly independent. Since $X$ is a basis, we can write

$$v_1 = \lambda_{11}x_1 + \cdots + \lambda_{1n}x_n$$

Moreover, since $v_1 \neq 0$, there exists $\lambda_1 \neq 0$. Without loss of generality, suppose $\lambda_{11} \neq 0$. Then

$$x_1 = \frac{1}{\lambda_{11}}(v_1 - \lambda_{12}x_2 + \cdots + \lambda_{1n}x_n)$$

That is, $\{v_1, x_2, \ldots, x_n\}$ spans $V$. Repeating the process $n$ times shows that $\{v_1, \ldots, v_n\}$ spans $V$. That is, $v_{n+1}$ is a linear combination of $\{v_1, \ldots, v_n\}$, which is a contradiction.

The theorem states that if a vector space $V$ is spanned by $n$ vectors, then every linearly independent subset of $V$ has at most $n$ vectors.

**Theorem 7-3.** If $V$ is an $n$-dimensional vector space, then any set of $n$ linearly independent vectors in $V$ is a basis for $V$.

**Proof.** We need to show that any vector in $V$ can be expressed as a linear combination of the linearly independent vectors. Let $X = \{x_1, \ldots, x_n\}$. By Theorem 7-2, for any $v \in V$, $\{v, x_1, \ldots, x_n\}$ is linearly dependent. That is, there is non-zero $(\lambda_0, \lambda_1, \ldots, \lambda_n)$ such that

$$\lambda_0 v + \lambda_1 x_1 + \cdots + \lambda_n x_n = 0$$

Since $\lambda_1 x_1 + \cdots + \lambda_n x_n = 0$ only when $(\lambda_1, \ldots, \lambda_n) = 0$, we must have $\lambda_0 \neq 0$. Therefore,

$$v = -\frac{1}{\lambda_0}(\lambda_1 x_1 + \cdots + \lambda_n x_n)$$

This is the definition of basis.

**Definition 7-7.** A set of vectors $x_1, \ldots, x_m$ is **orthogonal** if $x_i \cdot x_j = 0$ for all $i \neq j$. The vectors are **orthonormal** if, in addition, each is a unit vector.

**Example 7-4.** In Example 7-3, an orthogonal basis for $V$ is $\{(1,2,1), (-2,1,1)\}$ and an orthonormal basis is $\{(1/\sqrt{6},2/\sqrt{6},1/\sqrt{6}), (-2/\sqrt{6},1/\sqrt{6},1/\sqrt{6})\}$. 

---

Math for Economists, Sogang University
7.3 LINEAR MAP

In this section, we extend the domain of a function to general vector space.

**DEFINITION 7-8.** A function \( T : \mathbb{R}^n \to \mathbb{R}^m \) is said to be a **linear map** if the following two conditions are satisfied:

- **Additivity (addition):** \( T(x + y) = T(x) + T(y) \) for any \( x, y \in \mathbb{R}^n \)
- **Homogeneity of degree one (scalar multiplication):** \( T(cx) = cT(x) \) for any \( c \in \mathbb{R} \)

Often people write a linear function with upper case letters without parenthesis for the argument such as \( Tx \) when no confusion arises.

Notice that if the domain is a vector space, the linear map preserves the properties so that the range is also a vector space. As pointed in Section 3.2.1, a function whose graph is a straight line might not be linear. The graph should pass the origin.

The range of a linear map is the span of the image of the standard basis.

**THEOREM 7-4.** If \( T : \mathbb{R}^n \to \mathbb{R}^m \) is a linear map, then

\[
T(\mathbb{R}^n) = \text{span}(X), \quad \text{where } X = \{Te_1, \cdots, Te_n\}
\]

**Proof.** Since every \( x \in \mathbb{R}^n \) is a linear combination of standard basis, there exists \( \lambda_i \)'s such that \( x = \sum_{i=1}^{n} \lambda_i e_i \). Therefore, the range is the set of all points of the form

\[
Tx = T \left( \sum_{i=1}^{n} \lambda_i e_i \right) = \sum_{i=1}^{n} \lambda_i Te_i
\]

This is the definition of \( \text{span}(X) \).

The linear map could be used to characterize a different kind of vector space.

**DEFINITION 7-9.** Given a linear map \( T : \mathbb{R}^n \to \mathbb{R}^m \), the null space or kernel of \( T \) is

\( \ker(T) = \{ x \in \mathbb{R}^n | Tx = 0 \} \)
ker$(T)$ is the set of solutions to $Tx = 0$, and it is also a vector space.

**Theorem 7-5.** If $T: \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, then

(i) ker$(T)$ is a linear subspace of $\mathbb{R}^n$.

(ii) ker$(T) = \{0\}$ if and only if $T$ is injective.

**Proof.** (i) ker$(T)$ is closed under scalar multiplication because $T$ is linear. It is also closed under addition because of $Tx + Ty = 0$ for any $x, y \in \mathbb{R}^n$.

(ii) Suppose that $Tx = Ty$ for some $x, y \in \mathbb{R}^n$. Then, by linearity, $x - y \in \text{ker} T$. Since ker$(T) = \{0\}$, then it must have $x = y$. The converse is obvious because there is a unique $x$ such that $Tx = 0$. $\blacksquare$

If a linear map is from a set to itself, we have a stronger result.

**Theorem 7-6.** If $T: \mathbb{R}^n \to \mathbb{R}^n$ is a linear map, then the followings are equivalent.

(i) $T$ is a bijection.

(ii) $T$ is surjective.

(iii) $T$ is injective.

(iv) ker$(T) = \{0\}$.

(v) The set $X = \{Te_1, \ldots, Te_n\}$ is linearly independent.

The proof is in Section 8.4.

### 7.4 Flats

There are important linear objects in Euclidean space, so-called flats.

**Definition 7-10.** A flat is a subset of a Euclidean space that is congruent to a Euclidean space of lower dimension.
Flats are affine subspaces of Euclidean spaces, which means that they are similar to linear subspaces, except that they need not pass through the origin. A line is a flat of \( \mathbb{R}^2 \), and lines and planes are flats in \( \mathbb{R}^n \), \( n \geq 3 \).

### 7.4.1 Line

In \( \mathbb{R}^2 \) space, a line is either determined by two distinct points or a point and a direction (slope) with a general (standard) form of \( ax_1 + bx_2 + c = 0 \). In \( \mathbb{R}^n \) space with \( n \geq 3 \), a line cannot be represented by one equation as we usually do because a single equation typically has an \( n - 1 \) dimensional solution set while a line is an one dimensional object. With vectors as a slope, one can extend the concepts in \( n \)-dimensional space easily.

**Definition 7-11.** The points on a line that passes a point \( \mathbf{x}_0 \) with a direction, \( \mathbf{v} \) can be represented by the form \( \mathbf{x}_0 + t\mathbf{v} \) where \( t \) is a real number. This is the **parametric equation** for the line.

\[
\mathbf{v}: \text{a direction} \\
\mathbf{x}_0: \text{a point on the line}
\]

Given two distinct points, \( \mathbf{x}_0 \) and \( \mathbf{x}_1 \), the line through the points has the direction of \( \mathbf{v} = \mathbf{x}_0 - \mathbf{x}_1 \) and a parametric equation of \( \mathbf{x}_0 + t\mathbf{v} \). The line is the set, \{ \( \mathbf{z} \in \mathbb{R}^n \mid \mathbf{z} = \mathbf{x}_0 + t\mathbf{v} \) for all \( t \in \mathbb{R} \} \).

Instead of a vector representation, a line can be described by an affine combination such that

\[
\{ t\mathbf{x}_0 + (1 - t)\mathbf{x}_1 \mid t \in \mathbb{R} \}.
\]

**Example 7-5.** Let’s find an equation of a line passing \( \mathbf{a} = (1,2,0) \) and \( \mathbf{b} = (1,0,-1) \). The direction of the line is \( (\mathbf{a} - \mathbf{b}) \) or \( [0 \ 2 \ -1] \), and the parametric equation is \( (1,2,0) + t[0 \ 2 \ -1] \). One can rewrite it as a Cartesian equation if preferred.

\[
\begin{align*}
x_1 &= 1 + t \times 0, & x_2 &= 2 + t \times 2, & x_3 &= 0 + t \times 1 \\
&\quad x_1 = 1, & x_2 &= 2(x_3 + 1)
\end{align*}
\]
7.4.2 Plane

A plane is a two-dimensional flat surface and requires three pieces of information to define.

**Definition 7-12.** Given a vector \( \mathbf{x} \in \mathbb{R}^n \) and two linearly independent vectors \( \mathbf{v} \) and \( \mathbf{w} \) in \( \mathbb{R}^n \) (nonzero vectors \( \mathbf{v} \) and \( \mathbf{w} \) are linearly independent if there is no \( \lambda \) such that \( \mathbf{v} = \lambda \mathbf{w} \)), a **plane** is the set of all points,

\[
\mathbf{z} = \mathbf{x} + s\mathbf{v} + t\mathbf{w}, \quad s, t \in \mathbb{R}
\]

7.4.3 Hyperplane

The hyperplane is a flat of dimension 1 less than the ambient space or co-dimension one, and it separates a Euclidean space into two half-spaces. A general equation of a hyperplane in a \( n \)-dimensional space is

\[
a_1x_1 + \cdots + a_nx_n = b
\]

As we have seen before, we need two pieces of information to draw a line, and three for a plane. Since the equation involves \((n - 1)\) parameters, we need \( n \) independent information. In \( \mathbb{R}^n \), since a single linear equation has \((n - 1)\) dimensional solution, only a single equation is sufficient to describe a hyperplane.

Consider the equation \( \mathbf{a} \cdot \mathbf{x} \) where \( \mathbf{a} \) is a non-zero vector. In \( \mathbb{R}^2 \), the equation represents a line through the origin. In \( \mathbb{R}^3 \), the equation represents a plane through the origin. In both cases, the line or the plane is orthogonal to \( \mathbf{a} \) or \( \mathbf{a} \) is said to be **normal** to the line or the plane.

A hyperplane can also be described by a **normal equation** with a point on hyperplane and a direction perpendicular to the plane. Subtracting \( \mathbf{a} \cdot \mathbf{x}_0 = b \) from \( \mathbf{a} \cdot \mathbf{x} = b \) yields \( \mathbf{a} \cdot (\mathbf{x} - \mathbf{x}_0) = 0 \), and thus the coefficient vector is orthogonal to the vectors on the surface. Consequently the equation of a hyperplane with point \( \mathbf{x}_0 \) and normal or perpendicular direction \( \eta \neq 0 \) is \( \eta \cdot (\mathbf{x} - \mathbf{x}_0) = 0 \).

**Definition 7-13.** Given a vector \( \eta \in \mathbb{R}^n - \{0\} \) and a point \( \mathbf{x}_0 \in \mathbb{R}^n \), a **hyperplane** through the point \( \mathbf{x}_0 \) with normal \( \eta \) is the set of all points \( \mathbf{x} \) such that \( \eta \cdot (\mathbf{x} - \mathbf{x}_0) = 0 \).

The definition implies that a hyperplane consists of all of the \( \mathbf{x} \) such that the direction \( (\mathbf{x} - \mathbf{x}_0) \) is normal to \( \eta \). In \( \mathbb{R}^2 \), lines are hyperplanes. In \( \mathbb{R}^3 \), hyperplanes are “ordinary” planes.
**EXAMPLE 7-6.** A line in \( \mathbb{R}^2 \) has a general form of \( ax + by + c = 0 \) or

\[
\begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} = 0,
\text{ where } ax_0 + by_0 + c = 0
\]

Note that a line \( ax + by = 0 \) is the set of points that is orthogonal to \((a, b)\).

The direction of the line in **EXAMPLE 7-5** is \((0, 2, 1)\) and \((c, -1, 2)\) is orthogonal to the vector. Therefore, any vector that is on the line should be perpendicular to \((c, -1, 2)\) for all \( c \in \mathbb{R} \).

\[
\begin{pmatrix} c \\ -1 \\ 2 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \\ x_3 \end{pmatrix} = 0 \text{ or } c(x_1 - 1) - (x_2 - 2) + 2x_3 = 0
\]

Lines, planes, and hyperplanes are flats of \( \mathbb{R}^n \), and there are other flats of any dimension less than \( n \). Note that, although they have the same “shape” as Euclidean space, in general, they are not subspaces as they might not contain the origin.

**DEFINITION 7-14.** A linear manifold of \( \mathbb{R}^n \) is a set \( S \) such that there is a subspace \( V \) on \( \mathbb{R}^n \) and \( x_0 \in \mathbb{R}^n \) with \( S \subset V + \{x_0\} = \{x | x = v + x_0, \ v \in V\} \).

A linear manifold is an affine subspace of \( \mathbb{R}^n \).

### 7.5 Matrix Arithmetic

A matrix is a collection of numbers and is an efficient way to stack many numbers.

**DEFINITION 7-15.** An \( m \times n \) matrix is an rectangular array of numbers with \( m \) rows and \( n \) columns. If the element (entry, component, coefficient) in the \( i^{th} \) row and the \( j^{th} \) column is denoted by \( a_{ij} \), then \( A \) is often written as \([a_{ij}]\), and to be read "matrix whose \( i,j^{th} \) element is \( a_{ij} \). The \( i^{th} \) row of \( A \) is denoted by \( \text{row}_i A \) and the \( j^{th} \) column of \( A \) is by \( \text{col}_j A \).

\[
A = \begin{bmatrix}
\vdots & \cdots & \vdots \\
a_{11} & \cdots & a_{1n} \\
\vdots & \colon & \vdots \\
a_{m1} & \cdots & a_{mn}
\end{bmatrix}
\]

**DEFINITION 7-16.** Two \( m \times n \) matrices \( A = [a_{ij}] \) and \( B = [b_{ij}] \) are equal, written \( A = B \), if and only if \( a_{ij} = b_{ij} \) for all \( i \) and \( j \).

**DEFINITION 7-17 (Basic Operations)**
(i) **Scalar Multiplication:** For a constant $\lambda$, $\lambda A$ is an $n \times m$ matrix with elements $\lambda a_{ij}$.

(ii) **Addition:** If $A$ and $B$ are both $n \times m$ matrices, then so is $C = A + B$, whose entry is $c_{ij} = a_{ij} + b_{ij}$.

(iii) **Matrix Multiplication:** If $A$ is an $n \times k$ matrix and $B$ is a $k \times m$ matrix, then $M = A \times B$ is a $n \times m$ matrix and the element in row $i$ and column $j$ of $A \times B$ is the inner product of row $i$ of $A$ with column $j$ of $B$.

\[
c_{ij} = \text{row}_i A \cdot \text{col}_j B = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{h=1}^{k} a_{ih}b_{hj}
\]

Matrix multiplication is not defined for general matrices. The matrices should satisfy the condition on the sizes. In general, $A \times B \neq B \times A$.

**Example 7-7.** The left multiplication of $x$ to $A$ is the first row of $A$, and to $y$ is the sum of rows.

\[
x = [1 \ 0], \quad y = [1 \ 1], \quad A = [\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}]
\]

\[
xA = [1 \ 2], \quad yA = [4 \ 6]
\]

**Definition 7-18.** Let $A = [a_{ij}]$ be an $m \times n$ matrix.

(i) $A$ is a **square matrix** if $m = n$.

(ii) The **zero matrix**, written $0$, is the matrix of all zero elements.

(iii) The $n \times n$ **identity matrix** $I$ (or $I_n$) is defined by $I = [\delta_{ij}]$, where $\delta_{ij}$ is a Kronecker delta with properties that $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Clearly, a zero matrix is the additive identity and an identity matrix corresponds to a multiplicative identity.

**Definition 7-19.** For an $n \times n$ matrix $A$, the **principal diagonal** is the $n$ elements of $a_{ii}$, $i = 1, \ldots, n$. $A$ is a **diagonal matrix** if $a_{ij} = 0$ for $i \neq j$.

The power of a square matrix is written $A^r = A \times \cdots \times A$. Then $A$ is the $r^{th}$ root of $A^r$. With diagonal matrices, it is straightforward to calculate products, powers, roots, and inverses.
**Example 7-8.**

\[
A \times B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \times \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{bmatrix}
\]

\[
A^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1^2 & 0 & 0 \\ 0 & 2^2 & 0 \\ 0 & 0 & 3^2 \end{bmatrix}
\]

If the diagonal elements are nonnegative,

\[
A^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1^2 & 0 & 0 \\ 0 & 2^2 & 0 \\ 0 & 0 & 3^2 \end{bmatrix}
\]

If the diagonal elements are nonzero,

\[
A^{-1} = \begin{bmatrix} 1^{-1} & 0 & 0 \\ 0 & 2^{-1} & 0 \\ 0 & 0 & 3^{-1} \end{bmatrix}
\]

**Definition 7-20.** A square matrix \( A \) is upper (lower) triangular if and only if \( a_{ij} = 0 \) for \( i > j \) (\( i < j \)), and they are also called a right (left) triangular matrix.

In many occasions, a triangular matrix is a good alternative of a diagonal matrix. When we find the determinant or the eigenvalues, it requires the same amount of calculation as one with a diagonal matrix.

Like diagonal matrices, the product of triangular matrices is triangular. Suppose \( A \) and \( B \) are upper triangular and let \( C = AB \).

\[
c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}
\]

Since \( a_{ik} = 0 \) if \( i > k \) and \( b_{kj} = 0 \) if \( j < k \), \( c_{ij} \) could be non-zero only when \( i \leq k \leq j \) or \( i \leq j \). That is, \( C \) is upper triangular.

**Definition 7-21.** The transpose \( A^T \) or \( A' \) of \( A \) is \( A^T = [a_{ji}] \) or row\(_{i} A^T = \text{col}_{i} A \).

**Example 7-9.** Transpose is used to define a matrix multiplication as well as the dot product of vectors.
\[ A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \end{bmatrix} \quad \text{and} \quad AA^T = \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix} = \begin{bmatrix} 6 & -1 \\ -1 & 2 \end{bmatrix} \]

In case of a vector, we treat it as a matrix, typically a column vector.

\[ \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{x}^T \mathbf{x} = (x_1 \cdots x_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = x_1^2 + \cdots + x_n^2 = \mathbf{x} \cdot \mathbf{x} \]

**EXAMPLE 7-10.** A square matrix \( A \) is **symmetric** if \( A = A^T \).

Note that \( AA^T \) and \( A^T A \) are always well-defined and symmetric.

**DEFINITION 7-22.** Given symmetric matrices \( A \) and \( B \), \( B \) is called the **inverse** of \( A \) if \( AB = I \) and \( BA = I \). When \( A \) has an inverse, it is said to be **invertible** and typically written \( A^{-1} \).

An inverse matrix corresponds to the multiplicative inverse. If a matrix is invertible, it is said to be **nonsingular**; otherwise, the matrix is **singular**.

\[ \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

**DEFINITION 7-23.** A square matrix \( A \) is **idempotent** if \( A^2 = A \).

\[ \begin{bmatrix} 2 & -2 & -4 \\ -1 & 3 & 4 \\ 1 & -2 & -3 \end{bmatrix} \begin{bmatrix} 2 & -2 & -4 \\ -1 & 3 & 4 \\ 1 & -2 & -3 \end{bmatrix} = \begin{bmatrix} 2 & -2 & -4 \\ -1 & 3 & 4 \\ 1 & -2 & -3 \end{bmatrix} \]

There are two real numbers that are idempotent, but there are many in matrix form. Idempotent matrices play an important role in projection.

With a unique exception of the identity matrix, an idempotent matrix is singular. If not, we have

\[ A = IA = A^{-1}AA = A^{-1}A = I \]

that is a contradiction.

**DEFINITION 7-24.** A square matrix \( A \) is **orthogonal** if \( A^T A = AA^T = I \). \( A \) is invertible and \( A^{-1} = A^T \).
Note that the definition implies \( \text{row}_i A^T \cdot \text{col}_i A = 1 \) and \( \text{row}_i A^T \cdot \text{col}_j A = 0 \). Since \( \text{row}_i A^T = \text{col}_i A \), the rows and the columns of an orthogonal matrix are orthonormal vectors. As it is square, the columns or the rows constitute a basis for \( E^n \).

**THEOREM 7-7 (Basic Rules)**

(i) \( (\alpha + \beta)A = \alpha A + \beta A \)

(ii) \( \alpha(A + B) = \alpha A + \alpha B \)

(iii) \( A - A = 0 \)

(iv) associative rule: \( (A + B) + C = A + (B + C) \), \( (AB)C = A(BC) \)

(v) commutative rule: \( A + B = B + A \) (in general, \( AB \neq BA \))

(vi) distributive rule: \( A(B + C) = AC + AC \), \( (A + B)C = AC + BC \)

(vii) \( (A + B)^T = A^T + B^T \)

(viii) \( (AB)^T = B^T A^T \)

(ix) If either \( A \) or \( B \) is singular, \( AB \) and \( BA \) are also singular.

(x) Given \( n \times n \) matrices \( A \) and \( B \), \( AB = I \) if and only if \( BA = I \).

(xi) \( A^{-1} \) is unique if exists.

(xii) If \( A \) is symmetric and \( A^{-1} \) exists, then \( A^{-1} \) is symmetric.

(xiii) If \( A^{-1} \) and \( B^{-1} \) exist, then \( (AB)^{-1} = B^{-1}A^{-1} \).

(xiv) If \( A^{-1} \) exists, then \( (A^T)^{-1} = (A^{-1})^T \).

**Proof.** (i)-(vii) Find the left and right sides with \( A = [a_{ij}] \), \( B = [b_{ij}] \), \( C = [c_{ij}] \) and compare them as in the proof of (viii) below.

(viii) Let \( C = AB \) and \( D = (AB)^T \). \( c_{ij} = \text{row}_i A \times \text{col}_j B \) and \( d_{ij} = c_{ji} = \text{row}_i A \times \text{col}_j B \). Since \( \text{row}_i A = (\text{col}_i A^T)^T \), \( d_{ij} = c_{ji} = (\text{col}_j A^T)^T \times (\text{row}_i B^T)^T = \text{row}_i B^T \times \text{col}_j A^T \).

(ix) Suppose not so that \( (AB)^{-1} \) exists. let \( C = B(AB)^{-1} \) and \( D = (AB)^{-1} A \). Then
That contradicts to the singularity of \( A \).

(x) By (ix), \( A \) and \( B \) are invertible. \( AB = I \Rightarrow A^{-1}(AB)A = I \Rightarrow (A^{-1}A)BA = I \Rightarrow BA = I \).

(xi) Suppose not. Then there are two inverse matrices \( B \) and \( C \) such that \( BA = AB = I \) and \( CA = AC = I \).

(xii) \( AA^{-1} = (AA^{-1})^T = (A^{-1})^T A = (A^{-1})^TA = I \) and \( A^{-1}A = (A^{-1}A)^T = A^T(A^{-1})^T = A(A^{-1})^T = I \). Therefore, \( (A^{-1})^T \) is the inverse of \( A \). By (xi), \( A^{-1} = (A^{-1})^T \).

\[
B = BI = B(AC) = (BA)C = IC = C
\]

(xiii) Suppose that \( A \) and \( B \) are \( n \times n \) matrices.

\[
(AB)^{-1}A^{-1} = A(BB^{-1})A^{-1} = AA^{-1} = I \quad \text{and} \quad B^{-1}A^{-1}(AB) = B^{-1}(A^{-1}A)B = B^{-1}B = I
\]

On the other hand, let \( C = B(AB)^{-1} \) and \( D = (AB)^{-1}A \). Then

\[
AC = A(B(AB)^{-1}) = (AB)(AB)^{-1} = I \quad \text{and} \quad DB = ((AB)^{-1}A)B = (AB)^{-1}(AB) = I
\]

Therefore, \( (AB)^{-1} = B^{-1}A^{-1} \) by (ix).

(xiv) \( AA^{-1} = I \Leftrightarrow (AA^{-1})^T = I \Leftrightarrow (A^{-1})^TA = I \) and

\[
A^{-1}A = I \Leftrightarrow (A^{-1}A)^T = I \Leftrightarrow A^T(A^{-1})^T = I
\]

**DEFINITION 7-25.** The **trace** \( \text{tr}(A) \) of an \( n \times n \) matrix \( A \) is \( \text{tr}(A) = a_{11} + \cdots + a_{nn} \).

The following properties are direct from the definition.

**THEOREM 7-8.**

i) \( \text{tr}(A + B) = \text{tr}(A) + \text{tr}(A) \)

ii) \( \text{tr}(A^T) = \text{tr}(A) \)

iii) \( \text{tr}(cA) = c \times \text{tr}(A) \)

ii) \( \text{tr}(AB) = \text{tr}(BA) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ji} \) for \( A_{m \times n} \) and \( B_{n \times m} \).
8 SYSTEMS OF LINEAR EQUATIONS

Consider the following linear equation system.

\[
\begin{align*}
    a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
    a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{align*}
\]

Solving the system is to find the values of \(x_1, \ldots, x_n\) that satisfy every individual equation. In general, we are interested in

(i) the existence of the solution,

(ii) the number of solutions, and

(iii) efficient algorithms to compute the solution.

Note that a linear equation system can be rewritten in a matrix form such as

\[
\begin{bmatrix}
    a_{11} & \cdots & a_{1n} \\
    \vdots & \ddots & \vdots \\
    a_{m1} & \cdots & a_{mn}
\end{bmatrix}
\begin{bmatrix}
    x_1 \\
    \vdots \\
    x_n
\end{bmatrix}
= \begin{bmatrix}
    b_1 \\
    \vdots \\
    b_m
\end{bmatrix}, \quad \text{Ax} = \mathbf{b}
\]

In this way, a \(m \times n\) matrix \(A\) can be thought of as describing a function from \(\mathbb{R}^n\) to \(\mathbb{R}^m\), or specifically, \(f(\mathbf{x}) = A\mathbf{x}\). That is a system with \(n\) unknowns and \(m\) equations, \(f(x_1, \ldots, x_n) = (y_1, \ldots, y_m)\) and \(y_j = f_j(x_1, \ldots, x_n)\) where \(f_j: \mathbb{R}^n \to \mathbb{R}\).

**THEOREM 8-1.** Let \(A\) be \(m = n\), and consider the linear equation system \(Ax = \mathbf{b}\). If \(m = n\) and \(A\) is nonsingular, the unique solution is \(\mathbf{x} = A^{-1}\mathbf{b}\). Otherwise, there is no solution or infinitely many solutions.

**DEFINITION 8-1.** Given a linear equation system \(Ax = \mathbf{b}\), \(A\) and \(\mathbf{b}\) are called **coefficient matrix** and **constant matrix**, respectively. An **augmented matrix** is a matrix obtained by appending the columns of the coefficient matrix and constant vector, \((A|\mathbf{b})\).
8.1 GAUSS-JORDAN ELIMINATION METHOD

Note that if a system is derived from the original system by elementary equation operations such as adding a multiple of one equation to another and interchanging equations, then both systems have the same solutions. If a system of equations is transformed in a way that each equation has fewer variables than the previous one, the system can be solved by back substitution.

\[
\begin{bmatrix}
\alpha_{11} & \alpha_{12} & \alpha_{13} \\
0 & \alpha_{22} & \alpha_{23} \\
0 & 0 & \alpha_{33}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= 
\begin{bmatrix}
c_1 \\
c_2 \\
c_3
\end{bmatrix}
\]

\[
\begin{bmatrix}
\alpha_{11} & \alpha_{12} & \alpha_{13} \\
0 & 0 & \alpha_{23} \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= 
\begin{bmatrix}
c_1 \\
c_2 \\
c_3
\end{bmatrix}
\]

Such a solution method is called *Gaussian elimination method*.

Using all the three elementary equation operations including multiplying an equation by a nonzero scalar, we can have a simpler form such as

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= 
\begin{bmatrix}
d_1 \\
d_2 \\
d_3
\end{bmatrix}
\]

\[
\begin{bmatrix}
1 & \gamma_{12} & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
= 
\begin{bmatrix}
d_1 \\
d_2 \\
d_3
\end{bmatrix}
\]

Such a solution method is called *Gauss-Jordan elimination method*.

**DEFINITION 8-2 (Elementary Row Operation)**

(i) Multiplying the \(i^{th}\) row by a nonzero scalar \(c\).

(ii) Interchanging the \(i^{th}\) and \(j^{th}\) rows.

(iii) Adding the \(i^{th}\) row to \(j^{th}\) row.

**DEFINITION 8-3.** A row of a matrix is said to have \(k\) leading zeros if the first \(k\) elements of the row are all zeros and the \((k + 1)\)th element of the row is not zero.

**DEFINITION 8-4.** A matrix is in row Echelon form if each row has more leading zeros than the preceding row.

**DEFINITION 8-5.** The first nonzero entry in each row of a matrix in row echelon form is called a pivot.

**DEFINITION 8-6.** A reduced echelon matrix is a reduced row Echelon (or Gauss-Jordan) form such that
(i) Every pivot is 1, called the **leading 1**.

(ii) Every column containing a pivot contains no nonzero entries.

(iii) All nonzero rows, if any, precede all zero rows.

**Definition 8-7.** In the context of matrices with \( m \) rows, the **elementary row operation matrices**, or **elementary matrices**, are:

(i) \( E_M(i, c) \) matrix gotten from \( I_m \) by multiplying its \( i^{th} \) row by a nonzero scalar \( c \).

\[
E_M(i, c) = [e_1, \cdots, c \times e_i, \cdots, e_m]
\]

(ii) \( E_i(i, j) \) matrix gotten from \( I_m \) by interchanging its \( i^{th} \) and \( j^{th} \) rows.

\[
E_i(i, j) = [e_1, \cdots, e_j, \cdots, e_i, \cdots, e_m]
\]

(iii) \( E_A(i, j) \) matrix gotten from \( I_m \) by adding its \( i^{th} \) to its \( j^{th} \) row.

\[
E_A(i, j) = [e_1, \cdots, e_i, \cdots, e_i + e_j, \cdots, e_m]
\]

**Example 8-1.** For a system of \( m = 3 \) equations,

\[
E_i(1,2) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E_m(1,2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E_A(1,2) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}
\]

**Theorem 8-2.** Some basic row reduction results are:

(i) Any elementary row operation on a matrix \( A \) can be accomplished by pre-multiplying \( A \) by the corresponding elementary matrix.

(ii) \( E_i(i, j)^{-1} = E_i(i, j). \) \( E_M(i, c)^{-1} = E_M(i, 1/c). \) \( E_A(i, j)^{-1} = E_M(i, -1)E_A(i, j)E_M(i, -1). \)

(iii) Any matrix may be transformed into a reduced row Echelon form by a finite sequence of elementary row operations.

(iv) The only nonsingular reduced row Echelon form is \( I \).

**Proof.** (i-1) Apply the definitions and check the product. Since the \( i^{th} \) row of \( E_M(i, c) \) is \( ce_i^T \), \( ce_i \cdot col_j A = c \cdot a_{ij} \) for all \( j \) and \( e_k \cdot col_j A = a_{kj} \) for all \( k \neq i \). Therefore, \( row_iE_M(i, c) \cdot A = c(e_i)^T A = c \times row_i A \) and \( row_kE_M(i, c) \cdot A = row_k A \) for \( k \neq i \) as claimed.
(i-2) \( \text{row}_i E_i(i, j) \cdot A = \text{row}_j A \), \( \text{row}_j E_i(i, j) \cdot A = \text{row}_i A \), and \( \text{row}_k E_i(i, j) \cdot A = \text{row}_k A \) for \( k \neq i, j \)

(i-3) \( \text{row}_j E_A(i, j) \cdot A = \text{row}_i A + \text{row}_i A \) and \( \text{row}_k E_A(i, j) \cdot A = \text{row}_k A \) for \( k \neq i \)

(ii) As in (i), calculate the product of \( EE^{-1} \) to show \( EE^{-1} = I \).

(iii) Here is the procedure of Gaussian Elimination method.

(step 0) If the matrix is already in reduced row echelon form, the stop.

(step 1) If not, find the first column from the left with non-zero entry \( a \) and move the row containing that entry to the top of the rows being worked on.

(step 2) Multiply that row by \( 1/a \) to get leading 1.

(step 3) Subtract multiples of that row from the rows below it to make each entry below the leading 1 zero.

(step 4) Repeat steps 0-3 on the rows below.

For an \( n \times m \) matrix, the maximum number of operations for each column is \( 1 + 1 + (3 \times m) \) for the step 2 through 4.

(iv) Otherwise, it should have a zero along the main diagonal or a non-zero entry in off-diagonal. That is either not invertible or not in reduced row Echelon form.

This theorem implies that an invertible matrix can be expressed as a product of elementary matrices. We can find inverse matrices using this property.

**Corollary 8-3.** Let \( A \) be a nonsingular matrix and let \( E_1, \cdots, E_p \) be elementary matrices such that \( E_p \cdots E_1 A = R \), where \( R \) is a reduced row Echelon matrix. Then

(i) \( R = I \).

(ii) \( A^{-1} = E_p \cdots E_1 \). Hence \( A^{-1} \) may be computed by preforming the elementary row operations representing by the \( E_i \) on \( I \).

(iii) \( A = E_1^{-1} \cdots E_p^{-1} \). Hence any nonsingular matrix may be represented as a product of elementary matrices.
8.2 Real Numbers

**Example 8-2.**

\[
\begin{bmatrix}
-3 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 2 \\
3 & 4
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
1 & 0 \\
1 & 1
\end{bmatrix}
\begin{bmatrix}
-3 & -6 \\
3 & 4
\end{bmatrix}
= \begin{bmatrix}
-3 & -6 \\
0 & -2
\end{bmatrix}
\]

\[
\Rightarrow
\begin{bmatrix}
-1 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
-3 & -6 \\
0 & -2
\end{bmatrix}
= \begin{bmatrix}
1 & 2 \\
0 & -2
\end{bmatrix}
\]

\[
\Rightarrow
\begin{bmatrix}
1 & 1 \\
0 & -1/2
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & -2
\end{bmatrix}
= \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\]

That is,

\[
E_M(2, -1/2)E_A(2,1)E_M(1, -1/3)E_A(1,2)E_M(1, -3)A = I
\]

Therefore,

\[
A^{-1} = E_M(2, -1/2)E_A(2,1)E_M(1, -1/3)E_A(1,2)E_M(1, -3)
\]

\[
= \left(\begin{bmatrix}
1 & 0 \\
0 & -1/2
\end{bmatrix} \begin{bmatrix}
1 & 1 \\
0 & 1
\end{bmatrix}\right) \left(\begin{bmatrix}
-1/3 & 0 \\
0 & 1
\end{bmatrix} \begin{bmatrix}
1 & 0 \\
1 & 1
\end{bmatrix}\right) \begin{bmatrix}
-3 & 0 \\
0 & 1
\end{bmatrix}
\]

\[
= \begin{bmatrix}
1 & 1 \\
0 & -1/2
\end{bmatrix} \begin{bmatrix}
-1/3 & 0 \\
0 & 1
\end{bmatrix} \begin{bmatrix}
1 & 0 \\
-3 & 1
\end{bmatrix} = \begin{bmatrix}
-2 & 1 \\
3/2 & -1/2
\end{bmatrix}
\]

### 8.2 Solving a Linear System

Let \( A \) be an \( n \times m \) matrix and consider the linear equation system \( Ax = b \). The equations \( x_1 \text{col}_1A + \cdots + x_m \text{col}_mA \) are the linear combination of columns of \( A \). In a linear equation system \( Ax = b \), a solution exists only when the constant vector \( b \) is an element of the vector space spanned by \( A \). The rank condition for the existence of the solution is in Theorem 8-7.

Suppose that a solution exists. If the augmented matrix \( (A|b) \) can be transformed into \( (P|q) \) by elementary row operations, then the system \( Ax = b \) and \( Px = q \) have the same solutions. Thus, \( Ax = b \) may be solved as follows.

(i) transform \( (A|b) \) to a reduced echelon form \( (R|c) \) and consider the solution to \( Rx = c \).
(ii) Let \( k \) be the number of non-zero rows. Reorder the columns of \( R \) and the elements of \( x \) so that \( Rx = c \) partition as

\[
\begin{bmatrix}
I_k & R_0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} =
\begin{bmatrix}
c_0 \\
0
\end{bmatrix}, \quad \text{or} \quad x_1 = c_0 - R_0 x_2
\]

where \( R_0 \) is \( k \times (n-k) \), \( x_1 \) and \( c_0 \) are \( k \times 1 \), and \( x_2 \) is \((n-k) \times 1\).

(iii) Generate all possible solutions \( x \) by arbitrarily setting \( x_2 \) and computing \( x_1 \).

### 8.3 Kronecker Products and Stacking Operator

Consider the equation \( BXC = D \), where \( B, C, D \) and \( X \) are matrices conformable for the operations shown. View \( B, C, \) and \( D \) as given; and view \( X \) as a matrix of unknowns. Each scalar equation of the matrix equation \( BXC = D \) is linear in the elements of \( X \); so \( BXC = D \) is a system of linear equations in the elements of \( X \). Yet the system is not in the standard form \( Ax = b \). Problems of this general type appear fairly often in working with variance-covariance matrices.

**Definition 8.8.** Let \( A = [a_{ij}] = [A_1, \ldots, A_n] \) be an \( m \times n \) matrix, and let \( B \) be a matrix of any size. Then the **Kronecker product** of \( A \) and \( B \), written \( A \otimes B \), and the **stack** of \( A \), written \( \text{vec}(A) \), are defined by

\[
A \otimes B =
\begin{bmatrix}
a_{11}B & a_{12}B & \cdots & a_{1n}B \\
a_{21}B & a_{22}B & \cdots & a_{2n}B \\
\vdots & \vdots & \ddots & \vdots \\
a_{m1}B & a_{m2}B & \cdots & a_{mn}B
\end{bmatrix}
\quad \text{and} \quad
\text{vec}(A) =
\begin{bmatrix}
A_1 \\
A_2 \\
\vdots \\
A_n
\end{bmatrix}.
\]

**Theorem 8.4 (Properties)**

(i) \( \text{vec}(aA + \beta B) = a\text{vec}(A) + \beta\text{vec}(B) \).

(ii) \( \text{vec}(AYB) = (B^T \otimes A)\text{vec}(Y) \).

(iii) \( A \otimes (B \otimes C) = (A \otimes B) \otimes C \), \( (A \otimes B)(C \otimes D) = (A \otimes C)(B \otimes D) \).

(iv) \( A \otimes (B + C) = A \otimes B + A \otimes C \), \( (B + C) \otimes A = B \otimes A + C \otimes A \).

(v) \( (A \otimes B)^T = A^T \otimes B^T \).
(vi) \((A \otimes B)^{-1} = A^{-1} \otimes B^{-1}\).

(vii) \(\det(AB) = [\det(A)]^m \times [\det(B)]^n\), where \(A\) is \(m \times m\) and \(B\) is \(n \times n\).

(viii) \(\text{tr}(A \otimes B) = \text{tr}(A) \times \text{tr}(B)\).

(ix) \(r(A \otimes B) = r(A) \times r(B)\).

(x) Let \(A\) be an \(m \times m\) matrix with eigenvalues \((\lambda_1, \ldots, \lambda_m)\) and the corresponding eigenvectors \((x_1, \ldots, x^m)\). Let \(B\) be an \(n \times n\) matrix with eigenvalues \((\mu_1, \ldots, \mu_n)\) and corresponding eigenvectors \((z_1, \ldots, z^n)\). Then the \(mn \times mn\) matrix \(A \otimes B\) has eigenvalues \((\lambda_1, \ldots, \lambda_m) \otimes (\mu_1, \ldots, \mu_n)\) and corresponding eigenvectors \((x_1, \ldots, x^m) \otimes (z_1, \ldots, z^n)\). That is, for each \((i, j)\), \(\lambda_i \mu_j\) is an eigenvalue of \(A \otimes B\) with the corresponding eigenvector \(x^i \otimes z^j\).

Now we can easily unscramble \(BXC = D\). Apply the stacking operator to both sides to get \(\text{vec}(BXC) = \text{vec}(D)\). Next, apply (ii) to the left hand side. The result is \((C^T \otimes B)\text{vec}(x) = \text{vec}(D)\). This is of the form \(Ax = b\).

### 8.4 Matrix and a Linear Map

An \(n \times m\) matrix \(A\) can be thought of as a linear map from \(\mathbb{R}^n\) to \(\mathbb{R}^m\) such that \(y = Ax\).

**Theorem 8-5.** The function \(T: \mathbb{R}^n \rightarrow \mathbb{R}^m\) is linear if and only if there exists an \(n \times m\) matrix \(A\) such that \(Tx = Ax\) for all \(x \in \mathbb{R}\).

**Proof.** Define an \(n \times m\) matrix \(A\) by \(\text{col}_i A = Te_i\), \(i = 1, \ldots, n\). Since \(x = \sum_{i=1}^{n} x_i e_i\) for all \(x \in \mathbb{R}^n\), by linearity,

\[
Tx = \sum_{i=1}^{n} x_i Te_i = \sum_{i=1}^{n} x_i \text{col}_i A = Ax
\]

For the converse,

\[
T(ax + \beta y) = A(ax + \beta y) = Aax + A\beta y = aAx + \beta Ay = aTx + \beta Ty
\]

That is, \(T\) is linear.
There is a one-to-one correspondence between linear map and matrices, and the range of the map is typically written as \( \text{span}(A) \) instead of the range of \( T \).

\[
\text{span}(A) = \{ Ax | x \in \mathbb{R}^n \} = \text{span}(\text{col}_1 A, \ldots, \text{col}_n A)
\]

For this reason, it is called the column space of \( A \).

**Definition 8-9.** The column rank of a matrix is the dimension of its column space and the row rank of a matrix is the dimension of its row space.

**Theorem 8-6.** The column rank and the row rank are always equal.

**Proof.** We show \( \text{rank}(A) = \text{rank}(A^T) \). Let \( A \) be an \( m \times n \) matrix. Suppose that \( \text{rank}(A^T) = \dim(\text{span}(A^T)) = k \) and let \( V = \{ v_1, v_2, \ldots, v_k \} \) be a basis for \( \text{span}(A^T) \). Then each \( \text{col}_i A^T \) is a linear combination of \( V \) with the weights of \( c_{ij}, i = 1, \ldots, m, j = 1, \ldots, k \).

\[
\text{col}_i A^T = \begin{pmatrix}
\vdots \\
 a_{1i}
\end{pmatrix}
= c_{i1} \begin{pmatrix}
\vdots \\
 v_{11}
\end{pmatrix} + \cdots + c_{ik} \begin{pmatrix}
\vdots \\
 v_{k1}
\end{pmatrix}
\]

Taking only the \( l^{th} \) entry of \( \text{col}_i A^T \) gives

\[
a_{il} = c_{i1} v_{1l} + \cdots + c_{ik} v_{kl}
\]

And, by stacking terms over \( l \), we have

\[
\text{col}_i A = \begin{pmatrix}
\vdots \\
 a_{mi}
\end{pmatrix}
= v_{il} \begin{pmatrix}
\vdots \\
 c_{1i}
\end{pmatrix} + \cdots + v_{kl} \begin{pmatrix}
\vdots \\
 c_{mi}
\end{pmatrix}
= v_{il} c_1 + \cdots + v_{kl} c_k,
\]

where \( c_j = \begin{pmatrix}
\vdots \\
 c_{1j}
\end{pmatrix} \)

This shows that \( \text{span}\{\text{col}_1 A, \ldots, \text{col}_n A\} = \text{span}(A) \subset \text{span}\{c_1, \ldots, c_k\} \). Since the dimension of a vector space is no smaller than the dimension of its subspace,

\[
\text{rank}(A) = \dim(\text{span}(A)) \leq \dim \text{span}(c_1, \ldots, c_k) = k = \text{rank}(A^T)
\]

Applying the inequality to \( A^T \) gives \( \text{rank}(A^T) \leq \text{rank}(A) \). Therefore, \( \text{rank}(A^T) = \text{rank}(A) \). \( \blacksquare \)

Since those two ranks are equal, we call the number the **rank**.

If the rank of \( A \) is the same as the number of columns, \( A \) is said to have **full column rank**. Together with Theorem 8-7. (i), \( \dim(\text{span}(A)) \) is be smaller than the number of columns and thus the rank attains its maximum. Similarly, we can define the **full row rank**.
**THEOREM 8-7 (Properties of Rank)**

(i) for an \( m \times n \) matrix \( A \), \( \text{rank}(A) \leq \min(m, n) \).

(ii) \( \text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B)) \)

(iii) If \( P \) and \( Q \) are nonsingular, then \( \text{rank}(PAQ) = \text{rank}(PA) = \text{rank}(AQ) = \text{rank}(A) \).

(iv) If \( A \) has reduced Echelon form \( R \), then \( \text{rank}(A) = \text{rank}(R) = \text{number of nonzero rows in } R \).

(v) An \( n \times n \) matrix \( A \) is nonsingular if and only if \( \text{rank}(A) = n \).

(vi) A linear equation system \( Ax = b \) has:

(a) a solution if and only if \( \text{rank}(A|b) = \text{rank}(A) \).

(b) at most one solution if and only if \( \text{rank}(A) \) is the same as the number of unknowns.

**Proof.** (i) Since \( \dim(\text{span}(A)) \leq m \) and \( \dim(\text{span}(A^T)) \leq n \), the result follows from Theorem 8-6.

(ii) Let \( V = \text{span}(AB) \) and \( W = \text{span}(A) \). Since each matrix corresponds to a linear map, for any vector \( y \in V \), there exists a vector \( x \in \mathbb{R}^n \) such that \( y = (AB)x \). Let \( z = Bx \in \mathbb{R}^n \). Then \( y = A(Bx) = Az \), and thus \( y \in W \) and \( V \subset W \). Therefore,

\[
\text{rank}(AB) = \dim(\text{span}(AB)) \leq \dim(\text{span}(A)) = \text{rank}(A)
\]

(iii) By (ii), \( \text{rank}(P^{-1}(PA)) \leq \text{rank}(PA) \) and \( \text{rank}(PA) \leq \text{rank}(A) \). Therefore,

\[
\text{rank}(A) = \text{rank}(P^{-1}(PA)) \leq \text{rank}(PA) \leq \text{rank}(A)
\]

The other qualities can be shown in the same way.

(iv) Since the elementary row operator are nonsingular, the result follows from (iii)

(v) The reduced row Echelon form of a nonsingular matrix is an identity matrix by Theorem 8-2, (iv). By (iv), the rank of \( A \) is \( n \). Conversely, if \( A \) has full column rank, then the standard basis of the column space is the collection of columns of the identity matrix. By Theorem 7-7, (xii), the product of invertible matrices is invertible.
(v)(a) Let $R$ be the row Echelon form of $A$. Then the row Echelon form of $(A|b)$ should be $(R|c)$. Since $R$ and $(R|c)$ have the same column rank, $c$ can be expressed as a linear combination of the columns of $R$. The weight is a solution to the equation system.

(b) Suppose that $A$ is a $m \times n$ matrix. If $\text{rank}(A)$ and the number of unknowns are equal, $A$ should have full column rank, $n$, and $n \leq m$. If $m = n$, $A$ is invertible and the system has a unique solution. If $m > n$ and $\text{rank}(A|b) = n$, then its reduced row Echelon form is

$$\begin{bmatrix} I_n & c \\ 0 & 0 \end{bmatrix}$$

Again it has a unique solution. If $m > n$ and $\text{rank}(A|b) > n$, there is no solution. Note that it is not possible to have $m > n$ and $\text{rank}(A|b) < n$. ■

8.5 LINEAR MAP AND NULL SPACE

**Definition 8-10.** If the $i$th column of the row echelon form contains a pivot, $x_i$ is called basic or leading variable. Otherwise, we call $x_i$ a free variable.

According to Theorem 8-7 (iv), the rank of a matrix is the number of nonzero rows in its row Echelon form, and thus the rank of a matrix is the same as the number of basic variables. The number of free variables is related to its null space.

**Definition 8-11.** The null space or kernel of $A$ is the set of solutions to a linear equation system $Ax = 0$. Its rank is called nullity.

**Theorem 8-8 (Rank-Nullity Theorem)** Let $A$ be an $m \times n$ matrix. The null space of $A$ is a vector space of dimension $n - \text{rank}(A)$, or

$$\text{rank}(A) + \text{nullity}(A) = \text{dim}(E^n)$$

**Proof.** First, we prove the result for the row Echelon form and then show that it holds for the matrix.

The reduced row Echelon form of $A$ has the form of

$$R = \begin{bmatrix} I_k & R_0 \\ 0 & 0 \end{bmatrix},$$

where $R_0$ is a $k \times (n - k)$ matrix.
so that there are \( k \) basic variables and \( (n - k) \) free variables. Since the first through \( k^{th} \) elements of a basis vector for the null space should not have a non-zero element, \( \{e_{k+1}, \ldots, e_n\} \) is the standard basis for the null space of the reduced row Echelon form of \( A \). Therefore, the nullity is the number of free variables. Since the rank of \( A \) is the number of basic variables,

\[
\text{rank}(R) + \text{nullity}(R) = \dim(E^n)
\]

Finally, we need to show that the null space of a matrix \( A \) is the same as the null space of row Echelon form of \( A \).

Suppose \( x \in \ker A \) so that \( Ax = 0 \). The reduced row Echelon form of \( A \) is obtained by a sequence of elementary row operations, \( R = MA \) and \( M \) is invertible. Then \( Rx = MAx = M0 = 0 \) and thus \( x \in \ker R \). Similarly, if \( x \in \ker R \), \( Ax = M^{-1}Rx = M^{-1}0 = 0 \) and \( x \in \ker A \). Therefore, \( \ker A = \ker R \). Together with THEOREM 8-7 (iv),

\[
\text{rank}(A) + \text{nullity}(A) = \dim(E^n)
\]

**THEOREM 8-9.** Let \( A \) be an \( m \times n \) matrix.

(i) Let \( \alpha \) be a solution to \( Ax = b \). Then the set of all solutions to \( Ax = b \) is \( \{x|x = \alpha + \beta \text{ with } \beta \in \ker A\} \).

(ii) For any vector space \( S \) in \( \mathbb{E}^n \), there is a matrix \( A \) such that \( S \) can be represented as the solution set of \( Ax = 0 \), the null space of \( A \).

(iii) Let \( V \) be the vector space spanned by the columns of \( A^T \) (the row space). Then

(a) any vector in \( V \) is orthogonal to any vector in \( \ker A \).

(b) a basis for \( V \) together with a basis for \( \ker A \) forms a basis for \( E^n \).

As shown in Section 8.2, if the number of basic variables is \( k \), the solution has a form of \( (x_1, x_2) \in \mathbb{R}^n \) where \( x_1 = c_0 - R_0x_2 \in \mathbb{R}^k \) and \( x_2 \in \mathbb{R}^{n-k} \).

Of the basic variables has a unique solution while the free variables constitute the null space.

**Proof of THEOREM 7-6.**

Injection implies surjection: By THEOREM 7-5, \( T \) is injective if and only if \( \ker(T) = \{0\} \), that is, if and only if it the nullity is zero. Then by the rank-nullity theorem, \( \text{rank}(T) = n \) or
\[ \dim T(\mathbb{R}^n) = n. \] Since \( T(\mathbb{R}^n) \) is a subset of the domain and they have the full dimension, they should coincide. Therefore, the linear map is surjective.

Surjection implies injection: Since \( T(\mathbb{R}^n) = \mathbb{R}^n \), by the rank-nullity theorem, \( \dim \ker(T) = 0 \). Therefore, \( T \) is injective. By THEOREM 7-5.

These establish the equivalence of (i) through (iv).

Surjection \( \iff \{ T e_1, \ldots, T e_n \} \) is linearly independent: By THEOREM 7-4, \( T(\mathbb{R}^n) = \text{span}(X) \).

9 DETERMINANT

9.1 DEFINITION AND FORMULA

DEFINITION 9-1. A function \( f : \mathbb{R}^{[1, \ldots, n] \times [1, \ldots, n]} \to \mathbb{R} \), where \( \mathbb{R}^{[1, \ldots, n] \times [1, \ldots, n]} \) is the set of \( n \times n \) matrices, is a determinant function if it satisfies the following properties:

(i) Linearity: \( f(A) \) is linear in each columns of the matrix.

(ii) Alternating: \( f(A) = 0 \) if two columns are equal.

(iii) Normalization: \( f(I_n) = 1 \)

The determinant function is denoted by \( \det A \) or \( |A| \). If the properties hold for columns, then it also holds for rows. I use columns to save space in proofs.

For the sake of notation, let \( A_i = \text{col}_i A \). The linearity is

\[ \det(A_1, \ldots, A_{i-1}, \alpha A_i + \beta B_i, A_{i+1}, \ldots, A_n) = \alpha \det(A) + \beta \det(A_1, \ldots, A_{i-1}, B_i, A_{i+1}, \ldots, A_n) \]

where \( B_i \) is a \( n \times 1 \) vector.
The alternating property can be replaced with interchanging property that \( f(A) \) reverses sign if two columns are interchanged. With the linearity, we will show that each can be derived from the other.

Before we investigate the properties of the determinant function, let’s check how to calculate the determinant of a matrix.

Let \( S = (1, \ldots, n) \) and \( \sigma: S \to S \) is a permutation of \( S \), which is a bijective function. Let \( S_n \) denote the collection of all permutations. \( S_n \) has \( n! \) elements that is the number of all possible permutations for \( S \). Each permutation is simply a sequence of transposing two adjacent elements.

For instance, when \( n = 3 \), \( S = \{1, 2, 3\} \) and \( S_n = \{\sigma^1, \ldots, \sigma^n\} \) with

\[
\sigma^1 = (1, 2, 3), \sigma^2 = (1, 3, 2), \sigma^3 = (2, 1, 3), \sigma^4 = (2, 3, 1), \sigma^5 = (3, 1, 2), \sigma^6 = (3, 2, 1),
\]

The number of transposition to recover a given permutation is not unique, but it is unique up to whether it is even or odd. If the number of transpositions for a permutation is even/odd, the permutation is said to be even/odd. \( \text{sgn} \sigma = +1 \) if \( \sigma \) is even and \( \text{sgn} \sigma = -1 \) if it is odd.\(^6\)

In the example above, to recover \( S \) from \( (2, 3, 1) \), we have three transpositions:

\[
(2, 3, 1) \to (2, 1, 3) \to (1, 2, 3)
\]

Therefore, \( \text{sgn} \sigma^4 = -1 \).

**THEOREM 9-1. (Leibniz formula)**

\[
\det A = \sum_{\sigma \in S_n} (\text{sgn} \sigma) a_{1\sigma_1} a_{2\sigma_2} \cdots a_{n\sigma_n}
\]

The determinant can be defined in a recursive way.

\(^6\) Or equivalently, one can count the number of inversions: the number of pairs of elements of \( \sigma \), not necessarily adjacent, for which a larger integer precedes a smaller one.)
THEOREM 9-2. (Laplace formula) Let $A$ be an $n \times n$ matrix and $A_{ij}$ be $(n-1) \times (n-1)$ a submatrix of $A$ obtained by removing row $i$ and column $j$ from $A$. Then, for any $i$ or $j$,

$$
\det A = (-1)^{i+j+1}a_{i1}A_{1j} + \cdots + (-1)^{i+n}a_{in}A_{in} = \sum_{k=1}^{n} (-1)^{i+k}a_{ik}A_{ik}
$$

EXAMPLE 9-1. Let’s find the determinant of a $2 \times 2$ matrix with Leibniz formula. The permutations of column indices is $(1,2)$ and $(2,1)$. Since $(2,1)$ involves a single inversion, the determinant of the matrix is

$$
\det A = (a_{11} \times a_{22}) + (-1)(a_{12} \times a_{21})
$$

$$
\det A = \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = (1 \times 4) - (2 \times 3) = 4 - 6 = -2
$$

According to Laplace formula,

$$
\det A = \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = (-1)^{1+1} \times 1 \times 4 + (-1)^{1+2} \times 2 \times 3 = 4 - 6 = -2
$$

Observe that the first term is the product of the diagonal elements and the second is the product of the off diagonal elements.

$$
\begin{pmatrix} (+) & (-) \\ \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} & \det A = (1 \times 4) - (2 \times 3) = -2
\end{pmatrix}
$$

In a $3 \times 3$ matrix, the permutations are $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$, $(3,2,1)$. We need an odd number of inversion to get $(1,3,2)$, $(2,1,3)$, $(3,2,1)$ and even for $(1,2,3)$, $(2,3,1)$, $(3,1,2)$.

$$
\det A = (a_{11} \times a_{22} \times a_{33} + a_{12} \times a_{23} \times a_{31} + a_{13} \times a_{21} \times a_{32})
$$

$$
+ (-1)(a_{11} \times a_{23} \times a_{32} + a_{12} \times a_{21} \times a_{33} + a_{13} \times a_{22} \times a_{31})
$$

$$
\det A = \begin{vmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 2 & 2 & 1 \end{vmatrix}
$$

$$
= ((1 \times 1 \times 1) + (2 \times (-1) \times 2) + (1 \times 0 \times 2))
$$

$$
- ((1 \times (-1) \times 2) + (2 \times 0 \times 1) + (1 \times 1 \times 2))
$$

$$
= (1 + (-4)) - ((-2) + 2)) = -3
$$
Alternatively,
\[
\det A = \begin{vmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 2 & 2 & 1 \end{vmatrix} = (-1)^{2+1} \times 0 \times \begin{vmatrix} 2 \\ 1 \end{vmatrix} + (-1)^{2+2} \times 1 \times \begin{vmatrix} 1 \\ 1 \end{vmatrix} + (-1)^{2+3} \times (-1) \times \begin{vmatrix} 1 \\ 2 \end{vmatrix} = 0 + (1 - 2) + ((-1) \times (-1) \times (2 - 4)) = (-1) + (-2) = -3
\]

The rule of Sarrus is a mnemonic for the determinant of a $3 \times 3$ matrix.

9.2 Properties

Theorem 9-3. $\det(A) = \det(A^T)$ and $\det(\text{EA}) = \det(\text{AE})$ where $E$ is an elementary matrix.

Proof. Let $B = A^T$. Then $b_{ij} = a_{ji}$. The result is direct from the Leibniz formula.

\[
\det A = \sum_{\sigma \in S_n} (\text{sgn } \sigma) a_{1\sigma_1} \times a_{2\sigma_2} \times \cdots \times a_{n\sigma_n}
\]

\[
\det A^T = \sum_{\sigma \in S_n} (\text{sgn } \sigma) a_{\sigma_1 1} \times a_{\sigma_2 2} \times \cdots \times a_{\sigma_n n}
\]

Since $(\text{EA})^T = \text{AE}$, $\det(\text{EA}) = \det(\text{AE})$. 

The Laplace formula also can be applied to a column.

\[
\det A = (-1)^{1+j} a_{1j} |A_{11}| + \cdots + (-1)^{n+j} a_{nj} |A_{nj}| = \sum_{k=1}^{n} (-1)^{k+j} a_{kj} |A_{kj}|
\]

Definition 9-1 is equally valid for columns and thus any result for rows also holds for columns. In following, the elementary operations are applied to columns such that $\text{AE}$.

Theorem 9-4. Let $A$ and $B$ be $n \times n$ matrices. Then:
(i) \( \det(I) = 1 \)

(ii) \( \det E_M(i, c) = c, \ \det E_i(i, j) = -1, \ \det E_A(i, j) = 1 \)

(iii) \( \det(AE) = \det(A)\det(E) \) for all three elementary matrices.

**Proof.** (i) It is the normalization part of the definition.

(ii-1) By linearity and normalization,

\[
\det E_M(i, c) = \det[e_1, \ldots, c \times e_i, \ldots, e_m] = c \det[e_1, \ldots, e_i, \ldots, e_m] = c + \det I = c
\]

(ii-2) Consider a matrix with \( A_i + A_j \) at both \( i^{th} \) and \( j^{th} \) columns. By linearity,

\[
\det(e_1, \ldots, e_i, \ldots, e_j, \ldots, e_n) = \det(e_1, \ldots, e_j, \ldots, e_j, \ldots, e_n) + \det(e_1, \ldots, e_i, \ldots, e_j, \ldots, e_n)
\]

And by the alternating property

\[
\det(e_1, \ldots, e_i, \ldots, e_j, \ldots, e_n) = 0
\]

Therefore,

\[
\det(e_1, \ldots, e_i, \ldots, e_j, \ldots, e_n) + \det(e_1, \ldots, e_j, \ldots, e_j, \ldots, e_n) = 0
\]

(ii-3) By linearity and the alternating property

\[
\det(e_1, \ldots, e_i + \alpha e_j, \ldots, e_n) = \det(e_1, \ldots, e_i, \ldots, e_j, \ldots, e_n) + \alpha \det(e_1, \ldots, e_j, \ldots, e_j, \ldots, e_n) = 1
\]

(iii-1) By COROLLARY 8-3 (i) and linearity,

\[
\det AE_M(i, c) = \det(A_1, \ldots, c \times A_i, \ldots, A_n) = c \det(A_1, \ldots, A_i, \ldots, A_n) = c \det A
\]

(iii-2) Consider a matrix with \( A_i + A_j \) at both \( i^{th} \) and \( j^{th} \) columns. By linearity,

\[
\det(A_1, \ldots, A_i + A_j, \ldots, A_i + A_j, \ldots, A_n)
\]

By the alternating property,
\[
\det(A_1, \ldots, A_i + A_j, \ldots, A_i + A_j, \ldots, A_n) = 0
\]

and
\[
\det(A_1, \ldots, A_i, \ldots, A_i, \ldots, A_n) = \det(A_1, \ldots, A_j, \ldots, A_j, \ldots, A_n) = 0
\]

Therefore,
\[
\det(A_1, \ldots, A_i, \ldots, A_j, \ldots, A_n) + \det(A_1, \ldots, A_j, \ldots, A_i, \ldots, A_n) = 0
\]

Since \(\det(A_1, \ldots, A_j, \ldots, A_i, \ldots, A_n) = AE_i(i, j)\), \(\det A + \det AE_i(i, j) = 0\) and
\[
\det AE_i(i, j) = \det A \times (-1) = \det A \times \det E_i(i, j)
\]

(iii) By linearity and the alternating property,
\[
\det AE_A(i, j) = \det(A_1, \ldots, A_i, \ldots, A_i + A_j, \ldots, A_n)
\]
\[
= \det(A_1, \ldots, A_i, \ldots, A_i, \ldots, A_n) + \det(A_1, \ldots, A_i, \ldots, A_j, \ldots, A_n)
\]
\[
= \det A \times 1 = \det A \times \det E_A(i, j)
\]

Note that the proof of \(\det E_i(i, j) = -1\) show that interchanging property implies the alternating property. That is, if two columns are equal, an interchange of columns should not alter the determinant but they should have a different sign because zero is the unique number with such a property. The alternating property in the definition of determinant can be replaced with the interchanging property.

**Theorem 9-5.** Let \(A\) and \(B\) be \(n \times n\) matrices. Then:

(i) \(\det A = 0\) if and only if \(A\) is singular.

(ii) \(\det AB = \det A \det B = \det B \det A = \det BA\)

(iii) \(\det A^{-1} = 1/\det A\)

(iv) \(\det(A) = a_{11} \cdots a_{nn}\) if \(A\) is diagonal or triangular.

**Proof.** (i) To reach a contradiction, suppose that \(\det A = 0\) but \(A\) is invertible. By Theorem 8-2,
\[ A = E_1^{-1} \ldots E_p^{-1} = E_1 \ldots E_k \]

By applying (iv) sequentially,
\[
\det A = \det(E_1 \ldots E_{k-1}) \det E_k = \det(E_1 \ldots E_{k-2}) \det E_k = \ldots = \det(E_1) \ldots \det E_k
\]

However, since the determinant of every elementary matrix is not zero, \( \det A \) cannot be zero, which is a contradiction.

Conversely, if \( A \) is singular, all the elements in the last row of its row Echelon form are zero. Then, for any \( c \), we must have \( \det A = \det AE_M(n,c) = c \det A \) and \( \det A = 0 \).

(ii) Case 1: \( B \) is invertible. Then by the result of (i) and THEOREM 9-4 오류! 참조 원본을 찾을 수 없습니다. (iii),
\[
\det AB = \det AE_1 \ldots E_k = \det AE_1 \ldots E_{k-1} \det E_k = \det A \det E_1 \ldots \det E_k
\]

Similarly, \( B = \det E_1 \ldots E_k = \det E_1 \ldots \det E_k \). Therefore, \( \det AB = \det A \det B \)

Case 2: \( B \) is singular. By THEOREM 7-7 (viii), \( \det AB = 0 \) and \( \det A \times \det B = 0 \).

The last result follows from that \( \det AB = \det A \det B = \det B \det A = \det BA \).

(iii) \( \det A^{-1} A = \det A^{-1} \det A = 1 \)

(iv) According to Laplace formula

Since \( a_{ij} = 0 \) for all \( i > j \),
\[
\det A = (-1)^{i+j} a_{11} |A_{11}| + \ldots + (-1)^{i+n} a_{1n} |A_{1n}|
= a_{11} |A_{11}| = a_{11} a_{22} |A_{22}| = \ldots = a_{11} a_{22} \ldots a_{nn}
\]

\( \blacksquare \)

**EXAMPLE 9-2.** \( \det cA = c^n \det A \)

\[
\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = -2, \quad c \begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = c \begin{vmatrix} 2c & 2c \\ 3c & 4c \end{vmatrix} = 4c^2 - 6c^2 = -2c^2 \]

\( \square \)

**THEOREM 9-6.** Let \( X = [x_1, \ldots, x_n] \) with \( x_i \in \mathbb{R}^n, i = 1, \ldots, n \). The following are equivalent.

(i) The set of columns \( \{x_1, \ldots, x_n\} \) is linearly dependent.
(ii) \( Xb = 0 \) has a nonzero solution \( b \)

(iii) \( \det(X) = 0 \)

(iv) \( \operatorname{rank}(X) < n \)

**Proof.** (i) \( \iff \) (ii) The definition of linear dependence.

(i) \( \iff \) (iii) Since a column is a linear combination of other columns, by linearity and the alternating property, \( \det(X) = 0 \). Conversely, \( \det(X) = 0 \) implies that \( X \) is singular. Then the result follows from (i) \( \iff \) (ii).

(i) \( \iff \) (iv) \( \operatorname{rank}(X) \) is the dimension of its column space and \( \dim \text{span}(X) \) is the number of elements in a linearly independent set, which is less than \( n \).

\[ \blacksquare \]

**9.3 Cramer’s Rule**

**Theorem 9-7 (Cramer’s rule)** If a square matrix \( A \) is nonsingular, the unique solution for the system \( Ax = b \) is given by

\[ x_i = \frac{\det(\text{col}_1 A, \ldots, \text{col}_{i-1} A, b, \text{col}_{i+1} A, \ldots, \text{col}_n A)}{\det A} \]

**Proof.** Since \( A \) is invertible,

\[ b = x_1 A_1 + x_2 A_2 + \cdots + x_n A_n \]

Then

\[ \det(A_1, \ldots, A_{i-1}, b, A_{i+1}, \ldots, A_n) = \det \left( A_1, \ldots, A_{i-1}, \sum_{k=1}^{n} x_k A_k, A_{i+1}, \ldots, A_n \right) \]

\[ = \sum_{k=1}^{n} x_k \det(A_1, \ldots, A_{i-1}, A_k, A_{i+1}, \ldots, A_n) \]

\[ = x_i \det(A_1, \ldots, A_{i-1}, A_i, A_{i+1}, \ldots, A_n) \]

\[ + \sum_{k \neq i} x_k \det(A_1, \ldots, A_{i-1}, A_k, A_{i+1}, \ldots, A_n) \]

Where the second term in the last equation is zero by the alternating property.
\[
\begin{align*}
\det(A_1, \ldots, A_{i-1}, b, A_{i+1}, \ldots, A_n) &= \frac{x_i \det A}{\det A} = x_i \\
\end{align*}
\]

**EXAMPLE 9-3.** Consider an equation system,

\[
\begin{bmatrix}
1 & 2 \\
3 & 4
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
5 \\
6
\end{bmatrix}
\]

Applying Cramer's rule gives

\[
x_1 = \frac{1 \cdot 6 - 2 \cdot 3}{4 - 6} = \frac{6 - 6}{-2} = 3,
\quad x_2 = \frac{1 \cdot 3 - 5 \cdot 2}{4 - 6} = \frac{3 - 10}{-2} = \frac{7}{2}
\]

**9.4 ADJOINT MATRIX**

To calculate the determinant of a large matrix, it is efficient to use the Laplace formula, and it also helps to express inverse matrices systematically. Recall the Laplace formula.

\[
\det A = (-1)^{i+j} a_{ij} |A_{i1}| + \cdots + (-1)^{i+n} a_{in} |A_{in}| = \sum_{k=1}^{n} (-1)^{i+k} a_{ik} |A_{ik}|
\]

In the expression, \( A_{ij} \) is a \((n-1) \times (n-1)\) submatrix of \( A \) obtained by removing row \( i \) and column \( j \) from \( A \). \(|A_{ik}|\) is called minor.

**DEFINITION 9-2.** The \((i, j)\)th **minor** of \( A \) (**minor** of \( a_{ij} \)) is the determinant of the submatrix \( A_{ij} \), and is denoted by \( m_{ij} = \det A_{ij} \). The \((i, j)\)th **cofactor** of \( A \) (**cofactor** of \( a_{ij} \)) is \( c_{ij} = (-1)^{i+j} m_{ij} \).

For \( n = 1 \), the minor and the cofactor of \( a_{11} \) are one when \( a_{11} \neq 0 \), and the minor and the cofactor of \( a_{11} \) are zero when \( a_{11} = 0 \).

**THEOREM 9-8.** Let \( A \) be an \( n \times n \) matrix \( A \) with cofactors \( c_{ij} \). Then,

\[
\det A = a_{1j} c_{1j} + \cdots + a_{nj} c_{nj} = \sum_{i=1}^{n} a_{ij} c_{ij} = \text{col}_j C \cdot \text{col}_j A
\]

**Proof.** Since \( c_{ij} = (-1)^{i+j} m_{ij} = (-1)^{i+j} |A_{ij}| \) and
\[ \det A = (-1)^{i+1}a_{i1}|A_{i1}| + \cdots + (-1)^{i+n}a_{in}|A_{in}| \]

replacing terms yields \( \det A = a_{i1}c_{i1} + \cdots + a_{in}c_{in} \). \( \square \)

Given the cofactor matrix, the determinant is the dot product of columns with the same index in the cofactor and the matrix \( A \), or the rows.

**Example 9-4.** Let

\[
A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}
\]

\[
m_{11} = \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 4, \quad m_{12} = \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 3
\]

\[
m_{21} = \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 2, \quad m_{22} = \det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 1
\]

\[
c_{11} = (-1)^{1+1} \times m_{11} = 4, \quad c_{12} = (-1)^{1+2} \times m_{12} = -3
\]

\[
c_{21} = (-1)^{2+1} \times m_{21} = -2, \quad c_{22} = (-1)^{2+2} \times m_{22} = 1
\]

\[
\det A = a_{11} \cdot c_{11} + a_{12} \cdot c_{12} = (1 \times 4) + (2 \times (-3)) = -2
\]

A direct but tedious way to construct the inverse matrix is to compute every element by Cramer’s rule. For instance,

\[
\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

\[
b_{11} = \frac{\det \begin{bmatrix} 0 & 2 \\ 3 & 4 \end{bmatrix}}{\det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}} = \frac{4}{-2} = -2, \quad b_{21} = \frac{\det \begin{bmatrix} 1 & 1 \\ -2 & 0 \end{bmatrix}}{\det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}} = \frac{-3}{-2} = \frac{3}{2}, \quad b_{12} = 1, \quad b_{22} = -\frac{1}{2}
\]

The result can be generalized. Let \( A \) be an \( n \times n \) matrix and \( B \) be its inverse. Since \( AB = I \), applying Laplace formula to the numerator gives

\[
b_{kj} = \frac{\det(A_1, \cdots, A_{k-1}, e_j, A_{k+1}, \cdots, A_n)}{\det A} = (-1)^{j+k} |A_{jk}| \]

**Definition 9-3.** Given an \( n \times n \) matrix \( A \) with cofactors \( c_{ij} \), the **adjoint** (adjugate, adjunct) of \( A \) is the \( n \times n \) matrix with \( (i,j)^{th} \) element \( c_{ji} \), and denoted by \( A^+ \) or \( \text{adj} \ A \)
THEOREM 9-9. Let $A$ be an $n \times n$ matrix $A$ with the adjoint matrix $A^+$. Then,

$$AA^+ = A^+A = \det(A) I \quad \text{and} \quad A^{-1} = \frac{1}{\det A} \cdot A^+$$

**Proof.** By Laplace formula, \(\det A = (-1)^{i+1} a_{i1} |A_{i1}| + \cdots + (-1)^{i+n} a_{in} |A_{in}|\) for all $i$. Then

$$(-1)^{i+1} a_{k1} |A_{k1}| + \cdots + (-1)^{i+n} a_{kn} |A_{kn}| = \begin{cases} 0, & \text{for } i \neq k \\ \det A, & \text{for } i = k \end{cases}$$

Because if $i \neq k$, then it is the same as to calculate the determinant of a matrix that have two same rows so that the determinant is zero. Therefore,

$$\det A = \text{col}_i C \cdot \text{col}_i A = \text{row}_i A^+ \cdot \text{col}_i A \quad \text{and} \quad \text{col}_i C \cdot \text{col}_k A = \text{row}_i A^+ \cdot \text{col}_k A = 0 \quad \blacksquare$$
10 Eigenvalues and Eigenvectors

10.1 Eigenvalue

An eigenvalue of square matrix $A$ is a number $\lambda$ such that the following equation has a non-zero solution.

$$Ax = \lambda x$$

Considering $A$ as a linear map, $Ax$ is the image of $x$ under $A$. In general, applying a linear transformation usually tilts an input vector, but in this case, the condition selects those vectors that the transformation simply changes the vector only by a scalar factor $\lambda$.

Rearranging the terms shows that the equation defines the null space of $(A - \lambda I)$.

$$(A - \lambda I)x = 0$$

In order for the matrix $A - \lambda I$ to have a non-trivial null space, we should have $\det(A - \lambda I) = 0$. The element of the null space is called the eigenvector of $A$ associated with the eigenvalue $\lambda$.

**Definition 10-1.** $\lambda$ is an **eigenvalue** of an $n \times n$ matrix $A$ if

$$\det(A - \lambda I) = 0.$$ 

**Definition 10-2.** $\det(A - \lambda I) = 0$ is called the **characteristic equation** of $A$, and $\det(A - \lambda I)$ is called the **characteristic polynomial** of $A$.

**Example 10-1.**

$$A = \begin{bmatrix} 3 & 6 & -8 \\ 0 & 0 & 6 \\ 0 & 0 & 2 \end{bmatrix}$$

Since the eigenvalues are the solution to $\det(A - \lambda I) = 0$, the first step is to find the expression

$$A - \lambda I = \begin{bmatrix} 3 & 6 & -8 \\ 0 & 0 & 6 \\ 0 & 0 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} 3 - \lambda & 6 & -8 \\ 0 & -\lambda & 6 \\ 0 & 0 & 2 - \lambda \end{bmatrix}$$
\[
\det(A - \lambda I) = [(3 - \lambda)(-\lambda)(2 - \lambda) + (6)(6)(0) + (0)(0)(-8)] \\
-[(0)(-\lambda)(-8) + (0)(6)(3 - \lambda) + (0)(6)(2 - \lambda)] \\
= (3 - \lambda)(-\lambda)(2 - \lambda)
\]

Therefore, \(\lambda = (0, 2, 3)\).

**Theorem 10-1.** Let \(\lambda_1, \ldots, \lambda_n\) be the eigenvalues of \(A\) and \(B = aA^k + bl\) with \(a, b \in \mathbb{R}\). Then eigenvalues of \(B\) are \(a^n\lambda_1^k + b\).

**Proof.** It is sufficient to show that the result holds for \(aA\), \(A^k\), and \(A + bl\). The result follows from
\[
\begin{align*}
\det(aA - a\lambda I) &= a^n \det(A - \lambda I) \\
\det(A^k - \lambda^k I) &= \det \left( (A - \lambda I)(A^{k-1} + A^{k-2}\lambda + \cdots + A\lambda^{k-2} + \lambda^{k-1}) \right) \\
\det((A + bl) - (\lambda + b)I) &= \det(A - \lambda I)
\end{align*}
\]

**Theorem 10-2.** The characteristic polynomial of \(A\) is an \(n^{th}\) degree polynomial which may be written \(\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)\), where the \(\lambda_i\) depend on the elements of \(A\). Therefore, the characteristic equation has \(n\) solutions \(\lambda_1, \ldots, \lambda_n\). When \(A\) is real, complex eigenvalues must come in complex conjugate pairs.

**Proof.** Suppose that \(\lambda\) and \(x\) are a complex eigenvalue and corresponding eigenvector of real matrix \(A\) so that \(Ax = \lambda x\). Taking complex conjugates of this equation, we have \(\bar{A}x = \bar{\lambda}x\). Since \(A\) is real, \(\bar{A}x = \bar{\lambda}x\), and thus \(\bar{\lambda}\) and \(\bar{x}\) are the eigenvalue and eigenvector of \(A\).

**Example 10-2.** When \(A\) is \(2 \times 2\), the characteristic equation is
\[
\lambda^2 + \text{tr}(A)\lambda + \det(A) = (\lambda - \lambda_1)(\lambda - \lambda_2) = 0
\]
Therefore \(\lambda_1 + \lambda_2 = \text{tr}(A)\) and \(\lambda_1 \cdot \lambda_2 = \det(A)\)

If \(\text{tr}(A)^2 - 4\det(A) < 0\), the root is complex:
\[
\lambda = a \pm \sqrt{-1}b, \quad \text{tr}(A) = 2a, \quad \det(A) = a^2 + b^2
\]

**Theorem 10-3.** \(\det(A) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n\) and \(\text{tr}(A) = \lambda_1 + \cdots + \lambda_n\), where \(\lambda_1, \ldots, \lambda_n\) are the eigenvalues of \(A\).

**Proof.** The proof uses the fact that, in the characteristic polynomial, \(\det(A)\) is a constant and \(-\text{tr}(A)\) is the coefficient of \(\lambda^{n-1}\).
Let \( P(\lambda) = \det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda) \). \( P(0) = \det(A) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n \).

For the second part, we will show that the coefficient of \( \lambda^{n-1} \) is \(-\text{tr}(A)\). Firstly, it is trivial that its coefficient is \(- (\lambda_1 + \cdots + \lambda_n)\) because \( P(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda) \).

Secondly, by the definition of the determinant, the product of all the principal diagonal terms must be included in the expression of the determinant.

\[
(a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda)
\]

And if \( a_{ij} \) with \( i \neq j \) is included, \((a_{ii} - \lambda)\) and \((a_{jj} - \lambda)\) should be excluded from the expression of a term. Thus, every other possible product can contain at most \( n - 2 \) elements on the diagonal of \( A \) and contains at most \((n-2)\) of \( \lambda \)'s. Hence \( P(\lambda) \) can be written as

\[
P(\lambda) = (a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda) + q(\lambda),
\]

where \( q(\lambda) \) is a polynomial of degree \((n-2)\). Hence, the term of \( \lambda^{n-1} \) should be from \((a_{11} - \lambda) \cdots (a_{nn} - \lambda)\), and that is \(- (a_{11} + a_{22} + \cdots + a_{nn})\). Therefore, \( \text{tr}(A) = \lambda_1 + \cdots + \lambda_n \).

**COROLLARY 10-4.**

(i) \( A \) is nonsingular if and only if all its eigenvalues are nonzero.

(ii) \( A \) and \( A^T \) have the same characteristic polynomial and the same eigenvalues.

(iii) If \( A \) is triangular, its diagonal elements are its eigenvalues. If \( A \) is block triangular, the eigenvalues of the blocks along the diagonal are the eigenvalues of \( A \).

### 10.2 EIGENVECTOR

**DEFINITION 10-3.** Given a matrix \( A \) and an eigenvalue \( \lambda \) of \( A \), a nonzero vector \( x \) is an **eigenvector** of \( A \) corresponding to \( \lambda \) if and only if

\[
Ax = \lambda x.
\]

And the pair \((\lambda, x)\) is called **eigenpair**.

Since \( \det(A - \lambda I) = 0 \), \((A - \lambda I)\) is singular and \((A - \lambda I)x = 0\) has a non-zero null space. Therefore, eigenvectors cannot be uniquely determined and might not be real.
**Example 10-3.** Eigenvector is a solution to \((A - \lambda I)x = 0\). Since it is not unique, we take a vector that is easy to deal with.

First, for \(\lambda = 0\), \((A - \lambda I)x = Ax = 0\) and

\[
\begin{bmatrix}
3 & 6 & -8 \\
0 & 0 & 6 \\
0 & 0 & 2
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = 0
\]

From the third equation, \(x_3 = 0\). Since the first equation is \(3x_1 + 6x_2 = 0\),

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

For \(\lambda = 2\), \((A - 2I)x = 0\) and

\[
\begin{bmatrix}
3 & 6 & -8 \\
0 & 0 & 6 \\
0 & 0 & 2
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = \begin{bmatrix}
1 & 6 & -8 \\
1 & 6 & -8 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = 0
\]

\[
\begin{bmatrix}
1 & 6 & -8 \\
0 & 2 & 6 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 10 \\
0 & 1 & -3 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
\]

Since \(x_3\) could be anything but 0,

\[
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = \begin{bmatrix} -10 \\ 3 \\ 1 \end{bmatrix}
\]

Finally, for \(\lambda = 3\), \((A - 3I)x = 0\) and

\[
\begin{bmatrix}
3 & 6 & -8 \\
0 & 0 & 6 \\
0 & 0 & 2
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = \begin{bmatrix}
3 & 0 & 0 \\
0 & 3 & 0 \\
0 & 0 & 3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = \begin{bmatrix} 0 & 6 & -8 \\
0 & 0 & 0 \\
0 & 0 & -1
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix} = 0
\]

The last equation shows \(x_3 = 0\), and in turn the second gives \(x_2 = 0\).

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

**Example 10-4.** A matrix might not have real eigenvalues and eigenvector. The following matrix does not have a real eigenvalue for \(\sin \theta \neq 0\).
Let $A$ be an $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigenvectors $x_1, \ldots, x_n$. Also, $A = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and $X = [x_1, \ldots, x_n]$ so that $AX = X\Lambda$.

**Theorem 10-5.** If $A$ is nonsingular, the eigenvectors of $A^{-1}$ are the same as those of $A$; and the eigenvalues of $A^{-1}$ are the inverses of those of $A$.

**Proof.** Since $A$ is invertible, every eigenvalue is non zero and so is eigenvectors.

$$AX = \lambda X \iff A(\lambda^{-1}X) = X \iff (\lambda^{-1}X) = A^{-1}X \quad \Box$$

**Theorem 10-6.** $A$ is an idempotent matrix if and only if the eigenvalues of $A$ are zeros or ones. If $A$ is symmetric, then the rank of an idempotent matrix is $\text{tr}(A)$.

**Proof.** Let $\lambda$ be an eigenvalue of $A$ and $x$ be a corresponding eigenvector. Then,

$$\lambda x = Ax = AAx = A(\lambda x) = \lambda Ax = \lambda^2 x$$

$$(\lambda - \lambda^2)x = 0 \Rightarrow \lambda(1 - \lambda)x = 0$$

Therefore, $\lambda = 0$ or 1 because an eigenvector is never a zero vector. This also shows that $\det(A)$ is either 0 or 1, and if $A$ is nonsingular, $A = I$.

Since the eigenvalues of an idempotent matrix are zeros or ones and a symmetric matrix is diagonalizable, there exists a matrix $Q$ such that

$$Q^TQ = I \quad \text{and} \quad Q^T AQ = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

Therefore, its rank is $\lambda_1 + \lambda_2 + \cdots + \lambda_n = \text{tr}(A)$. \quad \Box

**10.3 Diagonalization**

Consider a geometric sequence such that $x_t = ax_{t-1}$ with a given $x_0$. Then the limit can be characterized by the equation $x_t = a^t x_0$. It is easy to show that $x_t \to 0$ independent of $x_0$ if $|a| < 1$. Now consider $x_t = Ax_{t-1}$ with $x \in \mathbb{R}^n$. It is usually not easy to characterize the behavior of $A^t$. In this situation, the notion of conjugacy could help. Two elements $f$ and $g$
are conjugate if there exists an element \( h \) in the group such that \( f = h^{-1}gh \). \( f \) is said to be a conjugate of \( g \) and vice versa. Then

\[
f^2 = f \circ f = h^{-1} \circ g \circ h^{-1} \circ g \circ h = h^{-1} \circ g^2 \circ h
\]

The conjugacy can be applied to any kind of groups as long as the inverse is well-defined such as the group of invertible matrices. In this case, the conjugacy relation is called matrix similarity.

**10.3.1 Eigen Decomposition**

**Definition 10-4.** \( A \) and \( B \) are **similar** if there exists a nonsingular matrix \( P \) such that

\[
B = P^{-1}AP.
\]

**Theorem 10-7.** Similar matrices have the same characteristic polynomial and thus the same eigenvalues.

**Proof.**

\[
det(B - \lambda I) = det(P^{-1}AP - \lambda I) = det(P^{-1}[A - \lambda I]P) = det(P^{-1}) det(A - \lambda I) det(P) = det(A - \lambda I)
\]

**Definition 10-5.** A square matrix is **diagonalizable** if it is similar to a diagonal matrix.

If there exists an invertible matrix \( P \) such that \( P^{-1}AP \) is a diagonal matrix. This process is known as the **eigendecomposition** of \( A \).

**Example 10-5.** Diagonalization is extremely useful for matrix power and matrix decomposition.

\[
A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}
\]

Let \( P = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \), then

\[
P^{-1}AP = \left( \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}
\]

Therefore,
\[
(P^{-1}AP)^4 = \begin{bmatrix}
1/2 & -1/2 \\
1/2 & 1/2
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
-1 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 81
\end{bmatrix} = \begin{bmatrix}
1 & 0 \\
81 & 1
\end{bmatrix}
\]
\[
A^4 = \begin{bmatrix}
1/2 & 1 \\
-1 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
0 & 81
\end{bmatrix}
\begin{bmatrix}
1 & -1 \\
1 & 1
\end{bmatrix} = \begin{bmatrix}
41 & 40 \\
40 & 41
\end{bmatrix}
\]

**Theorem 10-8.** Suppose \( n \times n \) matrix \( A \) has \( n \) linearly independent eigenvectors with eigenvalues \( \lambda_1, \ldots, \lambda_n \), and \( A = \text{diag}(\lambda_1, \ldots, \lambda_n) \). Then, there exists a nonsingular matrix \( P \) such that
\[
P^{-1}AP = \Lambda
\]
and \( P = X \). Conversely, the existence of such \( P \) implies that \( A \) has \( n \) linearly independent eigenvectors.

**Proof.** Since \( \text{rank}(X) = n \), \( X^{-1} \) exists.
\[
X \times \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = [\lambda_1 x_1, \ldots, \lambda_n x_n] = [Ax_1, \ldots, Ax_n] = AX.
\]
Therefore, \( \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) = X^{-1}AX \).

The converse is also true because \( X \) is invertible. \( \square \)

**Example 10-6.** Let’s eigendecompose matrix \( A \) in Example 10-5.

From \( \det\begin{bmatrix}
1 - \lambda & 2 \\
2 & 1 - \lambda
\end{bmatrix} = 0 \), the two eigenvalues of \( A \) are \( \lambda = -1 \) or \( 3 \). The eigenvectors are the solutions to
\[
\begin{bmatrix}
1 & 2 \\
2 & 1
\end{bmatrix}\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = -\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}
1 & 2 \\
2 & 1
\end{bmatrix}\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = 3\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\]
Or \( \begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = \begin{bmatrix}
1 \\
-1
\end{bmatrix} \) and \( \begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} = \begin{bmatrix}
1 \\
1
\end{bmatrix} \). Therefore, we have \( P = \begin{bmatrix}
1 & 1 \\
-1 & 1
\end{bmatrix} \), and it worked as we expected. \( \square \)

Be careful with the order of eigenvalues and eigenvectors. When you construct matrix \( P \), the first column is the eigenvector associated with the first eigenvalue in the diagonal matrix \( \Lambda \) and the same is for the second.

**10.3.2 Sufficient Conditions**

**Theorem 10-9.** For any \( A \) with \( n \) distinct eigenvalues (no two equal), \( A \) has \( n \) linearly independent eigenvectors.
Proof. Suppose they are linearly dependent. Without loss of generality, assume that \( x_1, \ldots, x_{k-1} \) are linearly independent but \( x_k \) is a linear combination of the first \( k-1 \) vectors. Or there exists non zero \( \alpha_i \)'s such that

\[
\sum_{i=1}^{k-1} \alpha_i x_i = x_k,
\sum_{i=1}^{k-1} \alpha_i Ax_i = Ax_k,
\sum_{i=1}^{k-1} \alpha_i \lambda_i x_i = \lambda_k x_k
\]

Multiply the first equation by \( \lambda_k \) and subtract it from the last equation to obtain

\[
\sum_{i=1}^{k-1} \alpha_i (\lambda_i - \lambda_k) x_i = 0
\]

Since \( \lambda_i - \lambda_k \neq 0 \) for all \( i = 1, \ldots, k-1 \), this implies that a nontrivial linear combination of linearly independent vectors is zero.

The condition is only sufficient. The identity matrix has \( n \) identical eigenvalues but the eigenvectors are linearly independent.

Example 10-7. Consider a Markov chain \( x_t = Ax_{t-1} \). Then

\[
x_t = Ax_{t-1} = A^2 x_{t-2} = \cdots = A^t x_0
\]

If eigenvectors are linearly independent, \( A \) is diagonalizable and

\[
AX = XA, \quad A = XAX^{-1}, \quad A^t = XA^t X^{-1}
\]

Since \( X^{-1} x_t = A^t X^{-1} x_0 \) or \( X_t = \Lambda X_0 \), we have \( X_{ti} = \lambda_{ti} X_{0i} \).

Example 10-8. Real eigenvectors do not exist for some matrices, and if eigenvectors are linearly dependent, they are not diagonalizable. If a real matrix is symmetric, it is diagonalizable.

Let's first find eigenvalues and eigenvectors.

\[
\lambda^2 - \lambda - 2 = 0, \quad \lambda = (-1, 2)
\]

\[
Ax = (x_1 + 4x_2, 0.5x_1) = (-x_1, -x_2), \quad x = (2, -1)
\]
\[ Ax = (x_1 + 4x_2, 0.5x_1) = (2x_1, 2x_2), \quad x = (4,1) \]

Since eigenvectors are linearly independent, it is orthogonally diagonalizable.

\[
A = \begin{bmatrix}
1/2 & 4 \\
-1 & 1
\end{bmatrix} = \begin{bmatrix}
2 & 4 \\
-1 & 0
\end{bmatrix}\begin{bmatrix}
2 & 4 \\
-1 & 0
\end{bmatrix}^{-1} = \begin{bmatrix}
2 & 4 \\
-1 & 0
\end{bmatrix}\begin{bmatrix}
1/6 & -4/6 \\
1/6 & 2/6
\end{bmatrix}
\]

\[
[\begin{bmatrix}
x_{t+1} \\
y_{t+1}
\end{bmatrix}] = \begin{bmatrix}
-1 & 0 \\
2 & 1
\end{bmatrix}\begin{bmatrix}
x_t \\
y_t
\end{bmatrix}, \quad [\begin{bmatrix}
x_t \\
y_t
\end{bmatrix}] = \begin{bmatrix}
1/6 & -4/6 \\
1/6 & 2/6
\end{bmatrix}\begin{bmatrix}
x_0 \\
y_0
\end{bmatrix}
\]

**THEOREM 10-10.** Any symmetric matrix \( A \) has an eigenvector.

**Proof.** (The proof uses multivariate optimization condition.) Let \( Q_A(x) = x^T A x \), which is a polynomial in \( x \) and thus differentiable.

\[ \nabla Q_A(x) = 2x^T A \]

Let \( S = \{ x \in \mathbb{R}^n \mid \| x \| = 1 \} \) be an unit sphere consisting of unit vectors. Since \( S \) is closed and bounded, the optimization problem, \( \max_{x \in S} Q_A(x) \), has an interior solution. Since the constraint is \( g(x) = x_1^2 + \cdots + x_n^2 = 1 \), the first order condition is

\[ \nabla Q_A(x) = \lambda \nabla g(x), \quad 2x^T A = \lambda 2x^T \]

where \( \lambda \in \mathbb{R} \) is a Lagrange multiplier. Since \( A \) is symmetric, \( A x = \lambda x \). \( \blacksquare \)

**THEOREM 10-11.** Let \( A \) be a symmetric real \( n \times n \) matrix. Then there exists a real matrix \( Q \) such that

\[ Q^T A Q = \Lambda \]

for some orthogonal matrix \( Q \). Conversely, if there exists a matrix \( Q \) such that \( Q^T A Q = \Lambda \), then it is symmetric.

**Proof.** We prove it by induction. When \( n = 1 \), it is obvious. Suppose the theorem holds for \( (n-1) \times (n-1) \), and let \( A \) be a \( n \times n \) symmetric matrix. By **THEOREM 10-10**, there exists an eigenvector

\[ A x = \lambda_1 x_1 \]
Then we can construct an orthonormal basis (using Gram Schmidt process) $S_1 = (\mathbf{x}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n)$ of $\mathbb{R}^n$ such that $S_1^T S_1 = I$. Then $S_1^T A S_1$ is symmetric and its first column can be expressed as $S_1^T A S_1 e_1$. Since $S_1 e_1 = \mathbf{x}_1$,

$$S_1^T AS_1 e_1 = S_1^T A \mathbf{x}_1 = S_1^T \lambda_1 \mathbf{x}_1 = \lambda_1 S_1^T \mathbf{x}_1 = \lambda_1 e_1.$$ 

Therefore, we can write

$$S_1^T A S_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}$$

where $A_1$ is a $(n-1) \times (n-1)$ symmetric matrix. By the induction hypothesis, there exists an $(n-1) \times (n-1)$ orthogonal matrix $S_2$ such that $S_2^T A_1 S_2$ is a $(n-1) \times (n-1)$ diagonal matrix. Let

$$\tilde{S}_2 = \begin{bmatrix} 1 & 0 \\ 0 & S_2 \end{bmatrix},$$

which is an $n \times n$ orthogonal matrix, and

$$S_2^T S_1^T A S_1 \tilde{S}_2 = \begin{bmatrix} 1 & 0 \\ 0 & S_2^T \\ S_2^T & S_2^T \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & S_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & S_2^T \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & S_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & S_2^T A_1 S_2 \end{bmatrix}.$$ 

Therefore, $Q = S_1 \tilde{S}_2$ is an orthogonal matrix, and $Q^T AQ = \begin{bmatrix} \lambda_1 & 0 \\ 0 & S_2^T A_1 S_2 \end{bmatrix}$ is a diagonal matrix. 

**Theorem 10-12.** Eigenvectors of distinct eigenvalues of a symmetric real matrix are orthogonal.

Let $A \mathbf{x}_1 = \lambda_1 \mathbf{x}_1$ and $A \mathbf{x}_2 = \lambda_2 \mathbf{x}_2$ with $\mathbf{x}_1$ and $\mathbf{x}_2$ are non-zero vectors. Pre-multiplying both sides of the first equation with $\mathbf{x}_2^T$ gives

$$\lambda_1 \mathbf{x}_2^T \mathbf{x}_1 = \mathbf{x}_2^T (A \mathbf{x}_1) = (A \mathbf{x}_2)^T \mathbf{x}_1 = (A \mathbf{x}_2) \mathbf{x}_1^T \mathbf{x}_1 = \lambda_2 \mathbf{x}_2^T \mathbf{x}_1$$

Therefore, $(\lambda_1 - \lambda_2) \mathbf{x}_2^T \mathbf{x}_1 = 0$ and $\lambda_1 \neq \lambda_2$ implies $\mathbf{x}_2^T \mathbf{x}_1 = 0$. 

**Theorem 10-13.** If $A$ is real and symmetric, its eigenvalues are real and its rank equals the number of nonzero eigenvalues.

**Proof.** Suppose eigenvalues are not real. Let conjugate pairs of $\lambda$ and $\bar{\lambda}$ be the eigenvalues of $A$. Since $A \mathbf{x} = \lambda \mathbf{x}$ and $A \bar{\mathbf{x}} = \bar{\lambda} \bar{\mathbf{x}}.$
\[ \bar{x}^T Ax = \lambda \bar{x}^T \bar{x}, \quad x^T A \bar{x} = \bar{\lambda} x^T \bar{x} \]

Subtracting one from the other gives
\[ \bar{x}^T A x - x^T A \bar{x} = (\lambda - \bar{\lambda}) x^T \bar{x}, \quad (\lambda - \bar{\lambda}) x^T \bar{x} = 0 \]

Since \( x^T \bar{x} = \|x\| \), it is strictly positive unless \( x = 0 \) and \( \lambda = \bar{\lambda} \).

Let’s check the rank. By THEOREM 10-10, \( A \) is diagonalizable. Therefore,
\[ \text{rank}(A) = \text{rank}(P^{-1} \Lambda P) = \text{rank}(\Lambda) = \text{the number of nonzero eigenvectors} \]

**THEOREM 10-14.** For any \( A \) and any positive constant \( \varepsilon \), there exists a nonsingular matrix \( P \) such that:

(i) \( P^{-1} AP \) is triangular.

(ii) \( P^{-1} AP \) has the eigenvalues of \( A \) on its diagonal.

(iii) \( P^{-1} AP \) has each above-diagonal element less than \( \varepsilon \) in modulus.

The intuition behind the theorem is that even if \( A \) has some repeated eigenvalues, perturbing \( A \) slightly should make eigenvalues distinct.

### 10.3.3 Matrix Norm

Recall the vector difference equation \( x_t = Ax_{t-1} \) that we have examined in the head of this subsection. Another way to characterize the behavior of \( A_t \) is using a matrix norm of \( A \) as an analogy of the absolute value of the coefficient in the scalar version, \( x_t = ax_{t-1} \).

Matrix norm is defined almost parallel to the vector norm.

**DEFINITION 10-6.** Let \( f \) be a real valued scalar function of a square matrix. \( f \) is a **matrix norm** if and only if, for any scalar \( c \) and any \( n \times n \) matrix \( A \) and \( B \), the following axioms hold.

(i) \( f(A) \geq 0 \)

(ii) \( f(A) = 0 \) if and only if \( A = 0 \).
(iii) \( f(A + B) \leq f(A) + f(B) \)

(iv) \( f(cA) = |c|f(A) \)

(v) \( f(AB) \leq f(A) \cdot f(B) \), sub-multiplicative property

The form of the property (v) resembles Cauchy-Schwarz inequality. The induced norm below is derived from a vector norm and satisfies all the properties. Otherwise, property (v) should be assumed.

**Theorem 10-15.** Let \( A \) be a \( n \times n \) matrix. Each of the following is a matrix norm.

(i) Induced norm: \( \max\{\|Ax\|: x \in \mathbb{R}^n \text{ with } \|x\| = 1\} \)

(ii) Maximum element norm: \( n \max_{i,j} |a_{ij}| \)

(iii) Holder norm: \( \left[ \sum_i \sum_j |a_{ij}|^q \right]^{1/q} \) for \( 1 \leq q \leq 2 \)

(iv) Euclidean norm: \( \left[ \sum_i \sum_j |a_{ij}|^2 \right]^{1/2} \)

(v) Element sum norm: \( \sum_i \sum_j |a_{ij}| \)

(vi) Column sum norm: \( \max_j \sum_i |a_{ij}| \)

(vii) Row sum norm: \( \max_i \sum_j |a_{ij}| \)

(viii) Weighted norm: \( f(A) = h(P^{-1}AP) \) where \( h \) is any matrix norm and \( P \) is any nonsingular matrix.

In case of the induced norm, we can see clearly that \( f(A) < 1 \) implies \( \|Ax\| < \|x\| \). In the difference equation \( \mathbf{x}_t = A\mathbf{x}_{t-1} \), therefore, any nonzero point is pulled closer to the origin.

Alternatively, we can check the condition with the eigenvalues.

**Theorem 10-16.** Let \( A \) be an \( n \times n \) matrix with eigenvalues of \( \lambda_1, \ldots, \lambda_n \) and let \( f \) be a any matrix norm function. Then \( |\lambda_i| \leq f(A) \) for all \( i = 1, \ldots, n \).
**Proof.** Let \( \lambda \) denote an eigenvalue of \( A \) of largest modulus and \( x \) be a corresponding eigenvector, and \( B \) be a square matrix \( B = [x, 0, \cdots, 0] \). Then \( Ax = \lambda x \) and \( AB = \lambda B \). These equations, together with norm axioms, yield:

\[
|\lambda| = \frac{f(B)}{f(B)} = \frac{f(\lambda B)}{f(B)} = \frac{f(A)f(B)}{f(B)} = f(A)
\]

**10.3.4 Matrix Power**

To check the stability of a linear dynamic system, it is necessary to know the limit behavior of matrix power and matrix exponential. This section presents some of the basic results.

**Theorem 10-17 (Matrix Power)** For an eigenvalue \( \lambda \) of \( A \) and an associated eigenvector \( x \),

\[
A^k x = \lambda^k x
\]

for any positive integer \( k \).

**Proof.** We prove it by induction. For \( k = 1 \), \( Ax = \lambda x \). Suppose that \( A^{k-1} x = \lambda^{k-1} x \) is true.

\[
\lambda^k x = \lambda \times \lambda^{k-1} x = \lambda \times A^{k-1} x = A^{k-1} (\lambda x) = A^{k-1} (Ax) = A^k x
\]

**Theorem 10-18.** The following statements about an \( n \times n \) matrix \( A \) are equivalent.

(i) \( A^t \rightarrow 0 \cdot I \) as \( t \rightarrow \infty \) (\( t \) integer)

(ii) The eigenvalues of \( A \) are all less than one in modulus.

(iii) The series \( I + A + A^2 + \cdots \) converges to \((I - A)^{-1}\).

**Proof.** (i) \( \Rightarrow \) (ii) By taking the limit on \( A^k x = \lambda^k x \), we have

\[
\left( \lim_{k \to \infty} A^k \right) x = x \lim_{k \to \infty} \lambda^k \text{ or } 0 \times x = x \lim_{k \to \infty} \lambda^k
\]

(ii) \( \Rightarrow \) (i) is immediate from Theorem 10-17.

(i) \( \Rightarrow \) (iii)
\[ S_k = I + A + A^2 + \cdots + A^{k-1} \]
\[ AS_k = A + A^2 + A^3 + \cdots + A^k \]
\[ (I - A)S_k = I - A^k \]

Since \( \lim_{k \to \infty} A^k = 0 \cdot I \), \( \lim_{k \to \infty} (I - A)S_k = (I - A) \lim_{k \to \infty} S_k = I \). Since \( \det(I - A) \neq 0 \),
\[ \lim_{k \to \infty} S_k = (I + A)^{-1} \]

(iii)\(\Rightarrow\)(i) Note that \( (I - A)S_k = I - A^k \) for any \( A \). Since \( \lim_{k \to \infty} (I - A)S_k = I \), \( \lim_{k \to \infty} A^k = 0 \cdot I \).

\[ \text{DEFINITION 10-7.} \] The matrix exponential \( e^A \) is defined as the infinite series
\[ e^A = I + A + \frac{A^2}{2!} + \cdots + \frac{A^r}{r!} + \cdots \]

\[ \text{EXAMPLE 10-9.} \] Let’s compute \( B = e^{θA} \) where
\[ A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \]
\[ A^2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A^4 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \]

Therefore, \( A^{4n+r} = A^r \).
\[ b_{11} = 1 + \frac{0}{1!}\theta^1 + \frac{(-1)}{2!}\theta^2 + \frac{0}{3!}\theta^3 + \frac{1}{4!}\theta^4 + \frac{0}{5!}\theta^5 + \cdots = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots = \cos \theta \]
\[ b_{12} = 0 + \frac{1}{1!}\theta^1 + \frac{0}{2!}\theta^2 + \frac{(-1)}{3!}\theta^3 + \frac{0}{4!}\theta^4 + \frac{1}{5!}\theta^5 + \cdots = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots = \sin \theta \]

Similarly, \( b_{21} = -\sin \theta \) and \( b_{22} = \cos \theta \) so that
\[ e^{θA} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \]

\[ \text{THEOREM 10-19.} \] Let \( A \) and \( P \) be \( n \times n \) matrices and \( P \) be invertible. Then,
\[ e^{P^{-1}AP} = P^{-1}e^AP \]

\[ \text{Proof.} \]
\[e^{P^{-1}AP} = I + P^{-1}AP + \frac{(P^{-1}AP)^2}{2!} + \frac{(P^{-1}AP)^3}{3!} + \ldots\]
\[= I + P^{-1}AP + \frac{P^{-1}A^2P}{2!} + \frac{P^{-1}A^3P}{3!} + \ldots\]
\[= P^{-1}\left(I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots\right)P = P^{-1}e^{AP}\]

**Theorem 10-20.** Let \(A\) and \(B\) be \(n \times n\) matrices. Then:

(i) The series \(e^A\) converges for any \(A\).

(ii) If \(AB = BA\), then \(e^{A+B} = e^Ae^B = e^Be^A\).

(iii) \(e^A\) is nonsingular, and \((e^A)^{-1} = e^{-A}\).

(iv) \(\lambda\) is an eigenvalue of \(A\) if and only if \(e^\lambda\) is an eigenvalue of \(e^A\).

(v) \(\det(e^A) = e^{\text{tr}(A)}\).

(vi) \(e^{At} \to 0\) as \(t \to \infty\) (real \(t\)) if and only if all eigenvalues of \(A\) have negative real parts.

**Proof.** (i) We will show the absolute convergence that, for any matrix norm,

\[1 + \|A\| + \left\|\frac{A^2}{2!}\right\| + \left\|\frac{A^3}{3!}\right\| + \ldots\]

converges. By DEFINITION 10-6 (v), we have

\[0 < \left\|\frac{A^n}{n!}\right\| = \left\|\frac{A^n}{n!}\right\| \leq \frac{\|A\|^n}{n!}\]

Therefore,

\[1 + \|A\| + \left\|\frac{A^2}{2!}\right\| + \ldots = e^{\|A\|}\]

Since \(e^{\|A\|}\) is finite, the series converges.

(ii) The Cauchy product of the infinite series of \(e^A\) and \(e^B\) is

\[e^Ae^B = \sum_{n=0}^{\infty} \frac{A^n}{n!} \times \sum_{n=0}^{\infty} \frac{B^n}{n!} = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{A^k}{k!} \frac{B^{n-k}}{(n-k)!}\]
\[
= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} A^k B^{n-k} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k}
\]

Since \(AB = BA\), we have

\[
(A + B)^n = \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k}
\]

Therefore,

\[
e^A e^B = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k} = \sum_{n=0}^{\infty} \frac{(A + B)^n}{n!} = e^{A+B}
\]

(iii) By (ii), \(e^A e^{-A} = e^{A+(-A)} = e^{0 \times I} = I\) as required.

(iv) If \(Ax = \lambda x\), then \(A^2 x = A(Ax) = A\lambda x = \lambda^2 x\). By induction, \(A^nx = \lambda^n x\)

\[
e^A x = \left(\sum_{n=0}^{\infty} \frac{A^n}{n!}\right) x = \sum_{n=0}^{\infty} \frac{A^n x}{n!} = \sum_{n=0}^{\infty} \frac{\lambda^n x}{n!} = \left(\sum_{n=0}^{\infty} \frac{\lambda^n}{n!}\right) x = e^\lambda x
\]

(v) By THEOREM 10-14 (i), there exists a \(P\) such that \(P^{-1}AP\) is a triangular matrix.

\[
det e^A = det(P^{-1}e^A P) = det e^A = e^{\lambda_1} \cdots e^{\lambda_n} = e^{\lambda_1 + \cdots + \lambda_n}
\]

(vi) If \(A\) is diagonalizable, \(e^{At} = P^{-1}e^{At} P\) and \(e^{At} = \text{diag}(\lambda_1^t, \cdots, \lambda_n^t)\).

\[\Box\]

10.4 MATRIX DECOMPOSITION

When a linear equation system \(Ax = b\) has an solution, we can transform the system into a form that requires a fewer resource for its computation. For instance, if the matrix can be decomposed as \(LU\) where \(L\) is a lower triangular matrix and \(U\) is an upper triangular matrix, we have \(LUx = b\) or \(Ux = L^{-1}b\). Eigendecomposition another good example of matrix decomposition for computationally efficiency.

DEFINITION 10-8. Let \(A\) be a square matrix. An LU factorization of \(A\) is to factorization of \(A\) with proper row and/or column permutations into two factors, a lower triangular matrix \(L\) and an upper triangular matrix \(U\).
Let’s factorize the following $2 \times 2$ matrix.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

$$1 = l_{11} \cdot u_{11}, \quad 2 = l_{11} \cdot u_{12}, \quad 3 = l_{21} \times u_{11}, \quad 4 = l_{21} \cdot u_{12} + l_{22} \cdot u_{22}$$

Note that there are 6 unknowns and 4 equations and thus the system is underdetermined. We can normalize by fixing the values of the two unknowns. Let $l_{11} = l_{22} = 1$.

$$u_{11} = 1, \quad u_{12} = 2, \quad l_{21} = 3, \quad u_{22} = -2$$

Therefore,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & -2 \end{bmatrix}$$

**Definition 10-9.** LU Factorization with partial pivoting (LUP) refers to LU factorization with row permutations only.

$$PA = LU$$

where $P$ is a row permutation matrix.

For a numerical solution of linear equation system of $Ax = b$, if $PA = LU$, we solve it with $LUx = Pb$, or $Ux = L^{-1}Pb$. This is an efficient way to calculate numerical solutions and useful in Monte Carlo simulations as well as computing inverse matrices and determinants.

**Definition 10-10.** LU factorization with full pivoting refers to LU factorization with row and column permutations.

$$PAQ = LU$$

where $P$ and $Q$ are row and column permutation matrices, respectively.

**Definition 10-11.** LDU decomposition is the decomposition of the form

$$A = LDU$$
where $D$ is a diagonal matrix and $L$ and $U$ are unit triangular matrices.

**Theorem 10-21 (Existence and Uniqueness)**

(i) Any square matrix admits a LUP factorization.

(ii) If $A$ is invertible, then it admits LU or LDU factorization if and only if its leading principal minors are nonzero. If LDU, it is unique.

(iii) If $A$ is a singular matrix of rank $k$, then it admits LU factorization if the first $k$ leading principal minors are nonzero. The converse is not true.

**Definition 10-12.** Cholesky decomposition is a decomposition of a positive definite matrix into the product of a lower triangular matrix and its transpose.

$$A = LL^T$$

Cholesky decomposition is a special case of LU decomposition, and always exists and unique.

A closely related variant is LDL decomposition of $A = LDL^T$ as a special case of LDU decomposition. Since

$$A = LDL^T = LD^{1/2}(D^{1/2})^T = LD^{1/2}(LD^{1/2})^T$$

we can found $L$ with all the diagonal elements of one. LDL decomposition avoids extracting square roots and negative entries in $D$, the LDL decomposition may be preferred in applications. This is also closely related to the diagonalization of a real symmetric matrix, $A = QAQ^T$.

**Definition 10-13.** A QR decomposition of a matrix is a decomposition of a matrix $A$ into $A = QR$ of an orthogonal matrix $Q$ and an upper triangular matrix $R$.

(i) QR decomposition of a square matrix

If $A$ is invertible, then the factorization is unique up to positive diagonal elements of $R$.

(ii) QR decomposition of a rectangular matrix

In general, a complex $m \times n$ matrix with $m \geq n$ can be factored as the product of an $m \times m$ unitary $Q$ and an $m \times n$ upper triangular matrix $R$. As the bottom $(m - n)$ rows of an $m \times n$ upper triangular matrix consist entirely of zeroes.
where \( R_1 \) is an \( n \times n \) upper triangular matrix, \( 0 \) is an \((m-n) \times n\) zero matrix, \( Q_1 \) is \( m \times n \), \( Q_2 \) is \( m \times (m-n) \), and \( Q_1 \) and \( Q_2 \) both have orthogonal columns.

Consider an underdetermined linear equation system \( A\mathbf{x} = \mathbf{b} \), where \( A \) is \( m \times n \) with \( m < n \) and rank of \( m \). The numerical solution can be calculated with the following steps.

\[
A = QR = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = [Q_1 \quad Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1
\]

In case of an overdetermined system with \( m > n \),

\[
A = QR = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = [Q_1 \quad Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1
\]

**Definition 10-14.** The singular value decomposition of an \( m \times n \) matrix is the factorization of the from:

\[
M_{m\times n} = U \Sigma V^T
\]

where

\( U \) is an \( m \times m \) orthogonal matrix such that \( U^T U = I \).

\( \Sigma \) is a diagonal \( m \times n \) matrix with non-negative real numbers on the diagonal.
$V$ is an $n \times n$ orthogonal matrix.

If $M$ is diagonalizable, the factor matrices can be constructed with eigenvalues and eigenvectors.

$$M^T M = V \Sigma^T U^T \Sigma V^T = V (\Sigma^T \Sigma) V^T$$

$$MM^T = U \Sigma V^T \Sigma U^T = U (\Sigma \Sigma^T) U^T$$

The columns of $V$ are eigenvectors of $M^T M$ and the columns of $U$ are eigenvectors of $MM^T$. The non-zero elements of $\Sigma$ are the square roots of the non-zero eigenvalues of $M^T M$ or $MM^T$. 
10.5 Orthogonal Projection

**Definition 10-15.** A projection is a linear transformation $P$ from a vector space to itself such that $P^2 = P$.

The property of $P^2 = P$ is known as idempotency. According to the definition, any idempotent matrix could be used to create a projection. Together with the orthogonality condition, we call it the orthogonal projection. In economics, we mostly use the orthogonal projection.

### 10.5.1 Orthogonal Projection

An orthogonal projection is a way to approximate a vector $y \in V$ with an element of its subspace $X \subset V$ such that

$$\hat{y} = \arg\min_{x \in X} \|y - x\|$$

**Figure 10-1. Orthogonal Projection**

**Definition 10-16.** Suppose that a vector space $V$ is a norm space. Then a projection $P$ is **orthogonal projection** if $x \cdot (y - Py) = 0$. $(y - Py)$ is called the **rejection** of $y$ on $Py$.

As in the figure, if the rejection is perpendicular to $X$, the length of rejection is minimized.

**Theorem 10-22.** Let $V$ be a vector space and $X \subset V$ be a subspace with $y \in V$ and $\hat{y} \in X$. If $\hat{y}$ solves $\min_{x \in X} \|y - x\|$, then $(y - \hat{y}) \perp x$ for all $x \in X$.

**Proof.** Since $\|y - \hat{y}\|^2 \leq \|y - x\|^2$ for all $x \in X$, the function...
$$F(t) = \|y - \hat{y} + tx\|^2$$

attains its minimum at $t = 0$ for $x \in X$, or $F'(0) = 0$. Since

$$F(t) = (y - \hat{y} + tx) \cdot (y - \hat{y} + tx) = \|y - \hat{y}\|^2 + 2t(y - \hat{y}) \cdot x + t^2\|x\|^2$$

and

$$F'(0) = 2(y - \hat{y}) \cdot x = 0$$

$(y - \hat{y}) \perp x$ for all $x \in X$. □

The proof is based on the triangular inequality depicted in Figure 10-1. The result also holds for non-differentiable norms.

**Alternative proof.** Suppose not. Then there exists a unit vector $\hat{x} \in X$ which is not orthogonal to the $y - \hat{y}$ such that $(y - \hat{y}) \cdot \hat{x} = \delta \neq 0$. Define a vector $z$ such that

$$z = \hat{y} + \delta \hat{x}$$

Cleary $z \in X$. Then

$$\|y - z\|^2 = \|y - \hat{y} - \delta \hat{x}\|^2$$

$$= \|y - \hat{y}\|^2 - 2(y - \hat{y}) \cdot (\delta \hat{x}) + \|\delta \hat{x}\|^2$$

$$= \|y - \hat{y}\|^2 - 2\delta^2 + \delta^2$$

$$= \|y - \hat{y}\|^2 - \delta^2$$

$$< \|y - \hat{y}\|^2$$

That contradicts to that $\hat{y}$ is a solution to $\min_{x \in X} \|y - x\|$. □

That is, the direction from $y$ to $\hat{y}$ is orthogonal to any vector in the subspace $X$. (In case of $y \in X$, it is still valid because the angle between a vector and a zero vector could be anything.) The converse is also true.

**Theorem 10-23.** Let $V$ be a vector space and $X \subset V$ be a subspace with $y \in V$ and $\hat{y} \in X$. If $(y - \hat{y}) \perp X$, then $\|y - \hat{y}\| \leq \|y - x\|$ for all $x \in X$. Moreover, $\|y - \hat{y}\| = \|y - x\|$ if and only if $\hat{y} = x$.

**Proof.**
\[ \|y - x\|^2 = \|(y - \hat{y}) + (\hat{y} - x)\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - x\|^2 + 2\|y - \hat{y}\| \times \|\hat{y} - x\| \times \cos \theta \]

Since \((y - \hat{y}) \perp x\) and \(x \in X\), \(\cos \theta = 0\) and thus

\[ \|y - x\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - x\|^2 \geq \|y - \hat{y}\|^2 \]

The first equality and \(\|y - \hat{y}\| = \|y - x\|\) implies \(\|y - x\|^2 = 2\|y - x\|^2\) or \(\|y - x\|^2 = 0\).

The converse is trivial. \(\blacksquare\)

That is, if \(\hat{y} \in X\) minimizes the distance, then \(y - \hat{y}\) is orthogonal to \(X\), and the minimum point \(\hat{y}\) is unique.

**Example 10-11.** To find the distance between a vector \(y = (1,1)\) and the vector space \(X = \{(x_1, x_2) | x_1 + 2x_2 = 0\}\), we need to find a vector \((x_1, x_2) - (1,1)\) that is orthogonal to every vector in the hyperplane \(x_1 + 2x_2 = 0\). Since \((2, -1) \in X\),

\[ (x_1 - 1, x_2 - 1) \cdot (2, -1) = 0 \quad \text{or} \quad 2x_1 - x_2 - 1 = 0 \]

Since \((x_1, x_2) \in X\), \((x_1, x_2) = \left(\frac{2}{5}, -\frac{1}{5}\right)\). The distance is \(\|(1,1) - \left(\frac{2}{5}, -\frac{1}{5}\right)\| = \frac{7}{5}\).

**Example 10-12.** Let \(y \in \mathbb{R}^n\) and \(X = \text{span}\{1\} \subset \mathbb{R}^n\). Then the orthogonal projection \(\hat{y} = \bar{y} \times 1\) with \(\bar{y} \in \mathbb{R}\) is such that

\[ (y - \hat{y}) \cdot 1 = (y - \bar{y} 1) \cdot 1 = 0 \]

\[ \sum_{i=1}^{n} y_i - \bar{y}(1 \cdot 1) = \sum_{i=1}^{n} y_i - \bar{y}n = 0 \quad \text{or} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]

The sample mean is an orthogonal projection of the sample point to a vector space spanned by one vector.

![Figure 10-2 Sample Mean as a Projection](image-url)
10.5.2 PROJECTION AND MAPPING

Because of the uniqueness of the orthogonal projection, the operation can be viewed as a function from $V$ to $X \subset V$.

$$P: V \rightarrow X$$

which we call a projection. The function is denoted by $P$ and the image $P(y)$ is the orthogonal projection $\hat{y}$. As we shall see shortly, the projection is a linear mapping and it is written $Py$.

**Theorem 10-24.** Let $X \subset \mathbb{R}^n$ be a subspace and the function $P: \mathbb{R}^n \rightarrow X$ be an orthogonal projection onto $X$. Then for any $y \in \mathbb{R}^n$ the $P$ has the properties:

(i) $P$ is a linear map.

(ii) $Px = x$

(iii) $P(Py) = Py$

(iv) $\|Py\| \leq \|y\|$

(v) A projection matrix is symmetric if and only if it is an orthogonal projection.

**Proof.** (i) According to Definition 7-8, it is sufficient to show that, for any $\alpha, \beta \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$,

$$P(\alpha x + \beta y) = \alpha Px + \beta Py$$

To verify the equality, we will show that $\alpha Px + \beta Py$ is an orthogonal projection of $\alpha x + \beta y \in \mathbb{R}^n$. Clearly, $\alpha x + \beta y \in \mathbb{R}^n$ because $x, y \in X$. For any $z \in X$,

$$(\alpha x + \beta y - (\alpha Px + \beta Py)) \cdot z = \alpha (x - Px) \cdot z + \beta (y - Py) \cdot z$$

By the definition of orthogonal projection,

$$\alpha (x + Px) \cdot z + \beta (y - Py) \cdot z = 0$$

(ii) Since $Px$ is the solution to $\min_{z \in X} \|x - z\|$ and $Px \in X$, the minimum is attained at $Px = x$. 
Real Numbers

10.5.3 Projection Matrix

Since an orthogonal projection is a linear map, there exists a \( n \times n \) symmetric matrix such that \( Px = Px \) or \( P(x) = Px \). The \( P \) in the left side is the linear map and the one in the right is a \( n \times n \) matrix. From now on, I will use them interchangeably.

**Theorem 10-25.** Let \( X \subset \mathbb{R}^n \) be a subspace. If \( P \) be the orthogonal projection onto \( X \) and \( \{x_1, \ldots, x_k\} \) is a basis for \( X \), then

\[
P = A (A^T A)^{-1} A^T
\]

where \( A \) is an \( n \times n \) matrix with \( \text{col}_i A = x_i \).

**Proof.** Since it is unique, we need to show that \( Py \in X \) and \( (y - Py) \perp X \).

\[
Py = A (A^T A)^{-1} A^T y = A ((A^T A)^{-1} A^T y)
\]

Since \( A \) is a basis of \( X \), \( A^T A \) is invertible by Theorem 10-26 below. And since \( A \) is a basis and \( (A^T A)^{-1} A^T y \) is an \( n \)-component vector, \( Py \) is a linear combination of the columns of \( A \) so that \( Py \in X \). Finally, for any \( z \in \mathbb{R}^k \), \( (y - Py) \perp Az \).
(Az)^T (y - A(A^T A)^{-1}A^T y) = z^T (A^T y - A^T A(A^T A)^{-1}A^T y) = z^T (A^T y - A^T y) = 0 \quad \blacksquare

If a subspace is generated by $A$, then the projection matrix is $A(A^T A)^{-1}A^T$. Note that if $A$ is not a basis or the columns are not linearly independent, $A^T A$ would be a singular matrix.

**Theorem 10-26.** Let $A$ be an $n \times k$ real matrix. Then,

(i) $A^T A$ is symmetric.

(ii) $\text{rank}(A^T A) = \text{rank}(AA^T) = \text{rank}(A)$.

(iii) $A^T A$ is positive definite if and only if $\text{rank}(A) = k$.

**Proof.**

(i) $(A^T A)^T = A^T (A^T)^T = A^T A$

(ii) Let $x \in \text{Ker}(A)$.

$$Ax = 0 \Rightarrow A^T Ax = 0 \Rightarrow x \in \text{Ker}(A^T A)$$

Thus $\text{Ker}(A) \subset \text{Ker}(A^T A)$.

Let $x \in \text{Ker}(A^T A)$.

$$A^T Ax = 0 \Rightarrow x^T A^T Ax = 0 \Rightarrow (Ax)^T (Ax) = 0 \Rightarrow Ax = 0 \Rightarrow x \in \text{Ker}(A)$$

Thus $\text{Ker}(A^T A) \subset \text{Ker}(A)$.

(iii) For every $x > 0$, $x^T A^T Ax > 0 \Rightarrow (Ax)^T (Ax) > 0 \Rightarrow Ax \neq 0$. Therefore, $A$ has $k$ linearly independent columns. Conversely, since $A$ has full column rank, $Ax \neq 0$ for every $x > 0$. \quad \blacksquare

**Example 10-13.** A direct way to find the projection vector is to use a projection matrix. In Example 10-11,

$$P = x(x^T x)^{-1}x^T = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix}^{-1} = \frac{1}{5} \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$$

Therefore,

$$Py = \frac{1}{5} \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \left( \frac{2}{5}, -\frac{1}{5} \right)$$
An equation system \( Ax = y \) might fail to have a solution if \( y \) cannot be expressed as a linear combination of the columns of \( A \). In such a case, it is conceivable to solve \( A\hat{x} = Py \) by replacing \( y \) with \( Py \in \text{span}(A) \). That is, \( A\hat{x} \) is the orthogonal projection of \( y \) onto the subspace spanned by the columns of \( A \).

Since \( A\hat{x} \) is the orthogonal projection of \( y \) on \( A \), \( (y - A\hat{x}) \perp A \). Since \( (y - A\hat{x}) \cdot \text{col}_i A = 0 \) for all \( i = 1, \ldots, k \),

\[
A^T(y - A\hat{x}) = 0, \quad A^T\hat{x} = A^Ty
\]

If \( A^T \) is invertible, the system has a unique projection matrix. Since \( \hat{x} = (A^TA)^{-1}A^Ty \),

\[
A\hat{x} = Py \Rightarrow A(A^TA)^{-1}A^Ty = Py \Rightarrow P = A(A^TA)^{-1}A^T
\]

Because the equality holds for every \( y \in \mathbb{R}^n \).

If \( A \) is square and invertible, the column space is \( \mathbb{R}^n \) and the original problem has a unique solution. The projection matrix in this case is an identity.

The overdetermined system is typical in estimations and we adopt the least square solution that is identical to the orthogonal projection.

**THEOREM 10-27.** Consider \( Ax = y \) with no solution and all arrays are real. Suppose \( A^TA \) is nonsingular and let \( x^* = (A^TA)^{-1}A^Ty \).

(a) \( x^* \) is closest to solving \( Ax = y \) in the least squares sense that \( x = x^* \) minimizes \( (Ax - y)^2 \).

(b) The least squares approximation \( Ax^* \) to \( y \) is orthogonal to \( Ax^* - y \) such that

\[
Ax^* = A(A^TA)^{-1}A^Ty
\]

(c) In the least squares approximation \( Ax^* = [A(A^TA)^{-1}A^T]y \), the matrix \( A(A^TA)^{-1}A^T \), often called the projection matrix, is symmetric and idempotent.

**Example 10-14.** Consider a linear equation system of \( Ax = b \) where

\[
A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad A^T A = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}
\]
Then,

$$\hat{y} = A(A^TA)^{-1}A^Ty = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1/2 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/3/2 \\ 3/2 \end{bmatrix}$$

**Figure 10-3. Orthogonal Projection**

### 10.5.4 Pseudoinverse

In the linear equation system $Ax = b$, if $A$ is nonsingular, then $x = A^{-1}b$. If $A$ is singular or rectangular, we can take $A^+b$ as a solution to the linear system such that $AA^+A = A$. $A^+$ is called generalized inverse or pseudoinverse of $A$.

**Definition 10-17.** For an $m \times n$ matrix $A$, a **pseudoinverse (Moore-Penrose Inverse)** of $A$ is defined as an $n \times m$ matrix $\tilde{A}$ satisfying the following properties.

(i) $AA^+A = A$, $A^+A$ need not be an identity matrix.

(ii) $A^+AA^+ = A^+$, $A$ is a pseudoinverse of $A^+$.

(iii) $(AA^+)^T = AA^+$ and $(A^+A)^T = A^+A$, symmetry

$A^+$ exists for every matrix, but if a matrix has full rank, it has a simple form.

If non-square matrix $A$ has a full row rank, the right inverse of $A$ is

$$A^+_R = A^T(AA^T)^{-1}$$
If non-square matrix $A$ has a full column rank, the left inverse of $A$ is

$$A_L^+ = (A^T A)^{-1} A^T$$

Let $A$ be of rank $r$. Without loss of generality, let $A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$, where $B_{r \times r}$ is the non-singular submatrix of $A$. Then the generalized inverse of $A$ is

$$A^+ = \begin{bmatrix} B^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

To both sides of linear system $A \bar{x} = P_A b$ multiplying $(A^T A)^{-1} A^T$ yields

$$\left[ (A^T A)^{-1} A^T \right] A \bar{x} = \left[ (A^T A)^{-1} A^T \right] A (A^T A)^{-1} A^T b$$

$$\bar{x} = (A^T A)^{-1} A^T b$$

Clearly, $P = AA^+$ and $Q = A^+ A$ are orthogonal projection operators.
11 MULTIVARIABLE CALCULUS

We extend the calculus to general functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. Since a function $F: \mathbb{R}^n \to \mathbb{R}^m$ can be interpreted as a system with $n$ unknowns and $m$ equations, $F(x_1, \cdots, x_n) = (y_1, \cdots, y_m)$ and $y_j = F_j(x_1, \cdots, x_n)$ where $F_j: \mathbb{R}^n \to \mathbb{R}$, we will focus on $F: \mathbb{R}^n \to \mathbb{R}$.

Recall that those definitions in Section 1.3 are for $n$-component vectors. Those properties regarding continuity and compact set in Sections 3.4 and 3.5 hold exactly in the same form with continuous multivariate functions. We start with a brief review of continuity.

11.1 SEQUENCES, LIMITS, AND CONTINUITY

Consider a sequence $\{x_i\}_{i=1}^{\infty}$ of vectors in $\mathbb{R}^n$, $x_i = (x_{i1}, x_{i2}, \cdots, x_{in}) \in \mathbb{R}^n$.

**Definition 11-1.** The limit of a sequence $\{x_i\}_{i=1}^{\infty}$, denoted by $\lim_{i \to \infty} x_i$, is equal to $x_0$ if the limit $L$ has the property, for each $\varepsilon > 0$, there is an $m$ such that $i > m$ implies

$$d(x_i, x_0) < \varepsilon.$$ 

The only difference from the scalar value sequence is the distance measure.

**Definition 11-2.** $L$ is the limit of a function $f: \mathbb{R}^n \to \mathbb{R}$ at $x_0 \in \mathbb{R}^n$, if for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$0 < d(x, x_0) < \delta \implies |f(x) - L| < \varepsilon$$

and the limit is denoted by

$$\lim_{x \to x_0} f(x) = L$$

The definition agrees with the definition of limits of univariate function with two twists. A general distance replaces absolute value. For our purpose, the standard Euclidean distance is used for the distance function. Second, we do not define one-sided continuity as there are infinitely many directions to approach $x_0$. 

Math for Economists, Sogang University
EXAMPLE 11-1. Find the limit of \( f(x) = x_1 x_2 \) at \((1,1)\) if exists.

We show for every \( \varepsilon > 0 \), there exists \( \delta > 0 \) such that

\[
0 < \| (x_1, x_2) - (1,1) \| < \delta \implies |f(x_1, x_2) - 1| < \varepsilon
\]

Note that

\[
\| (x_1, x_2) - (1,1) \| = \sqrt{(x_1 - 1)^2 + (x_2 - 1)^2}
\]

\[
|f(x_1, x_2) - 1| = |x_1 x_2 - 1| = |x_1(x_2 - 1) + (x_1 - 1)| \leq |x_1(x_2 - 1)| + |(x_1 - 1)|
\]

For any given \( \varepsilon > 0 \), let \( \delta = \min\left(\frac{1}{4} \varepsilon, 1\right) \). Then

\[
\| (x_1, x_2) - (1,1) \| < \frac{1}{4} \varepsilon
\]

and

\[
|(x_1 - 1)| < \frac{1}{4} \varepsilon, \quad |(x_2 - 1)| < \frac{1}{4} \varepsilon.
\]

Since \( x_1 \) is close to \( 1 \),

\[
|x_1(x_2 - 1)| + |(x_1 - 1)| < 2 \times \frac{1}{4} \varepsilon + \frac{1}{4} \varepsilon = \frac{3}{4} \varepsilon,
\]

\[
|f(x_1, x_2) - 1| < \frac{3}{4} \varepsilon < \varepsilon
\]

DEFINITION 11-3. A function \( f: \mathbb{R}^n \to \mathbb{R} \) is **continuous** at \( x_0 \in \mathbb{R}^n \) if

\[
f(x_0) = \lim_{x \to x_0} f(x).
\]

Equivalently, for any \( x_n \to x_0 \),

\[
f\left( \lim_{n \to \infty} x_n \right) = \lim_{n \to \infty} f(x_n).
\]

As in the univariate case, continuity implies that the graph of the function does not break up at \( x_0 \).
11.2 LINEAR FUNCTIONS

**Definition 11-4.** A function \( L: \mathbb{R}^n \to \mathbb{R}^m \) is **linear** if and only if

(i) for all \( x \) and \( y \), \( L(x + y) = L(x) + L(y) \).

(ii) for all scalars \( \lambda \), \( L(\lambda x) = \lambda L(x) \).

The first condition is that a linear function is additive, and the second condition requires the graph of a linear function should contain the origin. An affine function is not linear according to the definition. With these conditions, we can write a linear function in matrix form.

\[
L(x) = Ax, \quad \text{where } A \text{ is an } m \times n \text{ matrix}
\]

Furthermore, because the function has the form of matrix multiplication, the composition functions can be represented by the product of matrices. If \( L: \mathbb{R}^n \to \mathbb{R}^k \) and \( M: \mathbb{R}^k \to \mathbb{R}^m \), and \( L(x) = Ax \) and \( M(x) = Bx \), then \( L \circ M(x) = ABx \).

11.3 REPRESENTING FUNCTIONS

We use a graph to visualize a function. Since \( \text{Graph}(f) = \{(x, y) \mid y = f(x)\} \), the graph of \( f: \mathbb{R}^n \to \mathbb{R} \) is a subset of \( \mathbb{R}^{n+1} \). If \( n \geq 2 \), there is no way to visualize it. However, we can get insight into a surface by making a series of plots of curves in two dimension. One way to do this is to plot level curve.

**Definition 11-5.** A **level set** is the set of points where the function value is some constant value.

\[
\{x \in \mathbb{R}^n \mid f(x) = c\} \quad \text{for any } c \in \mathbb{R}
\]

A level set is generically a curve and called a level curve or contour. Familiar examples are indifference curve, budget line, and production possibility frontier.

**Definition 11-6.** The **upper contour set** of \( f \) is

\[
\{x \in \mathbb{R}^n \mid f(x) \geq c\} \quad \text{for any } c \in \mathbb{R}
\]

And the **lower contour set** is
\{x \in \mathbb{R}^n \mid f(x) \leq c\} \text{ for any } c \in \mathbb{R}.

Budget set, production possibility set, and better/worse sets are well-known examples. Do get confused with graph and level set.

|\begin{align*}
&\text{FIGURE 11-1 UPPER AND LOWER CONTOUR SETS} \\
&\text{11.4 PARTIAL AND DIRECTIONAL DERIVATIVES} \\
&\text{Consider a function } f: \mathbb{R}^n \to \mathbb{R}. \text{ When } n = 1, \text{ there are only two directions that the variable can move at a given point, right or left, and the slope is measured normalizing a direction to the right. When } n \geq 2, \text{ there are infinite directions to move. We cannot apply univariate calculus directly, and need to transform the problem into a univariate function.} \\
&\text{Fix } x_0 \in \mathbb{R}^n \text{ and direction } v \in \mathbb{R}^n. \text{ We can construct a function } F: \mathbb{R} \to \mathbb{R}, \text{ } F(t) = f(x_0 + tv). \\
&F'(0) = \lim_{t \to 0} \frac{F(t) - F(0)}{t} = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} \\
&F'(0) \text{ is the instantaneous change in } f(x_0) \text{ in the limit of } t, \text{ and that is what we call derivative.} \\
&\text{EXAMPLE 11-2. Find the derivative of } f(x) = x_1x_2 \text{ at } x = (2,3) \text{ in the direction of (1,2)} \text{ using } F(t) = f(x_0 + tv) = f((2,3) + t(1,2)). \\
&\text{Since } F(t) = f(x_1 + t, x_2 + 2t) = (x_1 + t)(x_2 + 2t) \\
&\quad \quad \quad \frac{d}{dt} (x_1 + t)(x_2 + 2t) = (x_2 + 2t) + (x_1 + t) \times 2 = 3 + 4 = 7 \\
&\text{Given } f: \mathbb{R}^n \to \mathbb{R}, \text{ consider a direction, a standard basis vector } e_i. \text{ Consider a function } g(t) = f(x_0 + te_i) \text{ for } x_0 \text{ fixed. This is one variable function of } t \text{ and those properties in univariate}
function can be applied, continuity, differentiability, and others. The partial derivative of \( f \) with respect to the \( i \)th variable at \( x_0 \) can be define as \( g'(0) \).

**Definition 11-7.** Let \( v \) be a unit vector in \( \mathbb{R}^n \). The **directional derivative** of \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) in the direction \( v \) at \( x_0 \) is defined as

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

For \( u = \lambda v \), we have \( D_u f = D_v f / \lambda \). Because of this rescaling property, directional derivatives are usually considered only for unit vectors.

**Definition 11-8.** The **partial derivative** of \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) with respect to \( x_i \) at \( x_0 \) is

\[
\frac{\partial f(x_0)}{\partial x_i} = \lim_{t \to 0} \frac{f(x_0 + te_i) - f(x_0)}{t} = \lim_{h \to 0} \frac{f(x_0, x_{02}, \ldots, x_{0,i-1}, x_{0,i} + h, x_{0,i+1}, \ldots, x_{0,n}) - f(x_0)}{h}
\]

There are several notations commonly used in Economics.

\[
\frac{\partial f}{\partial x_i}|_{x_0}, \quad \frac{\partial y}{\partial x_i}|_{x_0}, \quad f_{x_i}(x_0), \quad f_i(x_0), \quad D_{x_i} f(x_0), \quad D_i f(x_0), \quad \frac{d}{dx_i} f(x_i, x_{-i})
\]

### 11.5 Tangent Hyperplane

A natural extension of the tangent line to a function of two variables is tangent plane to the graph. It would have to satisfy two requirements.

(i) The tangent plane must pass a point \((x_0, y_0, z_0)\) where \( z_0 = f(x_0, y_0) \)

(ii) The tangent plane must contain every vector on it including those tangent lines in the \( x \) and \( y \) directions. The general equation of a plane (hyperplane in general) through \((x_0, y_0, z_0)\) is

\[
A(x - x_0) + B(y - y_0) + C(z - z_0) = 0.
\]

Assuming the plane is not vertical or \( C \neq 0 \), we can divide the equation with \( C \) to get

\[
z - z_0 = a(x - x_0) + b(y - y_0), \quad a = A/C, b = B/C
\]
Since $a$ is the slope of the plane in $x$ direction, that is the partial derivative of $f$ with respect to $x$ at $(x_0, y_0)$. $b$ can be calculated similarly. Therefore, the equation of the tangent plane to $z = f(x, y)$ at $(x_0, y_0, z_0)$ is

$$z - z_0 = \frac{\partial f(x_0, y_0)}{\partial x}(x - x_0) + \frac{\partial f(x_0, y_0)}{\partial y}(y - y_0)$$

Given the result, in a close neighborhood of $(x_0, y_0, z_0)$, the tangent plane should be a good approximation of the graph of the function.

$$z = z_0 + \frac{\partial f(x_0, y_0)}{\partial x}(x - x_0) + \frac{\partial f(x_0, y_0)}{\partial y}(y - y_0)$$

The expression in the right-hand side is often called the linearization of $f(x, y)$ at $(x_0, y_0, z_0)$, and sometimes it is used with $\Delta$ notation.

$$\Delta f(x) \approx \frac{\partial f(x_0, y_0)}{\partial x}\Delta x + \frac{\partial f(x_0, y_0)}{\partial y}\Delta y$$

Note that the whole argument is based on the assumption that the tangent plane is a good approximation to the surface at $(x_0, y_0, z_0)$. It is widely accepted that if the partial derivatives are continuous in some rectangle centered at $(x_0, y_0)$, $f(x, y)$ is “smooth” enough.

Note that, in a linear approximation, the order of changes does not matter at all since the hyperplane is a linear function whose “slope” is the same at any point on the plane.

### 11.6 Derivative

**Definition 11-9.** The function $f: \mathbb{R}^n \to \mathbb{R}$ is **differentiable** at a point $x_0$ if and only if there is a linear function $L: \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{x \to x_0} \frac{f(x) - f(x_0) - L(x_0) \times (x - x_0)}{\|x - x_0\|} = \lim_{x \to x_0} \left[ \frac{f(x) - f(x_0)}{\|x - x_0\|} - \frac{L(x_0) \times (x - x_0)}{\|x - x_0\|} \right] = 0$$

If $L(x_0)$ exists, we call it the **derivative** of $f$ at $x_0$ and denoted by $Df(x_0)$. Note that $L(x_0)$ has a form of $1 \times n$ matrix.
The definition does not provide a way to calculate a derivative. We first show how the derivative and the directional derivative are related, and then express the directional derivative in terms of the derivative.

Letting \( x = x_0 + tv \), the definition of derivative implies that if \( f \) is differentiable, then for any directions \( v \),

\[
\lim_{t \to 0} \left[ \frac{f(x_0 + tv) - f(x_0)}{|t| \cdot \|v\|} - \frac{Df(x_0)tv}{|t| \cdot \|v\|} \right] = 0
\]

\[
\lim_{t \to 0} \left[ \left( \frac{f(x_0 + tv) - f(x_0)}{t} - Df(x_0)v \right) \frac{t}{|t| \cdot \|v\|} \right] = 0
\]

Since \( \frac{t}{|t| \cdot \|v\|} \) does not converges, we have

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

If a function is differentiable at a point, then all directional derivative exists at the point. The definition and the property can be directly extended to \( f: \mathbb{R}^n \to \mathbb{R}^m \).

**Theorem 11-1.** If a function is differentiable at \( x_0 \), then all of the directional derivative exist. The directional derivative of \( f \) in direction \( v \) is

\[
D_v f(x_0) = Df(x_0)v = \sum_i \frac{\partial f(x_0)}{\partial x_i} v_i
\]

That is, if a function is differentiable, the matrix representation of the derivative is the matrix of partial derivatives. And a directional derivative is a weighted average of the partial derivatives.

**Proof.** The first equality is shown above. For the second, imagine a tangent hyperplane to the graph of the function at \( (x_0, f(x_0)) \). In the neighborhood of \( x_0 \),

\[
f(x_0 + tv) - f(x_0) \approx [f(x_0 + tv \cdot e_1) - f(x_0)] + \cdots + [f(x_0 + tv \cdot e_n) - f(x_0)]
\]

or

\[
\frac{f(x_0 + tv) - f(x_0)}{t} \approx \frac{f(x_0 + tv \cdot e_1) - f(x_0)}{t} + \cdots + \frac{f(x_0 + tv \cdot e_n) - f(x_0)}{t}
\]
Since
\[ \frac{\partial f(x_0)}{\partial x_i} = \lim_{t \to 0} \frac{[f(x_0 + te_i) - f(x_0)]}{t}, \]
we have
\[ \frac{\partial f(x_0)}{\partial x_i} = \lim_{t \to 0} \frac{[f(x_0 + tv_i e_i) - f(x_0)]}{tv_i} \]
or
\[ \frac{\partial f(x_0)}{\partial x_i} v_i = \lim_{t \to 0} \frac{[f(x_0 + tv_i e_i) - f(x_0)]}{t} \]
Therefore, we have
\[ \frac{f(x_0 + tv) - f(x_0)}{t} \approx \frac{\partial f(x_0)}{\partial x_1} v_1 + \frac{\partial f(x_0)}{\partial x_2} v_2 + \cdots + \frac{\partial f(x_0)}{\partial x_n} v_n \]
Taking limit gives the result,
\[ D_v f(x_0) = \left[ \frac{\partial f(x_0)}{\partial x_1} \quad \frac{\partial f(x_0)}{\partial x_2} \quad \cdots \quad \frac{\partial f(x_0)}{\partial x_n} \right] v. \]
Note that the converse is not true in general. For the converse of the statement, we need an additional condition.

**Theorem 11-2.** If the all partial derivatives exist and are continuous, then the function is differentiable.

**Theorem 11-3.** $F: \mathbb{R}^n \to \mathbb{R}^m$ with $F = (F_1, \ldots, F_m)$, is differentiable if each $F_j$ is differentiable.

**Definition 11-10.** The Jacobian matrix of a function $F: \mathbb{R}^n \to \mathbb{R}^m$ is a $m \times n$ matrix of all first-order partial derivatives, and is denoted by $DF$ or $J_F$.

\[ J_F = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}. \]
Note that each row contains the partial derivatives of a component function $F_j$ of $F$.

### 11.7 Chain Rule

**THEOREM 11-4.** Let functions $g, f: \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $x$. Then

1. $D(\lambda f)(x) = \lambda Df(x)$ for all $\lambda \in \mathbb{R}$
2. $D(f + g)(x) = Df(x) + Dg(x)$

For $m = 1$,

3. $D(f \cdot g)(x) = g(x)Df(x) + f(x)Dg(x)$
4. $D\left(\frac{f}{g}\right)(x) = \frac{g(x)Df(x) - f(x)Dg(x)}{(g(x))^2}$

**THEOREM 11-5 (Chain Rule)** If $G: X \to \mathbb{R}^k$, $X \subset \mathbb{R}^n$, is differentiable at $x \in X$ and $F: \mathbb{R}^k \to \mathbb{R}^m$, $Y \subset \mathbb{R}^k$, is differentiable at $y = G(x) \in Y$, then $F \circ G$ is differentiable at $x$ and

$$D(F \circ G)(x) = DF(y) \cdot DG(x)$$

**Proof.** The extension is straightforward. Let’s start with the linear approximation at $x$.

$$\Delta F_j(y) \approx \frac{\partial F_j(y)}{\partial y_1} \Delta y_1 + \frac{\partial F_j(y)}{\partial y_2} \Delta y_2 + \cdots + \frac{\partial F_j(y)}{\partial y_n} \Delta y_n$$

Now hold $x_{-i}$ constant and divide by $\Delta x_i$ to get

$$\frac{\Delta F_j(y)}{\Delta x_i} \approx \frac{\partial F_j(y)}{\partial y_1} \frac{\Delta y_1}{\Delta x_i} + \frac{\partial F_j(y)}{\partial y_2} \frac{\Delta y_2}{\Delta x_i} + \cdots + \frac{\partial F_j(y)}{\partial y_n} \frac{\Delta y_n}{\Delta x_i}$$

Taking limit gives the result and stack them up for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$ to get,

$$DF(y) = J_F(y) \times J_G(x)$$

where
\[ J_F = \begin{bmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_k} \end{bmatrix}, \quad J_g = \begin{bmatrix} \frac{\partial G_1}{\partial x_1} & \cdots & \frac{\partial G_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial G_k}{\partial x_1} & \cdots & \frac{\partial G_k}{\partial x_n} \end{bmatrix} \]

It is important to keep track of the dimensions.

\[ D(F \circ G)(x) = \frac{DF(G(x))}{m \times k} \frac{DG(x)}{k \times n} \]

**Example 11-3.**

\[ G: \mathbb{R}^2 \to \mathbb{R}^3, \quad G(x, y) = (x + y, 2x - 5, xy^2) \]

\[ F: \mathbb{R}^3 \to \mathbb{R}, \quad F(u, v, w) = u + v + 3vw \]

\[ F \circ G(x, y) = x + y + 2x - 5 + 3xy^2(2x - 5) \]

To get an explicit formula for \( D(F \circ G) \), The chain rule requires to evaluate \( DF(G(x_0, y_0)) \) and \( DG(x_0, y_0) \).

\[ DG = \begin{bmatrix} 1 \\ 2y^2 \\ -2xy \end{bmatrix}, \quad DF = \begin{bmatrix} 1 & 1 + 3w & 3v \end{bmatrix} \]

\[ D(F \circ G) = \begin{bmatrix} 3 + 12xy^2 - 15y^2 & 1 + 12x^2y - 30xy \end{bmatrix} \]

**Example 11-4.** Consider a function \( f(x, y) \) and \( y = g(x) \). By the chain rule, we have

\[ D_x f(x, y) = D_x f(x, y) + D_y f(x, y) D_x g(x) \]

This is the total derivative of \( f(x, y) \) with respect to \( x \).

**Theorem 11-6 (Leibnitz’s Rule)** Suppose that \( f(x, t) \) is continuous with a continuous derivative \( \frac{\partial f}{\partial t} \) in the rectangle \( x \in [a, b] \) and \( t \in [c, d] \) of the \( x-t \) plane. If \( A(t) \) and \( B(t) \) are continuously differentiable, and

\[ V(t) = \int_{A(t)}^{B(t)} f(x, t) \, dx \]

then
\[ V'(t) = f(B(t), t) \cdot B'(t) - f(A(t), t) \cdot A'(t) + \int_{A(t)}^{B(t)} \frac{\partial f(x,t)}{\partial t} \, dx. \]

**Proof.** Differentiate \( \int_{A(t)}^{B(t)} f(x,t)dx = F(B(t), t) - F(A(t), t). \) \[\blacksquare\]

### 11.8 **Matrix Differentiation**

**Definition 11-11 (Matrix Differentiation)**

\[
\frac{dA(z)}{dz} = \begin{bmatrix} da_{ij}(z) \end{bmatrix}, \quad \frac{dz(A)}{dA} = \begin{bmatrix} dz(A) \end{bmatrix}, \quad \frac{dz(x)}{dx} = \begin{bmatrix} dz_i(x) \end{bmatrix}
\]

Matrix differentiation refers to the derivative of a matrix function with respect to a scalar variable, the derivative of a scalar function with respect to a matrix, and the derivative of a vector function with respect to a vector.

The matrix derivatives provide a convenient way to keep track of partial derivatives in calculations.

**Theorem 11-7 (Matrix Differentiation Formulae)** For \( f: \mathbb{R}^n \to \mathbb{R}^m \) and \( g: \mathbb{R}^k \to \mathbb{R}^l \), and all vectors are column vectors.

(i) \( D(Af(x) + Bg(x)) = Adf(x) + Bdg(x), \quad A_{h \times m}, \quad B_{h \times l}, \quad \text{and} \quad n = k. \)

(ii) \( D\left([f(x)]^Tg(x)\right) = [g(x)]^Tdf(x) + [f(x)]dg(x), n = k \)

(iii) \( D(f \circ g(x)) = Df(g(x))dg(x), k = l = n \)

(iv) \( Df^{-1}(y) = [Df(x)]^{-1}, m = n \)

(v) \( D(Ax + b) = A \)

(vi) \( D(x^TA + b^T) = A^T \)

(vii) \( D(x^TAx) = x^T(A^T + A) \)

(viii) \( \frac{de^{At}}{dt} = Ae^{At} \)
When more than one variables are changed simultaneously or sequentially, we can compute their effects on the function value considering them as independent shocks.

**Definition 11-12.** The total differential of a function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) is the sum of the partial differentials with respect to all of the independent variables.

\[
 df(x) = f_1(x)dx_1 + \cdots + f_n(x)dx_n = D_x f(x)dx 
\]

\( f_i(x)dx_i \) is the magnitude of the change in \( f \) due to the changes in \( x_i \) holding others constant. Total differential can be thought of as a directional derivative without the requirement of direction being a unit vector. If we divide the expression by \( dx_1 \),

\[
 \frac{df(x)}{dx_1} = f_1(x) + f_2(x)\frac{dx_2}{dx_1} + \cdots + f_n(x)\frac{dx_n}{dx_1} = D_x f(x) \frac{dx(x_1)}{dx_1},
\]

the total effect is the sum of the direct effect measured by \( f_1(x) \) and the indirect effect by the others, \( dx_1 \rightarrow dx_i \rightarrow f_i(x) \).

### 11.10 Gradient

Partial derivatives are the rate of change of a function in the direction of standard basis vectors. So we can think of it as a compass directing the steepest ascent direction. To exploit this property, it is quite useful to interpret the derivative as a vector.

**Definition 11-13.** Given \( f: \mathbb{R}^n \rightarrow \mathbb{R} \), the gradient of \( f \) at \( x_0 \) is

\[
 \nabla f(x_0) = (Df(x_0)) = \left( \frac{\partial f(x_0)}{\partial x_1}, \frac{\partial f(x_0)}{\partial x_2}, \ldots, \frac{\partial f(x_0)}{\partial x_n} \right)
\]
We read this as “gradient of $f$” or “grad $f$”. It is worth to note that

(i) $\nabla$ takes a scalar function $f(x)$ and produces a vector $\nabla f(x)$.

(ii) The vector $\nabla f(x)$ lies in the hyperplane of the graph space.

11.10.1 GEOMETRIC INTERPRETATION

**THEOREM 11-8.** Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x_0$, then the direction that maximizes the directional derivative at $x_0$ is given by

$$v = \nabla f(x_0)$$

**Proof.** The slope (or rate of increase) of $f$ at $x_0$ in the direction of the unit vector $v$ is given by

$$D_v f(x_0) = v \cdot f'(x_0) = v \cdot \nabla f(x_0)$$

By Cauchy-Schwarz inequality,

$$\|D_v f(x_0)\| \leq \|v\| \cdot \|\nabla f(x_0)\| = \|\nabla f(x_0)\|$$

This shows that the marginal change in the value of the function cannot be greater than $\|\nabla f(x_0)\|$. But, taking $v = \nabla f(x_0)/\|\nabla f(x_0)\|$, we obtain that

$$D_v f(x_0) = \frac{\nabla f(x_0) \cdot \nabla f(x_0)}{\|\nabla f(x_0)\|} = \|\nabla f(x_0)\|$$

Therefore, $\nabla f(x_0)$ points in the direction of steepest ascent at $x_0$ and the magnitude of the slope is equal to $\|\nabla f(x_0)\|$.

11.10.2 LEVEL SETS

**THEOREM 11-9.** If $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x_0$, then the equation of the hyperplane tangent to the graph of $f(x)$ at $(x_0, y_0)$ is

$$y - y_0 = \nabla f(x_0) \cdot (x - x_0).$$
Proof. The graph of \( f(x) \) is given by \( \{(x, y) \mid y = f(x)\} \) or \( f(x) - y = 0 \). Let's define a new function \( F(x, y) = f(x) - y \). The surface given by \( y = f(x) \) is identical to the surface of \( F(x, y) = 0 \). Let \( z = (x, y) \) and define \( z(t) = z^0 + tv \) where \( v \in \mathbb{R}^{n+1} \) is the direction on the level curve or level surface. Since \( F(z(t)) = 0 \) for all \( t \). Differentiating \( F(z(t)) \) at \( t = 0 \) gives

\[
\nabla F(z^0) \cdot v = \langle \nabla f(x^0), -1 \rangle \cdot v = \nabla f(x^0) \cdot (x - x^0) - (y - y_0) = 0
\]

\[\Box\]

**Figure 11-2. Tangent Hyperplane**

**Theorem 11-10.** If \( f : \mathbb{R}^n \to \mathbb{R} \) is differentiable at \( x_0 \), then \( \nabla f(x_0) \) is perpendicular to the vector tangent to the level curve of \( f(x) \) at \( x_0 \).

**Proof.** Let \( F(x) = f(x) - f(x_0) \), and \( x(t) = x_0 + tv \) where \( v \in \mathbb{R}^n \) is the direction along the level curve. Then \( F(x(t)) = 0 \) for all \( t \).

\[
\nabla f(x_0) \cdot D_t x(0) = \nabla f(x_0) \cdot v = 0
\]

\[\Box\]
Figure 11-3. Gradient Orthogonal to the Tangent Line

Note that the gradient is perpendicular to the tangent plane and Theorem 11-9 and Theorem 11-10 states essentially the same property. That is, for a function \( y = f(x) \), let \( F(x, y) = f(x) - y \). Then the level set \( F(x, y) = 0 \) is the surface of \( f(x) \). Therefore, \( \nabla_{(x,y)} F(x, y) \) is perpendicular to the surface of \( f(x) \) and thus to the tangent hyperplane.

Corollary 11-11. If \( F: \mathbb{R}^{n+1} \to \mathbb{R} \) is differentiable at \( z_0 \) with \( F(z_0) = 0 \) and \( DF(z_0) \neq 0 \), then the equation of the hyperplane tangent to the surface of \( F(z) = 0 \) at the point \( z^0 \) is

\[
\nabla F(z_0) \cdot (z - z_0) = 0
\]

When we do not have an explicit expression of \( y = f(x) \) from \( F(z) = 0 \) with \( z = (x,y) \), it is still possible to find the equation of the tangent hyperplane.

Example 11-5. Consider \( f(x_1, x_2) = x_1^2 + x_2^2 \). \( f(x_1, x_2) = r^2 \) is a circle of radius \( r \) centered at \((0,0)\). At \((x_1, x_2) = (3,4)\), the level set is \( \{(x_1, x_2) : x_1^2 + x_2^2 = 5^2\} \). Since

\[
\nabla f(3,4) = (2x_1, 2x_2)|_{(3,4)} = (6,8)
\]

Therefore, the equation of the tangent line at \((x_1, x_2) = (3,4)\) is

\[
(x_1, x_2) = (3,4) + t(-4,3)
\]

And the equation of the tangent hyperplane at \((x_1, x_2) = (3,4)\) is

\[
(6,8,-1) \cdot ((x_1,x_2,y) - (3,4,25)) = 0
\]

\[
y - 25 = (6,8)((x_1, x_2) - (3,4))
\]

\[
y - 6x_1 - 8x_2 = 75
\]

11.11 Higher Order Derivatives

Every result for \( f(x) \) holds with \( Df(x) \).

Definition 11-14. The second derivative of \( f(x) \) is called Hessian or Hessian matrix of \( f(x) \) and written as
\[ Hf(x) = D^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \]

**Theorem 11-12 (Young’s Theorem)** Suppose that \( f(x) \) is twice continuously differentiable on \( X \). Then for all \( x \in X \) and for all \( i, j \),

\[ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i} \]

If these conditions hold, then \( Hf(x) \) is symmetric.

**Proof.** Let \((x_1, \ldots, x_n) \equiv (x, y, z) \equiv (x, y)\) so that we can concentrate on two variables.

Let \( g(a) = f(a, y + \Delta y) - f(a, y) \), and apply Mean Value Theorem to get

\[ g(x + \Delta x) - g(x) = g'(a^*)\Delta x \]

for some \( a^* \in (x, x + \Delta x) \). Then

\[ f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) - f(x, y + \Delta y) + f(x, y) = \left[ \frac{\partial f(a^*, y + \Delta y)}{\partial x} - \frac{\partial f(a^*, y)}{\partial x} \right] \Delta x. \]

Repeat the steps with \( h(b) = \frac{\partial}{\partial x} f(a^*, b) \), then

\[ h(y + \Delta y) - h(y) = h'(b^*) \]

for some \( b^* \in (y, y + \Delta y) \), and

\[ f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) - f(x, y + \Delta y) + f(x, y) = \frac{\partial^2 f(a^*, b^*)}{\partial x \partial y} \Delta x \Delta y. \]

Repeat the whole steps once again with \( k(a) = f(x + \Delta x, b^*) - f(x, b^*) \) and \( l(b) = \frac{\partial}{\partial y} f(a^*, b) \) to get

\[ f(x + \Delta x, y + \Delta y) - f(x + \Delta x, y) - f(x, y + \Delta y) + f(x, y) = \frac{\partial^2 f(a^{**}, b^{**})}{\partial y \partial x} \Delta x \Delta y. \]

Therefore, we have
\[
\frac{\partial^2 f(a^*, b^*)}{\partial x \partial y} = \frac{\partial^2 f(a^{**}, b^{**})}{\partial y \partial x}
\]

Note that \(a^*\) and \(a^{**}\) are between \(x\) and \(x + \Delta x\), and \(b^*\) and \(b^{**}\) are between \(y\) and \(y + \Delta y\). Because of continuity, in the limit of \(\Delta x, \Delta y \to 0\),

\[
\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x}
\]

### 11.12 Taylor’s Theorem

If \(f: \mathbb{R}^n \to \mathbb{R}\), the value of \(f(x)\) at \(x_0\) can be approximated in terms of its derivatives.

To apply one variable case, consider \(F(t) = f(x_0 + t(x - x_0))\) and \(F(0) = f(x_0)\) and \(F(1) = f(x)\). Provided that appropriate conditions hold, there exists \(c\) between \(x\) and \(x_0\) such that

\[
F(1) = F(0) + F'(0) + \frac{1}{2!} F''(0) + \cdots + \frac{1}{(n-1)!} F^{(n-1)}(0) + \frac{1}{n!} F^n(c)
\]

By substitution,

\[
f(x) = f(x_0) + F'(0) + \frac{1}{2!} F''(0) + \cdots + \frac{1}{(n-1)!} F^{(n-1)}(0) + \frac{1}{n!} F^n(c)
\]

\[
F'(t) = Df(x_0 + t(x - x_0))(x - x_0) = \sum_i D_i f(x_0 + t(x - x_0))(x_i - x_{0i})
\]

\[
F'(0) = \sum_i D_i f(x_0)(x_i - x_{0i})
\]

\[
F''(t) = \sum_i \sum_j D_j D_i f(x_0 + t(x - x_0))(x_i - x_{0i})(x_j - x_{0j})
\]

\[
F'''(t) = \sum_k \sum_j \sum_i D_k D_j D_i f(x_0 + t(x - x_0))(x_i - x_{0i})(x_j - x_{0j})(x_k - x_{0k})
\]

\[\vdots\]

Since the derivative of \(Df(x_0 + t(x - x_0))(x - x_0) = \sum_i D_i f(x_0 + t(x - x_0))(x_i - x_{0i})\) , the expansion has terms of the form \(D_i D_j f(x_0)\) that is the second derivative of \(f(x)\) defined in the "obvious" way: for \(D_i f: \mathbb{R}^n \to \mathbb{R}\), \(D_i D_j f(x_0)\) is just a partial derivative of \(D_i f(x_0)\). Similarly, for \(D_i D_k D_j D_i f(x_0)\), and so on.
Notation: We denote by $Hf(x_0)$ the $n \times n$ matrix of second partial derivatives whose $(i,j)^{th}$ entry of Hessian is $D_iD_j f(x_0)$.

**First order approximation**

$$A_1(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

$$\lim_{x \to x^0} \frac{|H_1(x, x^0)|}{\|x - x_0\|} = \lim_{x \to x^0} \frac{|f(x) - A_1(x)|}{\|x - x_0\|} = 0$$

where $H_1(x, x^0)$ is the approximation error that vanishes faster than $\|x - x_0\|$.

**Second order approximation**

$$A_2(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^THf(c)(x - x_0)$$

$$\lim_{x \to x^0} \frac{|H_2(x, x_0)|}{\|x - x_0\|^2} = \lim_{x \to x^0} \frac{|f(x) - A_2(x)|}{\|x - x_0\|^2} = 0$$

Note that

$$\sum_j \sum_i D_jD_i f(x_0 + t(x - x_0))(x_i - x_{0i})(x_j - x_{0j}) = (x - x_0)^THf(c)(x - x_0)$$

**Example 11-6.** Find the second order Taylor’s approximation of $f(x, y) = x^3y + 2xy$ around $(x, y) = (1, 1)$.

$$A_2(x, y) = f(1, 1) + \nabla f(1, 1)\left(\frac{x - 1}{y - 1}\right) + \frac{1}{2}\left(\frac{x - 1}{y - 1}\right)^THf(1, 1)\left(\frac{x - 1}{y - 1}\right)$$

$$Df(x, y) = [3x^2y + 2y \quad x^3 + 2y]$$

$$Df(1, 1) = [5 \quad 3]$$

$$D^2f(x, y) = \begin{bmatrix} 6xy & 3x^2 + 2 \\ 3x^2 + 2 & 0 \end{bmatrix}$$

$$D^2f(1, 1) = \begin{bmatrix} 6 & 5 \\ 5 & 0 \end{bmatrix}$$

$$A_2(x, y) = 3 + \left(\frac{5}{3}\right)(x - 1) + \frac{1}{2}(x - 1)^THf(1, 1)\left(\frac{x - 1}{y - 1}\right)$$

$$= 3 + \left(\frac{5}{3}(x - 1) + \frac{1}{2}(x - 1)^T\begin{bmatrix} 6 & 5 \\ 5 & 0 \end{bmatrix}(x - 1)\right)(y - 1)$$

$$= 3 + 5xy - 5 + 3(x - 3) + 3(x - 1)^2 + 5(x - 1)(y - 1)$$

$$= 3x^2 + 5xy - 4x - 2y + 3$$

$$A_2(1, 1) = 5$$

$$f(x, y) = 3$$
Theorem 11.13 (Taylor's Theorem) Suppose that a function \( f: \mathbb{R}^n \to \mathbb{R} \) is \( k \) times continuously differentiable on an open set containing the line segment connecting \( x^0 \) and \( x \). Then, there exists \( c \) on the segment such that

\[
f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2} (x - x_0)^T Hf(c)(x - x_0) + \ldots
\]

\[
+ \frac{D^i f(x_0)}{i!} + \frac{D^j f(x_0)}{(j + 1)!} + \ldots + \frac{D^k f(x_0)}{k!} + \frac{D^{k+1} f(c)}{(k + 1)!}
\]

And

\[
\lim_{x \to x^0} \frac{|H_k(x,x_0)|}{\|x - x_0\|^k} = 0 \quad \text{where} \quad H_k(x,x_0) = \frac{D^{k+1} f(c)}{(k + 1)!}.
\]

11.13 Implicit Function Theorem

We already learned that a function is invertible if it is one-to-one and onto. However, one can make any one-to-one function effectively invertible. That is, if \( f: X \to Y \) is one-to-one, then \( f^{-1}: f(X) \to X \) is always a function. If \( f \) is invertible, we can define \( g(f(x)) = x \) on \( B_\varepsilon(x_0) \).

Definition 11.15. A function \( f: \mathbb{R}^n \to \mathbb{R}^n \) is locally invertible at \( x_0 \) if there is a \( \varepsilon > 0 \) and a function \( g: B_\varepsilon(f(x_0)) \to \mathbb{R}^n \) such that

\[
f \circ g(y) \equiv y \quad \text{for} \quad y \in B_\varepsilon(f(x_0)), \quad \text{and} \quad g \circ f(x) \equiv x \quad \text{for} \quad x \in B_\varepsilon(x_0).
\]

Generally, a function is not invertible because it is not one-to-one. But if there is a subset of the domain where it is strictly monotone, it is one-to-one and therefore locally invertible.

For multivariable functions, linear theory says, roughly, if you have \( n \) linear equations with \( n \) unknowns, you expect a unique solution. If \( A \) is an \( n \times n \) matrix and invertible, then \( Ax = b \) has a unique solution \( x = A^{-1}b \). In general, to get a unique solution, the number of equations and the number of dependent variables should be the same. In general, a function \( f: \mathbb{R}^n \to \mathbb{R}^m \) would not be invertible unless \( n = m \).

Theorem 11.14 (Inverse Function Theorem) If \( f: \mathbb{R}^n \to \mathbb{R}^n \) is continuously differentiable at \( x_0 \) and \( Df(x_0) \) is invertible, then \( f(x) \) is locally invertible at \( x_0 \). Moreover, the inverse function, \( g(\cdot) \) is differentiable at \( f(x_0) \) and \( Dg(f(x_0)) = (Df(x_0))^{-1} \).
In the typical economic applications, a system of equations characterizes an economic equilibrium. We solve the equation for the variables under our interest as “functions” of parameters. The theorem shows that if the system can be solved at \( x_0 \) then also in the neighborhood of \( x_0 \). Furthermore, it gives expressions for the derivatives of the solution function.

The inverse function theorem does not help in this situation as its applications require the explicit form of the function. An explicit function is one in which dependent variable is given explicitly in terms of the independent variables while in implicit function the dependent variable is not expressed in terms of independent variables such as \( x^2 \ln y + y^2 e^x = 2 \).

If we are interested in the effect of a small change in \( x \) on \( y \) at \((x_0, y_0)\), we would like to solve the equation for \( y \), \( y = y(x) \). That looks impossible.

**Example 11-7.** Let \( F(x, y) = x^2 \ln y + y^2 e^x - 2 \) so that the surface passing \( 0 \) is the level set of \( F(x, y) = 0 \).

\[
2x \ln y + \frac{x^2}{y} \frac{dy}{dx} + 2y \frac{dy}{dx} e^x + y^2 e^x = 0
\]

For instance, in the neighborhood of \((x_0, y_0) = (1,1)\),

\[
(1 + 2e) \frac{dy}{dx} + 1 = 0, \quad \frac{dy}{dx} = \frac{1}{1 + 2e}
\]

Suppose that \( F: \mathbb{R}^{n+m} \to \mathbb{R}^m \) represents \( m \) equations in \( n + m \) variables. (Note that this is not a function or correspondence but a \( m \) equations, and we can always write \( F = 0 \).) We can think of the first \( n \) variables \( x = (x_1, \cdots, x_n) \) as parameters, and last \( m \) variables \( y = (y_1, \cdots, y_m) \) as endogenous variables. In principle, you can solve the equations for \( y \).

**Theorem 11-15 (Implicit Function Theorem)** Suppose \( F: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m \) is differentiable and \( D_x F(x_0, y_0) \) is not singular. There exists a function \( g: \mathbb{R}^n \to \mathbb{R}^m \) such that \( y = g(x) \) uniquely solves \( F(x, y) = 0 \) in the neighborhood of \((x_0, y_0)\). Furthermore, in the neighborhood of \((x_0, y_0)\),

\[
D_x g(x_0) = -\left(D_y F(x_0, y_0)\right)^{-1} D_x F(x_0, y_0).
\]

**Sketch of the proof.** Assume that \( g(x) \) exists (that is the most difficult part of the proof.)
Suppose
\[ F: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m, \quad g: \mathbb{R}^n \to \mathbb{R}^m \]

Let \( H(x) = F(x,Y(x)) \) so that \( H: \mathbb{R}^n \to \mathbb{R}^m \), \( H(x) = F \circ G(x) \), and \( G: \mathbb{R}^n \to \mathbb{R}^{n+m} \), \( G(x) = (x,Y(x)) \).

\[
\frac{DH}{m \times n} = \frac{DF}{m \times (n+m)} \times \frac{DG}{(n+m) \times n} = \begin{bmatrix} \frac{D_x F}{m \times n} & \frac{D_y F}{m \times m} \end{bmatrix} \begin{bmatrix} I_n \\frac{DY}{m \times n} \end{bmatrix} = D_x F + D_y F \times DY
\]

Since \( H(x) \) is identically zero by construction, \( DH = 0 \), and \( D_x F = -D_y F \times DY \). If \((D_y F)^{-1}\) exists, \( DY = -(D_y F)^{-1} D_x F \).

Inverse Function Theorem is a special case of implicit function theorem. For \( f: \mathbb{R}^n \to \mathbb{R}^n \), applying Implicit Function Theorem to \( F(x,y) = f(x) - y \) yields Inverse Function Theorem.

\[
F(x,y), y = 0, \quad D_y F(x(y),y) \times D_x x(y) + D_x F(x(y),y) = 0, \quad D_y x(y) = (D_y F)^{-1}
\]

**Example 11-8.** One of the most famous application in economics would be the marginal rate of substitution.

An indifference curve is defined as \( u(x) = u_0 \). The MRS of good \( i \) for good \( j \) is the opportunity cost of a good \( i \) in term of another good, \( j \). The consumption of all the other goods are assumed fixed. By implicit function theorem or differentiating with respect to \( x_j \),

\[
\frac{\partial u(x)}{\partial x_i} \frac{dx_i}{dx_j} + \frac{\partial u(x)}{\partial x_j} = 0, \quad \left( \frac{\partial u(x)}{\partial x_i}, \frac{\partial u(x)}{\partial x_j} \right) \cdot (dx_i, dx_j) = 0
\]

When \( x \in \mathbb{R}^2 \), the MRS \( i,j(x) \) is the slope of the tangent line to the indifference curve,

\[
\frac{dx_i}{dx_j} = \frac{\partial u(x)}{\partial x_j} / \frac{\partial u(x)}{\partial x_i}
\]

**11.14 Homogeneous Functions**

In many economic applications, the solution for the optimization problem often has a very specific functional form, homogeneous function.
**Definition 11-16.** $f: \mathbb{R}^n \to \mathbb{R}$ is **homogeneous** of degree $k \in \mathbb{Z}$ if $f(tx) = t^k f(x)$ for all $t > 0$.

For the properties to hold, the domain should be a cone and it is usually defined on the positive orthant, $\mathbb{R}^n_{++}$.

Homogeneity of degree one is weaker than linearity. All linear functions are homogeneous of degree one, but not conversely. For example, $f(x, y) = \sqrt{xy}$ is homogeneous of degree one but not linear.

**Example 11-9.** If $f(x, y)$ is homogeneous of degree one and we take $\lambda = 1/x$, and rearrange, we get $f(x, y) = xf(1, y/x) = xg(y/x)$. The value of the function is the product of a scale factor $x$ and the value of a function of the ratio $y/x$.

In Solow’s growth model, $F(L, K) = AL^\alpha K^{1-\alpha}$ with $0 < \alpha < 1$ so that the production function is homogeneous of degree 1. Dividing the function by $L$ gives

$$
\frac{F(L, K)}{L} = F \left(1, \frac{K}{L}\right) = f(k) = AL^{1-\alpha} K^{\alpha} = A \left(\frac{K}{L}\right)^\alpha = Ak^\alpha,
$$

so that the per capita output depends only on the capital per capita.

**Theorem 11-16 (Euler’s Theorem)** If $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x$ and homogeneous of degree $k \geq 1$, then $\nabla f(x) \cdot x = kf(x)$.

**Proof:** Differentiate $f(tx) = t^k f(x)$ with respect to $t$ (the equality should hold after differentiation because it is an identity for $t$ and $x$) yields

$$
Df(tx) \cdot x = kt^{k-1} f(x).
$$

The result follows with $t = 1$.

Since $f(tx) = t^k f(x)$ holds for every $t$ and $x$, we can also differentiate it with respect to $x$.

**Theorem 11-17.** If $f$ is differentiable and is homogeneous of degree $k$, its first partial derivatives are homogeneous of degree $k - 1$.

**Proof:** $Df(tx) \times t = t^k Df(x)$, $Df(tx) = t^{k-1} Df(x)$
**Theorem 11-18.** If $f$ is differentiable and homogenous, then the tangent planes to the level set of $f$ on each ray from the origin are parallel to each other. That is, the equation of the tangent plane is $\nabla f(x^0) \cdot (x - tx^0) = 0$ for all $x^0 > 0$ and $t > 0$.

**Proof.** $\nabla f(tx^0) = t^{k-1} \nabla f(x)$ and thus

$$\nabla f(tx^0) \cdot (x - tx^0) = t^{k-1} \nabla f(x) \cdot (x - tx^0) = 0, \quad \nabla f(x) \cdot (x - tx^0) = 0$$

\[\blacksquare\]

**Definition 11-17 (Homothetic Function)** A function $h(x)$ is homothetic if there exists a monotone function $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, and a homogeneous function $g: \mathbb{R}_+^n \rightarrow \mathbb{R}_+$ such that

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

**Theorem 11-19.** If $f$ is differentiable and homothetic, then the tangent planes to the level set of $f$ on each ray from the origin are parallel to each other. That is, the equation of the tangent plane is $\nabla f(x_0) \cdot (x - tx_0) = 0$ for all $x_0 > 0$ and $t > 0$. 
12 Quadratic Form

12.1 Quadratic Forms

**Definition 12-1.** A quadratic form on $\mathbb{R}^n$ is a real valued function of the form

$$Q(x) = \sum_{i \leq j} c_{ij}x_ix_j$$

Note that a quadratic form is a second-degree polynomial without a constant term.

**Theorem 12-1.** Every quadratic form $Q$ can be represented by a symmetric matrix $A$ so that

$$Q(x) = x^TAx$$

**Proof.** The coefficient of $x_i^2 = a_{ii}$ and $x_ix_j$ has the coefficient of $a_{ij} + a_{ji}$. Therefore, the matrix $A$ with $a_{ij} = a_{ji} = c_{ij}/2$ where $c_{ij}$ is the coefficient of $x_ix_j$ in the quadratic form $Q$ is symmetric.

Any twice continuously differentiable function has $n \times n$ symmetric Hessian matrix, and the second order term in a Taylor’s expansion is of a quadratic form. Quadratic form is closely related to the properties of second derivatives and the shape of functions.

**Definition 12-2.** A real and symmetric matrix $A$ is negative definite if, for all $x \neq 0$, $x^TAx < 0$.

Taking $x = e_i$, $x^TAx = a_{ii}$. Therefore, every element in the diagonal must be negative.

**Definition 12-3.** A real and symmetric matrix $A$ is negative semidefinite if, for all $x \neq 0$, $x^TAx \leq 0$.

Positive definite and positive semidefinite matrices are defined similarly.

**Example 12-1.** An ellipsoid is a set of the form

$$\{x|(x - x_0)^T A^{-1}(x - x_0) \leq 1\}$$
where $A$ is an symmetric positive definite matrix.

**Definition 12-4.** A real and symmetric matrix $A$ is in **definite** if $x^T A x > 0$ for some $x$ and $x^T A x < 0$ for some other $x$.

Note that quadratic forms involve only symmetric matrix $A$, and the theory of diagonalization gives a way to relate them. Since it is possible to write $A = P^T A P$,

$$Q(x) = x^T A x = x^T P A P^T x = (P^T x)^T A (P^T x)$$

Hence $Q(x)$ has a form of $Q(x) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2$ where $y_i$ is the $i$th element of $P^T x$. Since the definiteness requires the inequality to hold for all $x$, we have the following.

**Theorem 12-2.** A symmetric matrix $A$ is negative definite if every eigenvalue is negative, and is positive definite if every eigenvalue is positive.

There is a way to determine the definiteness without calculating eigenvalues.

**Definition 12-5.** Let $A$ be an $n \times n$ symmetric matrix. The $k$th order leading principal submatrix of $A$ is the matrix that is obtained by deleting the last $n - k$ rows and columns from $A$, and is denoted by $A_k$. Its determinant is called the $k$th order leading principal minor of $A$.

**Theorem 12-3.** Let $A$ be an $n \times n$ symmetric matrix. Then

(i) $A$ is positive definite if and only if its $n$ leading principal minors are (strictly) positive.

(ii) $A$ is negative definite if the sign of $k$th order leading principal minor is $(-1)^k$.

**Example 12-2.**

$$A = \begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix}$$

$k = 1, \quad n - k = 2, \quad (-1)^1 = -1, \quad \det A_1 = \det[-2] = -2$

$k = 2, \quad n - k = 1, \quad (-1)^2 = 1, \quad \det A_2 = \det\begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} = 4 - 1 = 3$

$k = 3, \quad n - k = 0, \quad (-1)^3 = -1, \quad \det A_3 = \det\begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix} = -8 - (-2 - 2) = -4$
**Definition 12-6.** A symmetric matrix \( A \) is negative semidefinite on \( \{ x \mid Bx = 0 \} \) if \( x^T Ax \leq 0 \) for all \( \{ x \mid Bx = 0 \} \).

**Theorem 12-4.** Let \( B \) be an \( k \times n \) matrix, where \( k < n \) and where the first \( k \) columns of \( B \) form a nonsingular matrix. Let \( C = \begin{bmatrix} 0 & B^T \\ B & A \end{bmatrix} \) and let \( C_r \) be the submatrix of \( C \) gotten by deleting all but the first \( r \) rows and columns. Then:

(i) \( x^T Ax > 0 \) for all \( x \neq 0 \) with \( Bx = 0 \) if and only if \( (-1)^k \det(C_{2k+i}) > 0 \) for \( i = 1, \cdots, n - k \). Alternatively, if \( \det(C) \) and the last \( n - k \) leading principal minors have the same sign as \( (-1)^k \), then \( x^T Ax > 0 \) for all \( x \neq 0 \) with \( Bx = 0 \).

(ii) \( x^T Ax < 0 \) for all \( x \neq 0 \) with \( Bx = 0 \) if and only if \( (-1)^{k+1} \det(C_{2k+i}) > 0 \) for \( i = 1, \cdots, n - k \). Alternatively, if \( \det(C) \) has the sign of \( (-1)^n \) and if the last \( n - k \) leading principal minors alternate in sign, then \( x^T Ax < 0 \) for all \( x \neq 0 \) with \( Bx = 0 \).

\( C \) is called a bordered matrix and is useful to check the definiteness of a symmetric matrix on a subset of \( \{ x \mid Bx = 0 \} \).

**Example 12-3.**

\[
C = \begin{bmatrix} 0 & \cdots & 0 & b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & b_{k1} & \cdots & b_{kn} \\ b_{11} & \cdots & b_{k1} & a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ b_{1n} & \cdots & b_{kn} & a_{n1} & \cdots & a_{nn} \end{bmatrix}
\]

The size of the bordered matrix is \((n + k) \times (n + k)\) and has \((n + k)\) leading principal minors.

\[
M_1, \cdots, M_k, M_{k+1}, \cdots, M_{2k-1}, M_{2k}, M_{2k+1}, \cdots, M_{n+k} = C
\]

The first \( k \) matrices \( M_1, \cdots, M_k \) are zero matrices and next \( k - 1 \) matrices \( M_{k+1}, \cdots, M_{2k-1} \) have zero determinant. The determinant of the next principle minor \( M_{2k} \) is \( \pm (\det(B_{k \times k}))^2 \) where \( B_{k \times k} \) is the matix with the first \( k \) rows and the first \( k \) columns. \( \det(M_{2k}) \) does not contain any information about \( A \).

The determinant of last \( n - k \) matrices \( M_{2k+1}, \cdots, M_{n+k} \) carry information about \( A \) and \( B \). Theorem 12-4 states that if the determinant of \( C = M_{n+k} \) has the sign \((-1)^n\) and the signs

---

Math for Economists, Sogang University
of the determinant of the last \((n + k)\) leading principal minors alternate in sign, then \(A\) is negative definite on \(\{x \mid Bx = 0\}\).

### 12.2 Convex Sets

A large part of optimization theory relies on the property of a convex set, and the convexity is indispensable for most of our results.

**Definition 12-7.** \(X \subset \mathbb{R}^n\) is convex if for every \(x, y \in X\) and \(\lambda \in [0,1]\),

\[
\lambda x + (1 - \lambda)y \in X.
\]

**Theorem 12-5.** If \(X \subset \mathbb{R}^n\) is convex, its closure is also convex.

**Proof.** Let \(\bar{X}\) be the closure of \(X\), \(x, y \in \bar{X}\) and \(\lambda \in [0,1]\). Choose sequences \(x_i, y_i \in X\) such that \(x_i \rightarrow x\) and \(y_i \rightarrow y\). Since \(X\) is convex and \(\bar{X}\) be the closure, \(\lambda x_i + (1 - \lambda)y_i \in X\), and

\[
\lim_{i \rightarrow \infty} \lambda x_i + (1 - \lambda)y_i = \lambda x + (1 - \lambda)y \in \bar{X}.
\]

**Theorem 12-6.** Let \(X\) be a nonempty closed convex subset of \(\mathbb{R}^n\), and \(x_0 \notin X\). Then there exist \(p \in \mathbb{R}^n \setminus \{0\}\) and \(c \in \mathbb{R}\) such that

\[
X \subset \{x \mid x \cdot p \geq c\} \quad \text{and} \quad x_0 \cdot p < c
\]

**Proof.** Consider the problem of minimizing the distance between \(x_0\) and \(X\). That is, find \(x^*\) to solve

\[
\min_{x \in X} \|x_0 - x\|
\]

While \(X\) is not necessarily bounded, since it is nonempty, we can find an element \(z \in X\) with which we can replace \(X\) in the problem by

\[
\{x \in X \mid \|x_0 - x\| \leq \|x_0 - z\|\}.
\]

so that the set is compact and a solution exists. Let \(p = x^* - x_0\) and \(c = x^* \cdot p\). This is possible since \(x_0 \notin X\) and \(p \neq 0\). \(x_0 \cdot p < c\) because \(c - p \cdot x^0 = (x^* - x_0) \cdot p = \|p\|^2\).

To complete the proof, we have to show that
\[ x \in X \implies x \cdot p \geq c. \]

This inequality is equivalent to

\[ (x - x^*) \cdot (x^* - x_0) \geq 0. \]

Since \( X \) is convex and \( x^* \) is the solution to the problem, it must be that

\[ \|\lambda x + (1 - \lambda)x^* - x_0\|^2 \]

is minimized when \( \lambda = 0 \) so that the derivative of \( \|\lambda x + (1 - \lambda)x^* - x_0\|^2 \) is non-negative at \( \lambda = 0 \). (This is a corner solution that we have not learned. If it is negative, we can make the objective smaller by increasing \( \lambda \).) Differentiation and evaluate at \( \lambda = 0 \) yields

\[ 2(x^* - x_0)^T(x - x^*) \geq 0 \]

This theorem can be refined for a boundary point with which the result is known as Supporting Hyperplane Theorem. Since \( x^0 \notin \text{int}(X) \), we can construct a sequence \( \{x^k\} \subseteq X \) such that \( x^k \) converges to \( x^0 \). For each \( x^k \), apply \textsc{Theorem 12-6} yields a sequence of \( p^k \) that satisfy the condition of the theorem. Since a subsequence of \( p^k \) should converge (\( p^k \) is bounded), the limit of \( p^k \) will satisfy the conclusion of the theorem.

\textbf{Definition 12-8.} For a \( X \subseteq \mathbb{R}^n \) and boundary point of \( X \) of \( x^0 \), the nonzero vector \( p \) is said to supports \( X \) at \( x^0 \) if \( p \cdot x \geq p \cdot x^0 \) for all \( x \in X \) (or if \( p \cdot x \leq p \cdot x^0 \) for all \( x \in X \)). The hyperplane \( \{x \mid p \cdot x = p \cdot x^0\} \) is a \textbf{supporting hyperplane} for \( X \) at \( x^0 \).

A supporting hyperplane divides the whole space into two half-spaces and \( X \) is entirely contained in one of the closed half-spaces bounded by the hyperplane.

\textbf{Theorem 12-7 (Supporting Hyperplane Theorem)} Let \( X \) be a convex subset of \( \mathbb{R}^n \) and \( x_0 \) is a point on the boundary of \( X \), then there exists a supporting hyperplane containing \( x_0 \).

\textbf{Proof.} Let \( \overline{X} \) be the closure of \( X \). Then \( \overline{X} \) is convex by \textsc{Theorem 12-5}. Since \( x_0 \) is a point on the boundary of \( X \), there is a sequence \( x_k \notin \overline{X} \) for every \( k \) and \( x_k \to x_0 \). By application of \textsc{Theorem 12-6}, for every \( k \),

\[ X \subset \{x \mid x \cdot p_k \geq c_k\} \quad \text{and} \quad x_k \cdot p_k < c_k \]
Without loss of generality, let \( \|p_i\| = 1 \) for every \( k \). Since the set of \( p_i \) is compact, there exists a convergent subsequence \( p_{i_k} \to p \). Note that the associated subsequence of \( x_i \) denoted by \( x_{i_k} \) converges to \( x_0 \). For every \( x \in X \),

\[
p_{i_k} \cdot x_{i_k} < c_k \leq p_{i_k} \cdot x
\]

implies

\[
p \cdot x_0 = \lim_{i_k \to \infty} p_{i_k} \cdot x_{i_k} \leq \lim_{i_k \to \infty} p_{i_k} \cdot x = p \cdot x
\]

Setting \( c = p \cdot x_0 \) completes the proof. \( \blacksquare \)

**Theorem 12-8 (Hyperplane Separation Theorem)** Let \( X \) and \( Y \) be nonempty, convex, disjoint subsets of \( \mathbb{R}^n \). Then there exists a nonzero \( p \) that separates \( X \) and \( Y \).

\[
\sup_{x \in X} p \cdot x \leq \inf_{y \in Y} p \cdot y
\]

\( p \cdot x = 0 \) is the hyperplane that separates the two sets.

**Proof.** Consider set \( Z \) with elements of \( z_{ij} = x_i - y_j \) or \( z = x - y \). This set is non-empty and convex. Since \( X \cap Y = \emptyset \), by the supporting hyperplane theorem, there exists \( p \in \mathbb{R}^n \) such that

\[
\sup_{z \in Z} p \cdot z \leq 0
\]

Therefore, \( p \cdot z \leq 0 \) for all \( z \in Z \). Because \( z = x - y \), we have

\[
p \cdot (x - y) \leq 0 \quad \text{or} \quad p \cdot x \leq p \cdot y
\]

Since the inequality holds for all \( x \) and \( y \), taking supremum over \( x \in X \) and infimum over \( y \in Y \) show the result. \( \blacksquare \)

### 12.3 Shape of Functions

Replacing \( x \in \mathbb{R} \) with \( x \in \mathbb{R}^n \), the definitions of concave and convex functions are identical to those in Section 5.6. Since the second derivative of a multivariable function is \( n \times n \) matrix, other results can be extended to multivariable functions with the definiteness of Hessian.
**Theorem 12-9.** Suppose that \( f(x) \) is twice continuously differentiable. Then it is concave if and only if its Hessian matrix is negative semidefinite. When \( x \in \mathbb{R} \), \( f(x) \) is concave if and only if \( D^2 f(x) \leq 0 \).

**Proof.** By Taylor theorem, there exists a \( c \in [x, x + \Delta x] \) such that

\[
f(x + \Delta x) - f(x) = \nabla f(x) \cdot \Delta x + \frac{1}{2} \Delta x^T D^2 f(c) \Delta x \leq \nabla f(x) \cdot \Delta x
\]

The level set of a univariate function is usually singleton or discrete, and level sets contain no useful information on the function’s property. For the optimization problem with constraints, the shape of a level set is critical for our approach.

The definitions of quasi-concave and quasi-convex function are based on the shape of level sets.

**Definition 12-9.** A function \( f(x) \) is quasi-concave if and only if, for all \( \lambda \in [0,1] \),

\[
f(\lambda x + (1 - \lambda) y) \geq \min[f(x), f(y)].
\]

**Theorem 12-10.** A continuously differentiable function \( f(x) \) is quasi-concave if and only if, for \( f(y) \geq f(x) \),

\[
\nabla f(x) \cdot (y - x) \geq 0
\]

**Proof.**

\[
f(\lambda x + (1 - \lambda) y) \geq \min[f(x), f(y)] \geq f(x)
\]

Let \( g(1 - \lambda) = f(x + (1 - \lambda)(y - x)) \). Then

\[
\frac{f(x + (1 - \lambda)(y - x)) - f(x)}{1 - \lambda} = \frac{g(1 - \lambda) - g(0)}{1 - \lambda} \geq 0
\]

Take limit of \( \lambda \to 1 \) to obtain

\[
g'(0) = \nabla f(x) \cdot (y - x) \geq 0.
\]

To prove the converse, let \( g(\lambda) = f(x + \lambda(y - x)) \). We need to show that \( g(\lambda) \geq g(0) \) when \( f(x) \leq f(y) \) implies \( \nabla f(x) \cdot (y - x) \geq 0 \). Suppose not so that there exists \( \lambda \in (0,1) \) such that \( g(0) > g(\lambda) \). Because \( g(1) \geq g(0) \), we can choose \( \lambda \) such that \( g'(\lambda) > 0 \). Letting \( z = x + \lambda(y - x) \),

Math for Economists, Sogang University
\[ \nabla f(z) \cdot (x - z) = \nabla f(z) \cdot (-\lambda)(y - x) \geq 0, \quad \nabla f(z) \cdot (y - x) \leq 0 \]

which contradicts to \( g'(\lambda) = \nabla f(z) \cdot (y - x) > 0 \).

**Theorem 12-11.** A function \( f \) is quasi-concave if and only if its upper contour sets \( \{ x | f(x) \geq t \} \) are convex for all \( t \).

Fix \( x_0 \). For any \( x, y \in \{ x | f(x) \geq f(x_0) \} \), \( f(x), f(y) \geq f(x_0) \) and the quasiconcavity implies \( f(\lambda x + (1 - \lambda)y) \geq \min[f(x), f(y)] \geq f(x_0) \). In turn, \( \lambda x + (1 - \lambda)y \in \{ x | f(x) \geq f(x_0) \} \).

Suppose \( \{ x | f(x) \geq f(x_0) \} \) is convex for every \( x_0 \). Consider \( x, y \in \{ x | f(x) \geq f(x_0) \} \), with \( f(x) \geq f(y) \). Letting \( y = x_0 \), \( \lambda x + (1 - \lambda)y \in \{ x | f(x) \geq f(x_0) \} \). Therefore, \( f(\lambda x + (1 - \lambda)y) \geq f(y) = \min[f(x), f(y)] \).

Quasi-concavity and quasi-convexity are global properties of a function. Unlike continuity, the differentiability, concavity, and convexity of a function cannot be defined at a point.

**Theorem 12-12 (Monotone Transformation)** Let \( f \) be a quasi-concave function.

(i) if \( g \) is increasing, \( g \circ f \) is quasi-concave.

(ii) if \( g \) is decreasing, \( g \circ f \) is quasi-convex.

Quasi-concavity is preserved under positive transformation, and thus all the ordinal properties of functions too.

**Theorem 12-13.** If a function \( f \) is quasi-concave, it is concave on \( \{ \Delta x \mid \nabla f(x) \cdot \Delta x = 0 \} \). If \( f \) is strictly quasi-concave, it is strictly concave on \( \{ \Delta x \mid \nabla f(x) \cdot \Delta x = 0 \} \).

**Proof.** At any point \( x_0 \in X \), since the upper contour set is convex, there is a separating hyperplane plane that \( \nabla f(x) \cdot (x_0 - x) \geq 0 \) where \( x \) is a point on the tangent line to the level set passing \( x_0 \). Along the tangent line, the function attains its global maximum at \( x_0 \) and the second derivative is negative semidefinite.

This result implies that quasi-concavity can be checked by finding the definiteness of the bordered Hessian.

**Theorem 12-14.** Let \( X \) be an open and convex set in \( \mathbb{R}^n \), and let \( f : X \to \mathbb{R} \) is twice continuously differentiable. If \( f \) is quasi-concave on \( X \), then \((-1)^k \det \bar{H}_k \geq 0\) for \( k = 1, \cdots, n \). Conversely, \((-1)^k \det \bar{H}_k > 0\) for \( k = 1, \cdots, n \), then \( f \) is quasi-concave on \( X \).
\[ H_k f(\mathbf{x}) = \begin{bmatrix}
0 & f_1(\mathbf{x}) & \cdots & f_k(\mathbf{x}) \\
f_1(\mathbf{x}) & f_{11}(\mathbf{x}) & \cdots & f_{1k}(\mathbf{x}) \\
\vdots & \vdots & \ddots & \vdots \\
f_k(\mathbf{x}) & f_{k1}(\mathbf{x}) & \cdots & f_{kk}(\mathbf{x})
\end{bmatrix} \]

In the theorem, weak inequality is not sufficient for concavity.

**Example 12-4**

Consider \( f(x, y) = (x - 1)^2(y - 1)^2 \) for \( x, y \geq 0 \), we have

\[
\begin{align*}
\det(\overline{H}_1) &= -4(x - 1)^2(y - 1)^4 \leq 0, \text{ for all } x, y \in \mathbb{R}_+^2 \\
\det(\overline{H}_2) &= 16(x - 1)^4(y - 1)^4 \geq 0, \text{ for all } x, y \in \mathbb{R}_+^2.
\end{align*}
\]

The equalities hold only for \( x = y = 1 \). Since \( f(0,0) = f(1,1) = 1 \), taking \( \lambda = 1/2 \) gives

\[
f\left(\frac{1}{2}(0,1) + \frac{1}{2}(2,2)\right) = f(1,1) = 0 < 1 = \min[f(0,0), f(2,2)]
\]

and thus \( f(x) \) is not quasi-concave.

**12.4 Transformation**

The key concepts in the analyses so far are continuity and differentiability of a function, and also the shape of functions plays important roles in optimization problems. This section shows how those properties are affected or preserved under functional transformations.

The cardinality is associated with counting and the ordinality is something to do with order. For instance, the cardinality of a set is the number of elements, and there is no definition for ordinality of a set. (An ordered set is a set over which an order is defined, and the concept is not directly related to the set itself.)

The cardinality is a property related to counting, and the property should be preserved only under positive scalar multiplications or positive linear transformations. The definition extends it to addition.

**Definition 12-10.** A property of a function is said **cardinal** if the property is preserved under a positive affine transformation.
That is, if a function \( f(x) \) is cardinal, then \( af(x) + b \) with \( a > 0 \) is also cardinal. Regarding the shape of a function, most properties are preserved under cardinal transformation. For an univariate function with \( a > 0 \), \( D(af(x) + b) = af'(x) \) and \( D^2(af(x) + b) = af''(x) \) have the same sign as \( f'(x) \) and \( f''(x) \), respectively. The property associated with the measure based on a specific unit is cardinal property, such as diminishing marginal utility and risk aversion.

**Definition 12-11.** A property of a function is said **ordinal** if the property is preserved under a strictly increasing transformation.

The ordinal property is about order structure of an ordered set and only cares the order not how much. The competition scheme in the Olympic games has an ordinal property because the silver medal goes to one whose performance is the second place. It does not matter how s/he was worse than the gold and how better than the bronze.

**Example 12-5.** Consider a concave function \( f(x) = \sqrt{x} \) and \( \rho(x) = x^4 \). On the domain of \( f(x) \), \( \rho(\cdot) \) is strictly increasing. \( \rho \circ f(x) = x^2 \) is convex. A concavity is not an ordinal property but cardinal.

On the other hand, suppose that \( f(x) \) is a quasi-concave function and \( \rho(\cdot) \) is strictly increasing. Then \( \{x|\rho \circ f(x) \geq c\} \) is the same as \( \{x|f(x) \geq \rho^{-1}(c)\} \) because \( \rho^{-1}(\cdot) \) is also strictly increasing. The quasi-concavity is an ordinal property.
13 UNCONSTRAINED OPTIMIZATION

An optimization problem without a constraint is written

\[ \text{UMP: } \max_{x \in \mathbb{R}^n} f(x; a), \min_{x \in \mathbb{R}^n} f(x; a) \]

max/min denotes the criterion. The vector of the variable(s) under max/min denotes the values to be determined to optimize the objective function \( f(x; a) \), and \( a \) is a parameters.

Since the solution to \( \max_x f(x) \) also solves \( \min_x [-f(x)] \) and the maximum is the same as negative of the minimum, we will focus on a maximization problem.

Note that even if a monotone transformation is applied to the objective function, the level curves remain unchanged. Therefore, \( \max_{x \in X} f(x) \) and \( \max_{x \in X} \rho(f(x)) \), where \( \rho(\cdot) \) is a positive monotone transformation, have the same solution for every \( X \), a subset of \( \mathbb{R}^n \).

13.1 CONDITIONS FOR OPTIMALITY

The basic characterization of local extrema is identical to the univariate case. They are reproduced using vector notation.

**Definition 13-1.** \( x^* \) is a local maximum if there exists \( \delta > 0 \) such that \( f(x^*) \geq f(x) \) for all \( x \in B_\delta(x^*) \).

**Theorem 13-1 (First Order Necessary Condition).** Suppose that \( f: \mathbb{R}^n \to \mathbb{R} \) is differentiable and \( f(x^*) \) attains its local maximum (local minimum) at \( x^* \). Then \( \nabla f(x^*) = 0 \).

**Proof.** Let \( h(t) = f(x^* + tv) \) for any \( v \in \mathbb{R}^n \) and \( t \in \mathbb{R} \). Since \( f(x^*) \geq f(x) \) for all \( x \in B_\epsilon(x^*) \) for some \( \epsilon \), for a sufficiently small \( t \) (\( t < \epsilon ||v|| \)), we have

\[ h(t) = f(x^* + tv) \leq f(x^*). \]

Since \( h(t) \) is maximized locally at \( t = 0 \), by **Theorem 5-1**, \( h'(0) = 0 \) or

\[ \nabla f(x^*) \cdot v = 0. \]
Since this must hold for all $v$, this implies $\nabla f(x^*) = 0$.

The intuition and the proof of the univariate case remain valid. Also, the second order condition is the concavity of the objective function.

**Theorem 13-2 (Second Order Sufficient Condition)** Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. If $\nabla f(x^*) = 0$ and $D^2f(x^*)$ is negative definite, then $f(x^*)$ is a unique local maximum.

**Proof.** As in Theorem 13-1, for every $v \in \mathbb{R}^n$ the second order condition in Theorem 5-2 should be satisfied in the maximization problem of $h(t) = f(x^* + tv)$ with respect to $t$

$$h''(0) = v^T D^2 f(x^*) v < 0.$$  

Since this must hold for all $v$, this implies $D^2 f(x^*)$ is negative definite.

The result can be shown using Taylor’s theorem. According to Taylor’s Theorem, there exists a $c$ between $x^*$ and $x$ such that

$$f(x) - f(x^*) = \frac{1}{2} (x^* - x)^T D^2 f(c)(x^* - x).$$

Since $D^2 f(c)$ is continuous and $(x^* - x)^T D^2 f(x^*)(x^* - x)$ is strictly negative, if $x$ is sufficiently close to $x^*$, $(x^* - x)^T D^2 f(c)(x^* - x)$ should be negative. Therefore, $f(x) - f(x^*) < 0$ for all $x \in B_\varepsilon(x^*)$.

**Theorem 13-3 (Second Order Necessary Condition)** Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. If $x^*$ is a local maximizer, then $D^2 f(x^*)$ is negative semidefinite.

**Proof.** Since $f(x) - f(x^*) = \frac{1}{2} (x^* - x)^T D^2 f(c)(x^* - x)$, if $(x^* - x)^T D^2 f(x^*)(x^* - x) > 0$, $x^*$ cannot be a local maximizer.

**Theorem 13-4.** If $f$ is concave, then any $f(x^*)$ with $\nabla f(x^*) = 0$ is a global maximum.

**Proof.** Direct from that, for every $dx$,

$$f(x^* + dx) = f(x^*) + \frac{1}{2} dx^T D^2 f(c) dx,$$ $f(x^* + dx) \leq f(x^*)$.

**Example 13-1.** Gradient Method
When we find the maximum value numerically, the most popular way is to move in the direction of the gradient. But the problem is the step size. It might be conceivable to move as long as the function value increases. Let $\mathbf{h}$ is a vector representing direction and step size.

$$\max_{\mathbf{h}} f(\mathbf{x}_0 + \mathbf{h}) \approx \max_{\mathbf{h}} f(\mathbf{x}_0) + Df(\mathbf{x}_0) \cdot \mathbf{h} + \frac{1}{2} \mathbf{h}^T D^2 f(\mathbf{x}_0) \cdot \mathbf{h}$$

The first order condition is

$$Df(\mathbf{x}_0) + \mathbf{h}^T D^2 f(\mathbf{x}_0) = \mathbf{0}$$

Since $\mathbf{h}$ is the difference between the starting point $\mathbf{x}_0$ and the end point $\mathbf{x}_1$, $\mathbf{h} = \mathbf{x}_1 - \mathbf{x}_0$,

$$\mathbf{h}^* = -[D^2 f(\mathbf{x}_0)]^{-1}[ Df(\mathbf{x}_0)]^T = -[D^2 f(\mathbf{x}_0)]^{-1} \nabla f(\mathbf{x}_0)$$

where $D^2 f(\mathbf{x}_0)$ is assumed symmetric.

To find the maximum, repeat the process $\mathbf{x}_{n+1} = \mathbf{x}_n - [D^2 f(\mathbf{x}_n)]^{-1} \nabla f(\mathbf{x}_n)$ until it converges.

### 13.2 Comparative Statics

In economics, comparative statics is to compare the economic outcomes before and after a change in some underlying exogenous variable. In particular, economists are often interested in the sensitivity of the maximized value of the objective function and the optimal value of choice variables to changes in the parameters of the problem.

The analyses are based on the first order conditions that are implicit functions of variables and parameters.

#### 13.2.1 Envelope Theorem

Suppose that $\mathbf{x}^* = \mathbf{x}(a)$ is a unique solution to $\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}; a)$. Define value function as

$$V(a) = \max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}; a) = f(\mathbf{x}(a), a)$$

for $a \in \mathbb{R}$. 
The function can be differentiated using chain rule if $f(x; a)$ and $x(a)$ are continuously differentiable, and the implicit function theorem gives the condition under which $x(a)$ is differentiable.

Since $x^*$ solves $\max_{x \in \mathbb{R}^n} f(x; a)$ if and only if $D_xf(x^*; a) = 0$, $x(a)$ is implicitly defined as a solution to a system of equations.

$$V'(a) = D_xf(x^*; a)D_\alpha x(a) + D_\alpha f(x^*, a).$$

The implicit function theorem shows that $D_\alpha x(a)$ exists and gives the formula for the derivative. In order to evaluate $V'(a)$, however, we do not need to know that value of $D_\alpha x(a)$ but only the existence because unconstrained optimization problems can have only interior solutions, $D_xf(x^*; a) = 0$.

$$V'(a) = D_\alpha f(x^*, a).$$

The condition for the existence of $D_\alpha x(a)$ is the non-singularity of $D^2_xf(x^*; a)$, which is the second order sufficient condition.

**THEOREM 13.5 (Envelope Theorem)** Suppose that the first order conditions of an optimization problem satisfy the conditions of the implicit function theorem. If the objective function is continuously differentiable at the optimum,

$$\frac{df(x(a), a)}{da} = \frac{\partial f(x^*, a)}{\partial a}.$$

The essence of envelope theorem is that the effect of any induced change in $x$ are locally negligible because the value function is locally independent of $x$ in the neighborhood of the solution.
13.2.2 Implicit Function Theorem

The first order condition, $Df(x(a)) = 0$, is identity in the parameter $a$. By implicit function theorem, these can be solved locally for $x$ as a function of $a$ if the matrix $D_x^2 f(x^*)$ is nonsingular. Since this is the Hessian of the objective function, this is not singular whenever the second-order sufficient conditions are satisfied. Thus,

$$D_x^2 f(x^*; a)D_a x(a) + D_a D_x f(x^*; a) = 0, \quad D_a x(a) = -[D_x^2 f(x^*; a)]^{-1} D_a \nabla f(x^*; a)$$

Some cases in economic applications such as Hotelling’s Lemma (input demand) and Shepherd’s Lemma (conditional input demand), the same information can be obtained either by envelope theorem or by implicit function theorem.

When the objective is linear in parameters, you can derive the comparative statics results from the value function. For instance, the supply function is the solution to profit maximization,

$$\pi(p) = \max_{y \in \mathcal{Y}} p \cdot y$$

Since the constraint is independent of $p$, by envelope theorem, we have $D_p \pi(p) = y(p)$. We can fully characterize $D_p y(p)$ with the Hessian of the value function. Or simply, because the profit function is convex, $\partial y_j(p)/\partial p_j > 0$ for all $j$. 
14 Optimization with Equality Constraints

Consider a constrained maximization problem:

$$\max_{x \in X} f(x; a)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$. In this problem, a solution $x^*$ should be a member of $X$ and $f(x^*) \geq f(x)$ for all $x \in X$.

When the feasible set $X$ is characterized by equations such as $X = \{x \in \mathbb{R}^n | G(x) = 0\}$ where $G(x) = (g_1(x), \ldots, g_m(x))$, we call it equality constraints. Since $G(x)$ is an arbitrary function, $G(x) = b$ can always be replaced by $G(x) - b = 0$. The general form of maximization problem with equality constraints is written as:

$$\max_{x \in \mathbb{R}^n} f(x; a)$$

subject to $G(x; a) = 0$

Since $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the problem has $m$ equality constraints and $n$ choice variables. Again $a \in \mathbb{R}$ is a parameter. If $m \geq n$, the problem usually has no solution or the solution is determined solely by the constraints. We consider only the problems with $m < n$.

14.1 Necessary Conditions

Intuitively, if we can solve the constraint equation system considering a group of choice variables as parameters, substituting them into the objective function converts MPE to UMP. For instance, if $g_i$’s are linear, a set of variables ($m$ variables) can be expressed with the other $n - m$ variables.

$$\max_{x_1, x_2, y} f(x_1, x_2, y)$$

subject to $x_1 + x_2 - y = 0$ \implies $$\max_{x_1, x_2} f(x_1 + x_2, x_1 + x_2)$$

Even when the constraints are nonlinear, the same trick can be applied using implicit function theorem.
Let’s start with a simple example.

\[
\max_{x_1, x_2} f(x_1, x_2) \quad \text{subject to} \quad h(x_1, x_2) = 0
\]

Since the feasible \(x_1\) and \(x_2\) lie on the level set of \(h(x_1, x_2)\), suppose that the equation can be solved for \(x_2 = x_2(x_1)\). The substitution gives an unconstrained problem.

\[
\max_{x_1} f(x_1, x_2(x_1))
\]

The first order condition is

\[
\frac{df(x_1, x_2(x_1))}{dx_1} = \frac{\partial f(x_1, x_2)}{\partial x_1} + \frac{\partial f(x_1, x_2)}{\partial x_2} \frac{dx_2}{dx_1} = 0
\]

By implicit function theorem,

\[
\frac{\partial g(x_1^*, x_2^*)}{\partial x_1} + \frac{\partial h(x_1^*, x_2^*)}{\partial x_2} \frac{dx_2}{dx_1} = 0, \quad \frac{dx_2}{dx_1} = -\frac{\partial h(x_1^*, x_2^*)/\partial x_1}{\partial h(x_1^*, x_2^*)/\partial x_2}
\]

Note that this requires one of the partial derivatives of \(g(x_1, x_2)\) is not zero, and this is called regularity condition. Then the first order condition becomes

\[
\frac{df(x_1^*, x_2^*)}{dx_1} + \frac{\partial f(x_1^*, x_2^*)}{\partial x_2} \left(-\frac{\partial h(x_1^*, x_2^*)/\partial x_1}{\partial h(x_1^*, x_2^*)/\partial x_2}\right) = 0
\]

Letting \(\lambda^* = \frac{\partial f(x_1^*, x_2^*)}{\partial x_2} \frac{\partial h(x_1^*, x_2^*)}{\partial x_1}\) we have

\[
\frac{\partial f(x_1^*, x_2^*)}{\partial x_1} - \lambda^* \frac{\partial h(x_1^*, x_2^*)}{\partial x_1} = 0
\]

On the other hand,

\[
\frac{\partial f(x_1^*, x_2^*)}{\partial x_2} = \frac{\partial f(x_1^*, x_2^*)}{\partial x_2} \frac{\partial h(x_1^*, x_2^*)}{\partial x_2} = \lambda^* \frac{\partial h(x_1^*, x_2^*)}{\partial x_2}
\]

Therefore, we have three equations and three unknowns to determine their values.

\[
\frac{\partial f(x_1^*, x_2^*)}{\partial x_1} - \lambda^* \frac{\partial h(x_1^*, x_2^*)}{\partial x_1} = 0, \quad \frac{\partial f(x_1^*, x_2^*)}{\partial x_2} - \lambda^* \frac{\partial h(x_1^*, x_2^*)}{\partial x_2} = 0, \quad g(x_1^*, x_2^*) = 0
\]
The exercise demonstrates some key features of the approach. To simplify the condition in the final result, we introduce **Lagrange multiplier** \( \lambda \). The Lagrange multiplier not only allows a simple representation of the condition, but it also provides extra information on the constraint.

Let’s write the constraint as \( h(x) - b = 0 \). We can interpret \( b \) as the allowance of the resource associated with the constraint, and have an interesting economic interpretation of the Lagrange multiplier.

\[
\lambda^* = \frac{\partial f(x^*)}{\partial h(x^*)/\partial x} = \frac{\partial f(x^*)}{\partial x} \times \frac{1}{\partial h(x^*)/\partial x} \quad \text{for all } i = 1, \ldots, n
\]

\( \partial f(x^*)/\partial x_i \) is the marginal contribution of \( x_i \) to the maximal objective function and \( 1/(\partial h(x^*)/\partial x_i) \) measures the changes in \( x_i^* \) with an additional unit of \( b \). Since the value of \( \lambda^* \) is the same for all \( i \), \( \lambda^* \) is the value of additional \( b \) in terms of the objective function value.

In **MPE**, we can write a constraint as \( h(x) - b = 0 \) or \( -h(x) + b = 0 \). The necessary condition we find gives the same characterization of the maximum point with either of the forms, but with a different sign of Lagrange multipliers. If we interpret the Lagrange multiplier properly, it does not matter how to write the constraints.

There is also an important technical condition to note. The expression of \( \lambda^* \) involves the inverse of \( \partial g(x^*)/\partial x_i \) for some \( i \). To make this operation possible, the implicit function argument requires the derivative of constraints to have full (row) rank or \( \nabla g(x^*) \neq 0 \). This condition is called **regularity condition**, **constraint qualification condition**, or **nondegeneracy condition**.

If the constraints are affine functions, no other condition is required. With nonlinear constraints, we need additional conditions.

**DEFINITION 14-1 (Linearity Independent Constraint Qualification, LICQ)** The gradients of functional constraints are linearly independent at the optimal point. That is, the Jacobian of \( G(x^*) \) has full rank.

There are several versions of regularity condition and LICQ is one of the most stringent ones. With the constraint qualification condition satisfied, it is straightforward to extend the result to general **MPE.**
**THEOREM 14.1 (Lagrange Multiplier Theorem)** Suppose that \( f: \mathbb{R}^n \to \mathbb{R} \) and \( G: \mathbb{R}^n \to \mathbb{R}^m \) are continuously differentiable at \( \mathbf{x}^* \) and the constraint qualification condition is satisfied. If \( \mathbf{x}^* \) solves MPE, then there exists \( \mathbf{\lambda}^* \) such that

\[
\nabla f(\mathbf{x}^*) = \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*)
\]

**Proof.** Let’s divide the choice variable \( \mathbf{x} \) into two groups, \( \mathbf{x} = (\mathbf{y}, \mathbf{z}) \), \( \mathbf{y} \in \mathbb{R}^{n-m} \) and \( \mathbf{z} \in \mathbb{R}^m \). Suppose that \( \mathbf{x}^* \) is the solution to MPE. Since \( DH(\mathbf{x}^*) \) has full rank, it is possible to solve \( H(\mathbf{x}^*) = \mathbf{0} \) for a function \( \mathbf{z} = \mathbf{z}(\mathbf{y}) \) such that \( \mathbf{z}^* = \mathbf{z}(\mathbf{y}^*) \) and \( H(\mathbf{y}^*, \mathbf{z}(\mathbf{y}^*)) = \mathbf{0} \) in the neighborhood of \( (\mathbf{y}^*, \mathbf{z}^*) \). The problem can be rewritten as

\[
\max_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{y}, \mathbf{z}(\mathbf{y}))
\]

Since the composition function has a critical point at \( \mathbf{y}^* \), we have the first order condition,

\[
D_y f(\mathbf{y}^*, \mathbf{z}^*) + D_z f(\mathbf{y}^*, \mathbf{z}^*) D_y \mathbf{z}(\mathbf{y}^*) = \mathbf{0}
\]

Furthermore, if \( D_z H(\mathbf{y}^*, \mathbf{z}^*) \) is invertible, we can apply implicit function theorem to \( H(\mathbf{y}^*, \mathbf{z}(\mathbf{y})) = \mathbf{0} \) to get the expression for \( D_y H(\mathbf{y}^*) \),

\[
D_y H(\mathbf{y}^*, \mathbf{z}(\mathbf{y})) = D_y H(\mathbf{y}^*, \mathbf{z}^*) + D_z H(\mathbf{y}^*, \mathbf{z}^*) D_y \mathbf{z}(\mathbf{y}^*) = \mathbf{0}
\]

Substituting \( D_y \mathbf{z}(\mathbf{y}^*) = -[D_z G(\mathbf{y}^*, \mathbf{z}^*)]^{-1} D_y G(\mathbf{y}^*, \mathbf{z}^*) \) into the first order condition gives

\[
\frac{D_y f(\mathbf{y}^*, \mathbf{z}^*)}{1 \times (n-m)} = \frac{D_z f(\mathbf{y}^*, \mathbf{z}^*)}{1 \times m} \frac{[D_z H(\mathbf{y}^*, \mathbf{z}^*)]^{-1}}{m \times m} \frac{D_y H(\mathbf{y}^*, \mathbf{z}^*)}{m \times (n-m)}
\]

Let \( \mathbf{\lambda}^T = D_z f(\mathbf{x}^*) [D_z H(\mathbf{y}^*, \mathbf{z}^*)]^{-1} \) to have \( D_y f(\mathbf{x}^*) = \mathbf{\lambda}^T D_y H(\mathbf{y}^*, \mathbf{z}^*) \) or

\[
\nabla_y f(\mathbf{y}^*, \mathbf{z}^*) = \sum_{i=1}^{m} \lambda_i^* \nabla_y h_i(\mathbf{y}^*, \mathbf{z}^*)
\]

On the other hand,

\[
\frac{D_z f(\mathbf{y}^*, \mathbf{z}^*)}{1 \times m} = \frac{D_z f(\mathbf{y}^*, \mathbf{z}^*)}{1 \times m} \frac{[D_z H(\mathbf{y}^*, \mathbf{z}^*)]^{-1}}{m \times m} \frac{D_y H(\mathbf{y}^*, \mathbf{z}^*)}{m \times m} = \mathbf{\lambda}^T D_z H(\mathbf{y}^*, \mathbf{z}^*) = \sum_{i=1}^{m} \lambda_i^* \nabla_z h_i(\mathbf{y}^*, \mathbf{z}^*)
\]

Combining two equalities completes the proof.  

---

Math for Economists, Sogang University
The condition derived consists of \( n \) equations with \( m + n \) unknowns of \((\lambda, x)\). In principle, together with \( m \) equality constraints, we can determine the unique values of \((\lambda^*, x^*)\).

14.2 LAGRANGIAN METHOD

The substitution method always works as long as the constraint qualification conditions are met. However, the most popular approach to a constrained optimization problem is the method of Lagrange multiplier or Lagrangian method as an easier way to check optimality conditions, in particular, the second order sufficient condition.

**DEFINITION 14-2 (Lagrangian)** Lagrangian for the maximization problem is the function

\[
\mathcal{L}(\lambda, x) = f(x) - \lambda \cdot H(x) = f(x) - [\lambda_1 h_1(x) + \cdots + \lambda_m h_m(x)]
\]

The Lagrangian is the sum of the original objective function and the functional constraints weighted by Lagrange multiplier \( \lambda \).

The form of Lagrangian shows that a feasible solution to MPE also maximizes the Lagrangian for any \( \lambda \). On the other hand, to identify an optimal point directly from the Lagrangian, we need the converse. However, the converse is not true because \( \max_{x \in \mathbb{R}^n} \mathcal{L}(\lambda, x) \) depends on the value of \( \lambda \). We need \( \lambda^* \) such that

\[
\max_{x \in \mathcal{X}} f(x) = \max_{x \in \mathbb{R}^n} \mathcal{L}(\lambda^*, x)
\]

Let’s think of \( \lambda \) as a penalty for violating constraints. Clearly, the solution would be affected by \( \lambda \) and we can write the solution as a function of \( \lambda \), \( x = x(\lambda) \). If there exists \( \lambda^* \) with \( x^* = x(\lambda^*) \) such that \( \mathcal{L}(\lambda^*, x^*) = \max_{x \in \mathbb{R}^n} \mathcal{L}(\lambda, x) \) and \( H(x^*) = 0 \), \( x^* \) is also the solution to the original problem. The idea is formalized in the following theorem.

**THEOREM 14-2.** Suppose that there exists \( x^* \) and \( \lambda^* \) such that \( x^* \) maximizes \( \mathcal{L}(\lambda^*, x) \) with \( H(x^*) = 0 \). Then \( x^* \) solves the maximization problem.

**Proof.** Let \( X = \{x | H(x) = 0\} \).

\[
\max_{x \in X} f(x) = \max_{x \in \mathbb{R}^n} [f(x) - \lambda^* H(x)] \leq \max_{x \in \mathbb{R}^n} [f(x) - \lambda^* H(x)]
\]

The inequality holds because \( X \subset \mathbb{R}^n \). Since \( x^* \) maximizes \( \mathcal{L}(\lambda^*, x) \),
\[ \max[f(x) - \lambda^* H(x)] = f(x^*) - \lambda^* H(x^*) = f(x^*) \]

If the value of \( \lambda^* \) is properly determined, the maximum of \( \text{MPE} \) is identical to \( \max L(\lambda^*, x) \).

Note that the first order condition for \( \max L(\lambda^*, x) \) is

\[ D_xL(\lambda^*, x^*) = \nabla f(x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0 \]

Furthermore, we have

\[ D_xL(\lambda^*, x^*) = -H(x^*) = 0 \]

Therefore, the result is equivalent to that if \( DL(\lambda^*, x^*) = 0 \) and \( x^* \) is in fact the maximum point, then \( x^* \) is the solution to \( \text{MPE} \). **THEOREM 14-2** can be viewed as the converse of **THEOREM 14-1**: if \( x^* \) solves \( \text{MPE} \), there exists \( \lambda^* \) such that \( D_xL(\lambda^*, x^*) = 0 \). Therefore, we have \( DL(\lambda^*, x^*) = 0 \) at the maximum point, and the condition is necessary for the solution to \( \text{MPE} \) as well as the "optimal" Lagrange multipliers. Note that \( \lambda^* \) is not chosen to maximize \( L(\lambda, x^*) \) with respect to \( \lambda \). \( DL(\lambda^*, x^*) = 0 \) is simply an alternative way to write \( G(x^*) = 0 \).

The necessary condition can be stated with the derivative of Lagrangian.

**THEOREM 14-3 (Method of Lagrange Multiplier, Lagrange Method)** Suppose that \( f: \mathbb{R}^n \to \mathbb{R} \) and \( H: \mathbb{R}^n \to \mathbb{R}^m \) are continuously differentiable at \( x^* \) and the constraint qualification condition is satisfied. If \( x^* \) solves \( \text{MPE} \), then there exists \( \lambda^* \) such that \( DL(\lambda^*, x^*) = 0 \).

An alternative proof is provided to show directly that \( \nabla f(x^*) \) is a linear combination of \( \nabla G(x^*) \). For the sake of intuition, the proof considers \( \text{MPE} \) with a single constraint, but can be extended easily with projection matrix. Let’s start with a lemma that shows the relationship between two parallel vectors.

**LEMMA 14-4.** Consider a non-zero vector \( u \in \mathbb{R}^n \setminus 0 \) and let \( S = \{ s \in \mathbb{R}^n | s \cdot u = 0 \} \). If \( s \cdot v = 0 \) for all \( s \in S \), then there exists \( \lambda \in \mathbb{R} \) such that \( v = \lambda u \).

**Proof.** Let’s write \( v = v_0 + v_1 \). Since \( s \cdot u = 0 \) and \( s \cdot v = 0 \), choosing \( v_0 = \lambda u \) makes \( v_1 \cdot u = 0 \) or \( v_1 \in S \). Therefore,

\[ 0 = v_1 \cdot v = v_1 \cdot (v_0 + v_1) = v_1 \cdot v_0 + \|v_1\|^2 = v_1 \cdot (\lambda u) + \|v_1\|^2 = \|v_1\|^2 \]
This implies that $\|v_1\| = 0$ and $v = v_0 = \lambda u$. ■

$S$ is a hyperplane orthogonal to vector $u$. Therefore, any vector orthogonal to $S$ points in the same direction as $u$. Also, note that the lemma determines a unique value of $\lambda$. Because $v$ is in the span of $u$, the scalar projection of $v$ onto $u$ only scales the vector. Therefore, $\lambda = v \cdot u/\|u\|^2$.

**Proof of Theorem 14-1.** Let $x(t) = x^* + tv$ be a vector on the constraint surface through $x^*$ such that $x^* = x(0)$. By construction, the objective function attains its maximum at $t = 0$. Therefore,

$$Df(x(0)) = 0 \quad \text{or} \quad \nabla f(x^*) D x(0) = \nabla f(x^*) \cdot v = 0$$

On the other hand, since the tangent hyperplane to the surface at $x^*$ is also orthogonal to the gradients of constraints, $\nabla h(x^*)(x - x^*) = 0$ or $\nabla h(x^*) D x(0) = \nabla g(x^*) \cdot v = 0$. By constraint qualification condition, $\nabla h(x^*)$ has full rank or $\nabla h(x^*) \neq 0$. By Lemma 14-4, there exists $\lambda^*$ such that

$$\nabla f(x^*) = \lambda^* \nabla h(x^*) \quad \text{where} \quad \lambda^* = \frac{\nabla h(x^*)}{\|\nabla g(x^*)\|^2}$$

Constraint qualification condition is indispensable for the result. In the proof of Theorem 14-1, it is required to apply implicit function theorem while it is necessary to span a subspace containing $\nabla f(x^*)$ in this proof.

**Figure 14-1. Constraint Qualification Condition**

In Figure 14-1, the first figure shows that $\nabla f(x^*) = \lambda^* \nabla g(x^*)$ with $\lambda^*$. The second is that case that $\nabla f(x^*) = \lambda_1^* \nabla g_1(x^*) + \lambda_2^* \nabla g_2(x^*)$ with $\lambda_1^* < 0$ and $\lambda_2^* > 0$. And, in case of the third, it is not possible to express $\nabla f(x^*)$ with a linear combination of $\nabla g_1(x^*)$ and $\nabla g_2(x^*)$ and the necessary condition fails to describe the optimal point.
EXAMPLE 14-1 by Vincent Crawford

$$\max_{x_1, x_2} -(x_1^2 + x_2^2) \quad \text{subject to} \quad (x_1 - 1)^3 - x_2^2 = 0$$

Since $x_2^2 \geq 0, \ (x_1 - 1)^3 < 0$ or $x_1 \leq 1$. The solution is $(1,0)$. However, since $\nabla h(x^\ast) = (3(x_1^\ast - 1)^2, -x_2^\ast) = (0,0)$, constraint qualification is not met as shown in the first axes in Figure 14-2. First order conditions are $D_1 \mathcal{L} = 2x_1^\ast - 3\lambda^\ast(x_1^\ast - 1)^2 = 0$ and $D_2 \mathcal{L} = 2x_2^\ast + 2\lambda^\ast x_2^\ast = 0$. From $D_2 \mathcal{L} = 0$, $x_2^\ast = 0$ or $\lambda^\ast = -1$. However, there is no real root for $3(x_1^\ast)^2 - 4x_1^\ast + 3 = 0$. Lagrangian method fails.

**Figure 14-2. Violation of Constraint Qualification Condition**

14.3 SUFFICIENT CONDITIONS

Like unconstrained optimization problems, if an objective function is concave in the neighbor of the point satisfying the first order conditions, it is a local maximum point. However, since choices are limited to the constraints, the condition could be relaxed to the extent that the objective is concave on the feasible set around the solution or on the tangent lines to constraints at the solution.

The equations of tangent hyperplanes to the constraints are $DH(x^\ast)(x - x^\ast) = 0$ or $\nabla h_i(x^\ast)(x - x^\ast) = 0, \ i = 1, \cdots, m$. Note that what we want is $(x - x^\ast)^TA(x - x^\ast)$ is negative definite locally, and THEOREM 12-4 provides a way to check the concavity of a function on the set of $\{x|DH(x^\ast)(x - x^\ast) = 0\}$. The second derivative of the objective along the vector $(x - x^\ast)$ is given by the following theorem.
**THEOREM 14-5.** Assume that \( x^* \) and \( \lambda^* \) satisfies the conditions in THEOREM 14-3. Define \( x(t) \) such that \( H(x(t)) = 0 \) with \( x(0) = 0 \) and \( x'(0) = v \), then second order sufficient condition is

\[
\frac{d^2}{dt^2} f(x(0)) = v^T \left[ D^2 f(x^*) - \sum_{i=1}^{n} \lambda_i^* D^2 h_i(x^*) \right] v = v^T D^2 \mathcal{L}(\lambda^*, x^*) v < 0
\]

**Proof.**

\[
\frac{d}{dt} f(x(t)) = Df(x(t))x'(t) = \sum_{i=1}^{n} \frac{\partial f(x(t))}{\partial x_i} x'_i(t)
\]

\[
\frac{d^2}{dt^2} f(x(t)) = \sum_{i=1}^{n} \left[ \frac{d}{dt} \frac{\partial f(x(t))}{\partial x_i} \right] x'_i(t) + \sum_{i=1}^{n} \frac{\partial f(x(t))}{\partial x_i} x''_i(t)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f(x(t))}{\partial x_i \partial x_j} x'_i(t) x'_j(t) + Df(x(t)) x''(t)
\]

\[
\frac{d^2}{dt^2} f(x(0)) = [x'(0)]^T D^2 f(x(0)) x'(0) + \sum_{i=1}^{n} Dg_i(x(0)) x''(0)
\]

Repeat the steps to \( h_i(x(t)) = 0 \) to get

\[
\frac{d}{dt} h_i(x(t)) = \sum_{i=1}^{n} \frac{\partial h_i(x(t))}{\partial x_i} x'_i(t) = 0
\]

\[
\frac{d^2}{dt^2} h_i(x(t)) = \sum_{i=1}^{n} \left[ \frac{d}{dt} \frac{\partial h_i(x(t))}{\partial x_i} \right] x'_i(t) + \sum_{i=1}^{n} \frac{\partial h_i(x(t))}{\partial x_i} x''_i(t)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 h_i(x(t))}{\partial x_i \partial x_j} x'_i(t) x'_j(t) + Dh_i(x(t)) x''(t)
\]

\[
\frac{d^2}{dt^2} h_i(x(0)) = [x'(0)]^T D^2 h_i(x(0)) x'(0) + Dh_i(x(t)) x''(t) = 0
\]

\[
\lambda_i^* D h_i(x(t)) x''(t) = -\lambda_i^*[x'(0)]^T D^2 h_i(x(0)) x'(0)
\]

Since \( x(0) = x^* \) and \( x'(0) = v \), substituting this equality into the expression for the second derivative of \( f(x(t)) \) gives the result. \( \square \)

To check the second order condition on Lagrangian, since \( x'(0) = v \), let’s apply THEOREM 14-5 to
\[ D^2 \mathcal{L}(\lambda^*, x^*) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_1^2} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_1 \partial \lambda_m} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_1 \partial x_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_1 \partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_m \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_m \partial \lambda_m} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_m \partial x_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial \lambda_m \partial x_n} \\ \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_1 \partial \lambda_m} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_n \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_n \partial \lambda_m} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_n \partial x_n} \end{bmatrix} \]

The expressions in border parts correspond to \( x'(0) = v \) and \( \forall h_i(x^*) \cdot v = 0 \) the matrix that we apply the criteria in \textsc{Theorem 12-4} is:

\[
 D^2 \mathcal{L}(\lambda^*, x^*) = \begin{bmatrix} 0 & \cdots & 0 & -\frac{\partial h_1(x^*)}{\partial x_1} & \cdots & -\frac{\partial h_1(x^*)}{\partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ 0 & \cdots & 0 & -\frac{\partial h_m(x^*)}{\partial x_1} & \cdots & -\frac{\partial h_m(x^*)}{\partial x_n} \\ -\frac{\partial h_1(x^*)}{\partial x_1} & \cdots & -\frac{\partial h_m(x^*)}{\partial x_1} & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_1^2} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \ddots & \ddots \\ -\frac{\partial h_1(x^*)}{\partial x_n} & \cdots & -\frac{\partial h_m(x^*)}{\partial x_n} & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 \mathcal{L}(\lambda^*, x^*)}{\partial x_n \partial x_n} \end{bmatrix}
\]

Note that the requirement is slightly less stringent than quasi-concavity. If a simple monotone transformation can convert the objective function into a concave function, that is sufficient for a local maximum. For instance, the log transformation of the Cobb-Douglas function is always concave as long as the domain is \( \mathbb{R}_{++}^n \).

\textbf{Example 14-2.}

\[
\max_{xy} x^2 + 4y^2 \\
\text{subject to } x^2 + y^2 = 1
\]

The first order condition is

\[
-(x^*)^2 + (y^*)^2 - 1 = 0 \\
2x^* - \lambda^*2x^* = 0 \\
8y^* - \lambda^*2y^* = 0
\]

Math for Economists, Sogang University
Solutions to the equation system are \( x^* = 0, \ y^* = \pm 1, \ \lambda^* = 4 \) or \( x^* = \pm 1, \ y^* = 0, \ \lambda^* = 1 \).

Since the Lagrangian is
\[
L(x, y, \lambda) = x^2 + 4y^2 - \lambda(x^2 + y^2 - 1)
\]
The bordered Hessian is
\[
D^2L(x, y, \lambda) = \begin{bmatrix}
0 & -2x & -2y \\
-2x & 2 - 2\lambda & 0 \\
-2y & 0 & 8 - 2\lambda
\end{bmatrix}
\]
Note that the way to write the constraint does not affect the result.

At the first pair of the solution,
\[
D^2L(0, \pm 1, 4) = \begin{bmatrix}
0 & 0 & \mp 2 \\
0 & -6 & 0 \\
\mp 2 & 0 & 0
\end{bmatrix}, \quad \text{det}(D^2L) = -(6 \times (\mp 2)^2) = -24 < 0.
\]
And thus \((x^*, y^*, \lambda^*) = (0, \pm 1, 4)\) are local maximum points.

At the second pair of the solution,
\[
D^2L(\pm 1, 0, 1) = \begin{bmatrix}
0 & \mp 2 & 0 \\
\mp 2 & 0 & 0 \\
0 & 0 & 6
\end{bmatrix}, \quad \text{det}(D^2L) = -(6 \times (\mp 2)^2) = 24 > 0.
\]
And thus \((x^*, y^*, \lambda^*) = (\pm 1, 0, 1)\) are local minimum points.

The values of the function at those points are \( f(0, \pm 1) = 4 \) and \( f(\pm 1, 0) = 1 \).

### 14.4 Comparative Statics

Because of constraints, changes in parameters could have a nontrivial effect on the choice set. In a standard utility maximization problem, a price increase shrinks the budget set and changes relative prices. Unlike the unconstrained case, \( Df(x^*) \) is not zero anymore, and the secondary effect usually survives.
\[
V'(a) = D_xf(x^*; a)D_a x(a) + D_a f(x^*, a)
\]
By the first order condition, we have $Df(x^*) = \lambda^*DH(x^*)$. And differentiating the constraints of $H(x(a), a) = 0$ gives

$$D_xH(x^*; a)D_a x(a) + D_a H(x^*, a) = 0$$

By combining the two equations, we have

$$D_xf(x^*; a)D_a x(a) = \lambda^*DH(x^*)D_a x(a) = -\lambda^* D_a H(x^*, a).$$

Substitution yields

$$V'(a) = D_a f(x^*, a) - \lambda^* D_a H(x^*, a).$$

The final expression is similar to the case of unconstrained maximization problems replacing the objective with the Lagrangian.

On the other hand, suppose that a constraint is changed from $h_i(x, a) = 0$ to $h_i(x, a) = b_i$. We can apply the envelope theorem to approximate the effect of the change on the maximum value of the problem. Since $h_i(x) - b_i = 0$, $D_b f(x^*) = 0$ and $D_b H(x^*) = -1$,

$$V'(b_i) = \lambda_i^*$$

This equation confirms the properties of $\lambda_i^*$ discussed in Section 11.1. That is, the decision maker is willing to pay $\lambda_i^*$ for an additional unit of the $i$th resource. For this reason, $\lambda^*$ is called **shadow price** or **imputed value** of the resource.

Although the sign of Lagrange multiplier depends on the way to write the Lagrangian, there is a convention. In economic applications, objective functions usually have positive derivatives and the larger $b$ we have the greater the maximum. The first order condition is written in a way that $\nabla g_i(x^*)$ directs the upper contour of the objective so that $\lambda_i^* > 0$. As a matter of fact, such interpretation is natural in inequality constraints, with which the more choices we have, we never get worse off.

**Implicit Function Theorem**

The process is identical to those for the unconstrained problems. Do not forget to consider all variables including $\lambda$. Otherwise, the result could be misleading.
15 Optimization with Inequality Constraints

Consider a maximization problem:

\[
\max_{x \in \mathbb{R}^n} f(x; a)
\]

\[
\text{MPI :}
\begin{align*}
\text{subject to } & G(x; a) \leq b \\
& H(x; a) = c
\end{align*}
\]

where \( f: \mathbb{R}^n \to \mathbb{R} \), \( G(x) = (g_1(x), \ldots, g_k(x)) \) and \( H(x) = (h_1(x), \ldots, h_m(x)) \) are continuously differentiable functions, and \( b \in \mathbb{R}^k \) and \( c \in \mathbb{R}^m \). \( k \) is the number of inequality constraints, and \( K \) denotes the set of inequality constraints. The analysis in this chapter often ignores equality constraint because an extension to general problem is straightforward.\(^7\)

To see the nature of inequality constraint clearly, let \( x^* \) be an optimum and divide the constraints into the ones with \( g_i(x^*) < b_i \) and with \( g_j(x^*) = b_j \). If \( g_i(x^*) < b_i \) at the solution, the constraint does not play any role and a problem without the constraint would yield the same result. If the solution is an interior point of the constraint set, we say that the constraint is not binding or it is a non-binding constraint. If constraints satisfy with equality, we call them binding or active constraints. Constructing MPI only with binding constraints is an equality constrained problem.

If the binding constraints can be identified a priori, we can convert MPI into MPE excluding non-binding constraints. Then we should have the first order conditions identical to those of MPE, or

\(^7\) An equality constraint can be expressed with two inequalities of \( g_i(x) \leq 0 \) and \( -g_i(x) \leq 0 \), and thus inequality constraint is a more general form. Or alternatively, we can combine multipliers such as \( \mu_j^+ g_i(x) + \mu_j^- (-g_i(x)) = \mu_j g_i(x) \), and consider them as a single inequality constraint. In any case, however, because of the constraint qualification condition, equality and inequality constraints are usually considered separately.
\[ \nabla f(x^*) = \sum_{i \in j} \mu_i^* \nabla g_i(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x^*) \]

and

\[ g_i(x^*) = 0 \quad \text{for all } i \in B \]
\[ h_j(x^*) - c_j = 0 \quad \text{for all } j \in M \]

where \( B \subset K \) is the set of binding constraints. Although there is no obvious way to identify binding constraints, in principle, we can solve MPI by solving MPE with all possible combinations of (binding and nonbinding) constraints, say \( 2^m \) problems, and selecting a feasible maximum point with the largest objective function value. Obviously, however, this approach could be not only costly but also ill-organized to identify binding constraints.

**Example 15-1.**

\[ \max_{x, y \in \mathbb{R}} -(x - a)^2 - 2(y - 1)^2 \]
subject to \( x + 2y \leq 3 \)
\[ x \geq y \]

We rearrange the constraint to have “right” signs of the Kuhn-Tucker multipliers.

\[ \max_{x, y \in \mathbb{R}} -(x - 4)^2 - 2(y - 1)^2 \]
subject to \( x + 2y \leq 3 \)
\[ -x + y \leq 0 \]
\[ \max_{x, y \in \mathbb{R}} -(x - 4)^2 - 2(y - 1)^2 \]
subject to \( x + 2y \leq 3 \)
\[ x \geq y \]

We rearrange the constraint to have “right” signs of the Kuhn-Tucker multipliers.
subject to \( x + 2y \leq 3 \)
\(-x + y \leq 0 \)

We have 4 possible combinations of binding constraints. Let's solve them one by one to find the optimal point.

case 1. With no constraint

\[
\max_{x,y \in \mathbb{R}} - (x - 4)^2 - 2(y - 1)^2
\]

The first order condition is

\[
-2(x^* - 4) = 0
\]
\[-4(y^* - 1) = 0 \]
\[x^* = 4, \quad y^* = 1 \]

That does not satisfy the first and second constraints. There is a binding constraint.

Case 2. only the first constraint is binding, \( x + 2y = 3 \)

\[
\max_{x,y \in \mathbb{R}} - (x - 4)^2 - 2(y - 1)^2 \]

subject to \( x + 2y - 3 = 0 \)

The first order condition is

\[
-(x^* + 2y^* - 3) = 0
\]
\[-2(x^* - 4) - \lambda^* = 0 \]
\[-4(y^* - 1) - 2\lambda^* = 0 \]
\[x^* = 3, \quad y^* = 0, \quad \lambda^* = 2 \]

The solution is not feasible.

Case 3. only the second constraint is binding, \( x = y \)

\[
\max_{x,y \in \mathbb{R}} - (x - 4)^2 - 2(y - 1)^2 \]

subject to \( x - y = 0 \)

The first order condition is
\[ -(x^* - y^*) = 0 \]
\[ -2(x^* - 4) - \lambda^* = 0 \]
\[ -4(y^* - 1) + \lambda^* = 0 \]
\[ x^* = y^* = 2, \quad \lambda^* = 4 \]

The solution is not feasible.

Case 4. Both are binding.

\[
\max_{x, y \in \mathbb{R}} -(x - 4)^2 - 2(y - 1)^2
\]
subject to \( x + 2y - 3 = 0 \)
\( -x + y = 0 \)

The first order condition is
\[ -(x^* + 2y^* - 3) = 0 \]
\[ -(x^* - y^*) = 0 \]
\[ -2(x^* - 4) - \lambda_1^* - \lambda_2^* = 0 \]
\[ -4(y^* - 1) - 2\lambda_1^* + \lambda_2^* = 0 \]
\[ x^* = y^* = 1, \quad \lambda_1^* = 2, \quad \lambda_2^* = 4 \]

The maximum value is \(-9\).

Therefore, the maximum is \(-1\) and only the first constraint is binding. We can conclude the solution is \(x^* = 3, \ y^* = 0, \ \lambda_1^* = 2, \ \lambda_2^* = 0\). (The objective function is concave and the second order condition is satisfied.)

In principle, MPI can be solved with the Lagrange method. One might think that the rest of the results in this chapter is for computational advantages, but there are more. They provide a significantly better understanding of optimization problems, especially about its dual structure and multipliers.

Under appropriate conditions, the first order condition for optimality can be written as a pair of similar conditions known as **Karush-Kuhn-Tucker condition**.
\[
\frac{\nabla f(x^*)}{f} = \sum_{i=1}^{k} \mu_i^* \nabla g_i(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x^*)
\]

and

\[
\mu_i^*(g_i(x^*) - b_i) = 0 \quad \text{for all } i \in K
\]

\[
\mu_i^* \geq 0 \quad \text{for all } i \in K
\]

\[
g_i(x^*) - b_i \leq 0 \quad \text{for all } i \in K
\]

\[
h_j(x^*) - c_j = 0 \quad \text{for all } j \in M
\]

The second condition is satisfied if either \( \mu_i^* = 0 \) or \( g_i(x^*) = 0 \). Now we have \( n + k + m \) equations with the same number of unknowns.

The second condition is called a **complementary slackness condition**. The **Kuhn-Tucker multiplier** \( \mu_i \) has a same interpretation of Lagrange multipliers in MPE. The multipliers associated with non-binding constraints are zero. Also, if your willingness-to-pay is positive for more resource \( (\mu_i > 0) \) it must be the case that the constraint is binding. It is possible but rare to have both \( \mu_i = 0 \) and \( g_i(x^*) = 0 \). This happens when the \( i \)th resource is “just enough” and cannot be used anymore. In any case, we have the same expression both for binding and nonbinding constraints.

**Example.** Let’s solve the exercise above using the new conditions.

\[
\max_{x,y \in \mathbb{R}} -(x - 4)^2 - 2(y - 1)^2
\]

subject to

\[
x + 2y \leq 3
\]

\[
-x + y \leq 0
\]

The optimality condition is

\[
-2(x^* - 4) = \mu_1^* + \mu_2^*
\]

\[
-4(y^* - 1) = 2\mu_1^* - \mu_2^*
\]
And

\[ \mu_1^* (x^* + 2y^* - 3) = 0 \\
\mu_2^* (x^* - y^*) = 0 \]

Compare to the case 4 in the previous example, the two equality constraints, \( h_i(x^*) - c_i = 0 \), are replaced with complementary slackness conditions, \( \mu_i^* (g_i(x^*) - b_i) = 0 \). In practice, to identify the set of binding constraints, we use \( \mu_i^* \) by checking whether it is zero or positive.

Case 1. \( \mu_1^* = \mu_2^* = 0 \) \( \Rightarrow \) \( x^* = 4, y^* = 1 \) which does not satisfy constraints.

Case 2. \( x^* + 2y^* - 3 = 0 \) and \( \mu_2^* = 0 \) \( \Rightarrow \) \( x^* = 3, y^* = 0, \mu_2^* = 2 \), which violates the second.

Case 3. \( \mu_1^* = 0 \) and \( x^* - y^* = 0 \) \( \Rightarrow \) \( x^* = y^* = 2, \mu_1^* = 4 \), which violates the first.

Case 4. \( x^* + 2y^* - 3 = 0 \) and \( x^* - y^* = 0 \) \( \Rightarrow \) \( x^* = y^* = 1, \mu_1^* = 2, \mu_2^* = 4 \)

In terms of computation, there is not much improvement.

The key result is derived over three sections. In Section 1, the necessary conditions for optimality are derived. Section 2 provides the fundamental results in linear programming to demonstrate the nature of the optimization problem with inequality constraints. The results are extended to MPI in Section 3.

15.1 NECESSARY CONDITIONS

The optimality essentially requires that there is no improvement by changing choice variables in the feasible neighborhood. In UMP, the optimality condition is based only on the slope of the objective function because a choice can be changed to all possible directions. In MPE, a feasible set is an intersection of level sets and the tangent hyperplane is well described by the gradients. In MPI, it is an intersection of upper or lower contour sets. Since there is an additional dimension to move toward, we need additional information to describe the possible changes of choice variables in addition to the gradients of constraints.
**Definition 15-1 (Feasible Vector)** Given a set $X$ and a point $x_0 \in X$, a vector $v$ is feasible at $x_0$ in $X$ if there exists $\varepsilon > 0$ such that $x_0 + tv \in X$ for all $t \in [0, \varepsilon)$.

![Diagram showing feasible vectors for nonlinear and linear constraints]

**Figure 15-1 Feasible Vectors for Nonlinear and Linear Constraints**

The existence of a feasible vector requires an interior point in the feasible set and there might not exist a feasible vector in a choice set with a nonlinear equality constraint. This is one of the reasons why equality and inequality constraints are considered separately.

In MPE, the choice set is often described by $x = x_0 + tv$ with $\nabla H(x_0) \cdot v = 0$ or $v$ is orthogonal to $\nabla h_i(x_0)$ for all $i = 1, \ldots, m$. With inequality constraints, for any feasible vector at a boundary point, it makes an obtuse angle with the gradient of every binding constraint.

**Theorem 15-1.** Suppose that the constraints are linear such that the feasible set is $X = \{x | Ax \leq b\}$. A vector $v$ is feasible in $X$ at the boundary point of $x_0$ if and only if $A_i \cdot v \leq 0$ for every binding constraints, where $A_i$ is the $i$th row of $A$ representing the $i$th constraint.

A feasible $v$ at $x_0$ implies $A_i \cdot (x_0 + tv) \leq b_i$. At the boundary, $A_i \cdot x_0 = b_i$ and thus

$$A_i \cdot v \leq 0.$$

Conversely, for every binding constraint $i$, $A_i \cdot x^* = b_i$. Since $A_i \cdot v \leq 0$, $A_i \cdot (x^* + tv) \leq b_i$.

On the other hand, if the $i$th constraint is not binding, it must be that $A_i \cdot x^* < b_i$ and $A_i \cdot (x^* + tv) < b_i$ for a sufficiently small $t$ and every $v$.

When $x^*$ is a local maximum point, then it is obvious that moving toward a feasible direction should not increase the maximal value. That is, $D_v f(x^*) \leq 0$ or $\nabla f(x^*) \cdot v \leq 0$ for any
feasible vector $v$. The next theorem identifies the relative position of $\nabla f(x^*)$ to the span of $\nabla G(x^*)$.

**THEOREM 15-2 (Farkas’ Lemma)** Let $A \in \mathbb{R}^{n \times m}$ and $w \in \mathbb{R}^n$. Then exactly one of the following two statements holds:

(i) There exists $\lambda \in \mathbb{R}^m_+$ such that $A\lambda = w$.

(ii) There exists $p \in \mathbb{R}^n$ such that $A^T p \geq 0$ and $w \cdot p < 0$.

**Proof.** Consider a set $Z = \{z | z = A\lambda, \lambda \geq 0\}$. $w \in \mathbb{R}^n$ must be either in $Z$ or not in $Z$. The first condition states that $w \in Z$.

If $w \notin Z$, $A\lambda = w$ has no solution. Since $Z$ is convex, there is a hyperplane $p \cdot (z - w) = c$ that has $Z$ on the one side and the vector $w$ on the other side. That is, $p \cdot z \geq 0$ for $z \in Z$ and $w \cdot p < 0$. To complete the proof, we have to show $A^T p \geq 0$. Note that $p \cdot z = p \cdot A\lambda = A^T p \cdot \lambda \geq 0$ for all $\lambda \geq 0$. If the $i$th component of $A^T p$ is negative, for a sufficiently large $a > 0$, we have $A^T p \cdot \lambda < 0$ with $\lambda = ae_i$.

Note that they cannot hold simultaneously. If both hold, there exist a vector $p \in \mathbb{R}^n$ such that $A^T p \geq 0$ and $w \cdot p < 0$, and a $\lambda \in \mathbb{R}^m_+$ such that $A\lambda = w$. Then we have a contradiction that $p \cdot A\lambda = p \cdot w \geq 0$ because $A^T p \geq 0$ and $\lambda \geq 0$.

The theorem implies either that $w$ belongs to the convex cone, $Z = \{z | z = A\lambda, \lambda \geq 0\}$ or that there exists a hyperplane that separates the convex cone and $w$.

Let

$$A_{n \times m} = [\nabla g_1(x^*) \ldots \nabla g_m(x^*)], \quad w = \nabla f(x^*).$$

Since the second condition is ruled out by THEOREM 15-1, THEOREM 15-2 shows that $\nabla f(x^*)$ is a vector in the convex cone of $\nabla G(x^*)$. That is, the gradient of the objective can be expressed as a positive linear combination of gradients of constraints.

**THEOREM 15-3 (Necessary Conditions)** Suppose $x^*$ solves MPI with $G(x) = Ax$ and $f(x)$ is differentiable at $x^*$. Then there exists $\lambda_i^* \geq 0$ such that

$$\nabla f(x^*) = \sum_{i \in M} \lambda_i^* \nabla g_i(x^*) - b_i = \sum_{i \in M} \lambda_i^* \nabla g_i(x^*)$$

or
\[ \nabla f(x^*) = \sum_{i \in M} \lambda_i^* \nabla(A_i x^* - b_i) = \sum_{i \in M} \lambda_i^* A_i \]

**Proof.** If the maximum point does not satisfy the property, then by **THEOREM 15-2**, there is a \( p \in \mathbb{R}^n \) such that \( A^T(-p) \leq 0 \) and \( \nabla f(x^*) \cdot (-p) > 0 \). Since \( A_i(-p) \leq 0 \) for all \( i \), \((-p)\) is feasible by **THEOREM 15-1**. The property of derivative imply that, for sufficiently small \( t > 0 \),

\[ A_i(x^* + t(-p)) \leq b_i \quad \text{and} \quad f(x^* + t(-p)) > f(x^*) \quad \text{for all} \ i \]

which is contrary to the hypothesis that \( f(x^*) \) is the maximum. \( \square \)

This theorem gives the necessary conditions for the solution to MPI but limited to the case of linear constraint set because **THEOREM 15-1** considers only linear constraints. In fact, it is not directly applicable to nonlinear constraints.

**EXAMPLE 15-2 JOEL SOBEL AND JOEL WATSON, UCSD**

\[
\max_{x_1, x_2} f(x_1, x_2) = x_1 \\
\text{subject to} \quad -x_1 \leq 0 \\
\quad -x_2 \leq 0 \\
\quad -(1 - x_1)^3 + x_2 \leq 0
\]

It is apparent that the solution to this problem is \((x_1^*, x_2^*) = (1,0)\). By **THEOREM 15-3**,

\[
\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda_1^* \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \lambda_2^* \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \lambda_3^* \begin{pmatrix} 3(1 - x_1^*)^2 \\ 1 \end{pmatrix} \\
\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \lambda_1^* \begin{pmatrix} -1 \\ 0 \end{pmatrix} + \lambda_2^* \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \lambda_3^* \begin{pmatrix} 0 \\ 1 \end{pmatrix}.
\]

But it has no non-negative solution for \( \lambda_i^* \)'s when \((x_1^*, x_2^*) = (1,0)\). The complimentary slackness condition implies that \( \lambda_1^* = 0 \) when \( x_1^* = 1 \), which again means that the conditions in **THEOREM 15-3** need not hold if \( X \) is not described by linear inequalities.
The problem arises from the characterization of feasible vectors. The gradients of binding constraints at \((1,0)\) are \((0,1)\) from \(x_2 \geq 0\) and \((0,-1)\) from \(x_2 - (1 - x_1)^3 \leq 0\). One of the feasible vectors \(v = (1,0)\) makes a zero dot product with both of these gradients, but does not satisfy the definition of a feasible direction. We simply rule out such peculiar cases. The following two conditions are popular ones.

**Definition 15-2 (Mangasarian-Fromovitz Constraint Qualification Condition, MFCQ)** Suppose \(G(x)\) and \(H(x)\) are differentiable and \(X = \{x \in \mathbb{R}^n | G(x) \leq b\}\). The constraint qualification is satisfied at \(x^* \in X\) if the gradients of equality constraints are linearly independent and there exists a a feasible vector \(v\) such that \(\nabla h_j(x^*) \cdot v = 0\) and \(\nabla g_i(x^*) \cdot v < 0\) for all \(i = 1, \cdots, k\) and \(j = 1, \cdots, m\).

**Definition 15-3 (Slater Constraint Qualification Condition)** For MPI where \(f(x)\) is concave, \(g_i(x)\)'s are convex, and \(h_j(x)\)'s are affine functions, there exists \(x \in \mathbb{R}^n\) such that \(H(x) = 0\) and \(G(x) < 0\). Simply, \(X = \{x \in \mathbb{R}^n | G(x) \leq b, H(x) = c\}\) has an interor point.

LICQ implies MFCQ, and Slater condition restricts the functional forms. One of LICQ, MFCQ, and Slater’s condition is sufficient for the results in this chapter, and we refer to them simply as a constraint qualification condition.

Note that some constraint qualification condition will not hold if an equality constraint is presented in inequality constraints (that is, one constraint is of the form \(g_i(x) \leq b_i\) and
another is \(-g_i(x) \leq -b_i\). This is why the necessary conditions for optimality considers equality constraints separately.

Clearly, the example does not satisfy the constraint qualification. The constraint qualification holds if the gradients of binding constraints are linearly independent. Also, if a feasible set is convex and has an interior point, the condition is also qualified.

**THEOREM 15-4 (Karush-Kuhn-Tucker Condition)** Suppose \(x^*\) solves \(\text{MPI}\) and \(f(x), G(x),\) and \(H(x)\) are differentiable at \(x^*\). If the constraint qualification is satisfied at \(x^*\), then there exists \(\mu^* \geq 0\) and \(\lambda^*\) such that

\[
\begin{align*}
\text{Stationarity condition:} \quad & \nabla f(x^*) = \sum_{i=1}^k \mu_i^* \nabla g_i(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) \\
\text{Complementary slackness condition:} \quad & \mu^* \cdot (G(x^*) - b) = 0 \\
\text{Primal feasibility condition:} \quad & G(x^*) - b \leq 0 \\
& H(x^*) - c = 0 \\
\text{Dual feasibility condition:} \quad & \mu^* \geq 0
\end{align*}
\]

To characterize the solution of \(\text{MPI}\) with nonlinear constraints, constraint qualification condition is essential. Note that the condition is different from the one for the equality constrained problems. If there is no inequality constraint, the conditions are the same as those for the Lagrangian method.

**Proof.** The proof of the stationarity condition and dual feasibility condition is identical to the proof of **THEOREM 15-3** replacing \(A_i\) and **THEOREM 15-1** with \(\nabla g_i(x^*)\) and the constraint qualification condition, respectively.

The complementary slackness condition is equivalent to that a component of \(\mu^*_i\) is zero only if the corresponding constraint is not binding. Since \(x^*\) is a local maximum point, it is also a local maximum point over the set of binding constraints and \(\mu^*_i \geq 0\) for all \(i \in B\). Therefore, to have the stationarity condition, we must have \(\mu^*_i = 0\) for all \(j \in K \setminus B\) or \(g_j(x^*) < 0\). ■
The next two sections examine the properties of MPI in detail to explore the implications of the conditions.

15.2 **LINEAR PROGRAMMING**

Duality is a relationship between a given linear programming problem and another related linear programming problem. The theory is important for those who want to understand the shadow price interpretation of multipliers and to take advantage of computational efficiency. Also the duality provides a good guide to understand MPI.

A linear programming problem is an optimization problem with a linear objective function with linear inequality constraints and non-negativity condition that choice variables cannot be negative.

**EXAMPLE 15-3.** Bradley, Hax, and Magnanti, 1977, *Applied Mathematical Programming*

There is a firm producing three products. The revenue and resource constraints are given by:

\[
\begin{align*}
\max_{x_1, x_2, x_3} & \quad 6x_1 + 14x_2 + 13x_3 \\
\text{subject to} & \quad x_1/2 + 2x_2 + x_3 \leq 24 \\
& \quad x_1 + 2x_2 + 4x_3 \leq 60 \\
& \quad x_1 \geq 0, \; x_2 \geq 0, \; x_3 \geq 0
\end{align*}
\]

The Lagrangian is

\[
\mathcal{L}(\mu, \eta, x) = 6x_1 + 14x_2 + 13x_3 - \mu_1(x_1/2 + 2x_2 + x_3 - 24) - \mu_2(x_1 + 2x_2 + 4x_3 - 60) - \eta_1(-x_1) - \eta_2(-x_2) - \eta_3(-x_3)
\]

Since constraints are linear, by **THEOREM 15-3**, if \( x^* \) is the solution, there exists \( \mu^* \geq 0 \) and \( \eta^* \geq 0 \) such that

\[
6 = \mu_1^*/2 + \mu_2^* - \eta_1^*
\]
14 = 2 \mu_1^* + 2 \mu_2^* - \eta_2^*
13 = \mu_1^* + 4 \mu_2^* - \eta_3^*

And

\mu_1^*(x_1^*/2 + 2x_2^* + x_3^* - 24) = 0
\mu_2^*(x_1^* + 2x_2^* + 4x_3^* - 60) = 0
\eta_1^*(-x_1^*) = 0
\eta_2^*(-x_2^*) = 0
\eta_3^*(-x_3^*) = 0

The solution is

x_1^* = 36, \quad x_2^* = 0, \quad x_3^* = 6, \quad \mu_1^* = 11, \quad \mu_2^* = \frac{1}{2}, \quad \eta_1^* = \eta_3^* = 0. \quad \eta_2^* = 20

With the solution, we can check the profit maximization condition, \( MR_i = MC_i \). First, the coefficients of the objective function are the price of products, or marginal revenue. According to the shadow price interpretation, the marginal cost of product \( i \) is

\[ MC_i = \sum_{j=1}^{2} a_{ij} \mu_j^* \]

where \( a_{ij} \) is the amount of input \( j \) to produce a unit of product \( i \). Therefore, the stationarity conditions are \( MR_1 = MC_1 \) and \( MR_2 = MC_2 \). The shadow prices also allow to calculate the value of total resources:

\[ v = 11 \times 24 + \frac{1}{2} \times 60 = 294 \]

Which is exactly equal to the maximized value of the objective function, the firm’s revenue. The shadow price shows that the first resource and the second have an imputed worth of 264 and 30, respectively. A decision maker can use an internalize pricing system to value resources and to allocate them accordingly.

Now let’s consider if it is possible to determine shadow prices directly without solving the production decision problem. Imagine that the firm does its own resources but has to purchase them so that the price is the same as the shadow price. The firm’s objective turns to minimize the cost under the constraints that marginal cost is not less than the marginal revenue. Note
that this is the way to make a decision in a market. S/He likes to find the upper bound of the resource price that is acceptable or makes at least break-even.

\[
\begin{align*}
\min_{\mu_1, \mu_2} & \quad 24\mu_1 + 60\mu_2 \\
\text{subject to} & \quad \mu_1/2 + \mu_2 \geq 6 \\
& \quad 2\mu_1 + 2\mu_2 \geq 14 \\
& \quad \mu_1 + 4\mu_2 \geq 13 \\
& \quad \mu_1 \geq 0, \mu_2 \geq 0
\end{align*}
\]

Interestingly, if we use \(x_i\) for the multiplier for the first three constraints, we have the Lagrangian identical to the original problem except those parts for the last two non-negativity conditions.

\[
-\mathcal{L}(\mathbf{x}, \xi, \mu) = 24\mu_1 + 60\mu_2 - x_1(\mu_1/2 + \mu_2 - 6) - x_2(2\mu_1 + 2\mu_2 - 14) - x_3(\mu_1 + 4\mu_2 - 13) - \xi_1(-\mu_1) - \xi_2(-\mu_2)
\]

One can verify that we have the same “solution” of \((\mu^*, \mathbf{x}^*)\) to the original maximization problem such that \(\mu_1 = 11, \mu_2 = 1/2\), and \(v = 294\). These are exactly the desired values of the shadow prices, and the value of \(v\) reflects that the firm’s contribution is fully allocated to its resources. Essentially, the dual determines reservation price for the resources that would allow the firm to break even, in the sense that its total contribution would exactly equal the total rental value of its resources. However, the firm in fact owns its resources, and so the shadow prices are interpreted as the breakeven rates for acquiring the additional resource.

The result in the example also holds in general linear programming problems. Let’s write the conditions in matrix form.

\[
\begin{align*}
\max_{\mathbf{x} \in \mathbb{R}^n} & \quad \mathbf{c} \cdot \mathbf{x} \\
\text{subject to} & \quad \mathbf{A} \mathbf{x} \leq \mathbf{b} \\
& \quad \mathbf{x} \geq 0
\end{align*}
\]

If \(\mathbf{x}^*\) is the solution, there exists \(\lambda^* \geq 0\) and \(\eta^* \geq 0\) such that

\[
\mathbf{c} - \mathbf{A}^T \lambda^* + \eta^* = \mathbf{0}
\]
and, letting $A_i$ be the $i$th row of $A$, the complementary slackness conditions are

$$\lambda_i^*(A_i x^* - b_i) = 0 \quad \text{for all } i = 1, \ldots, m$$

$$\eta_i^*(-x_i^*) = 0 \quad \text{for all } i = 1, \ldots, n$$

The Lagrangian is

$$\mathcal{L}(\lambda, \eta, x) = c \cdot x - \lambda \cdot (A x - b) - \eta \cdot (-x)$$

The original maximization problem is called a Primal, and the associated minimization problem is called a Dual. The relationship is summarized in the following table.

<table>
<thead>
<tr>
<th>Primal</th>
<th>variables</th>
<th>constraints</th>
<th>objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual</td>
<td>$x_1 \geq 0, x_1 \geq 0, \ldots, x_1 \geq 0$</td>
<td>$\geq b_1, \geq b_2, \ldots, \geq b_m$</td>
<td>$\max c \cdot x$</td>
</tr>
<tr>
<td>Variables</td>
<td>$\lambda_1 \geq 0$</td>
<td>$a_{11}, a_{12}, \ldots, a_{1n}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_2 \geq 0$</td>
<td>$a_{21}, a_{22}, \ldots, a_{2n}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_m \geq 0$</td>
<td>$a_{m1}, a_{m2}, \ldots, a_{mn}$</td>
<td>$\geq b_m$</td>
</tr>
<tr>
<td>constraints</td>
<td>$\geq c_1, \geq c_2, \ldots, \geq c_n$</td>
<td>$\min b \cdot \mu$</td>
<td></td>
</tr>
<tr>
<td>objective</td>
<td>$\min b \cdot \mu$</td>
<td>$A^T \mu \geq c, \mu \geq 0$.</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 15-3 Primal and Dual Relationship**

In practice, the multipliers for the non-negativity condition of $x \geq 0$ is suppressed by assuming the domain is $\mathbb{R}^n_+$.  

**Theorem 15-5 (The Strong Duality Property of Linear Programming)**

For a Primal Linear Programming Problem

$$\max_x c \cdot x \quad \text{subject to } Ax \leq b, \quad x \geq 0,$$

there is an associated Dual Linear Programming Problem

$$\min_{\lambda} b \cdot \mu \quad \text{subject to } A^T \mu \geq c, \quad \mu \geq 0.$$
If the primal has a finite optimal solution, then so does the dual and \( b \cdot \mu^* = c \cdot x^* \). Moreover, the converse is also true.

**Proof.** Dual is constructed using the relationship summarized in Figure 15-3.

First part: Since the solution to the primal satisfies 
\[
\begin{align*}
\mathcal{L}(\lambda, \eta^*, x^*) &= c \cdot x^* - \lambda \cdot (Ax^* - b) + \eta^* \cdot x^* \\
&= (c - A^T \lambda^* + \eta^*) \cdot x^* + \lambda b = \lambda b
\end{align*}
\]
On the other hand, by the necessary condition of 
\[
c - A^T \lambda^* + \eta^* = 0,
\]
we have
\[
\mathcal{L}(\lambda^*, \eta^*, x^*) = c \cdot x - \lambda^* \cdot (Ax - b) + \eta^* \cdot x = (c - A^T \lambda^* + \eta^*) \cdot x + \lambda^* b = \lambda^* b
\]
Combining these two equations yields
\[
\lambda b \geq \mathcal{L}(\lambda, \eta^*, x^*) \geq \mathcal{L}(\lambda^*, \eta^*, x^*) = \lambda^* b
\]
That is, \( \lambda^* \) is the minimum point for all \( \lambda \geq 0 \).

Second part: Note that \( \mathcal{L}(\lambda^*, \eta^*, x^*) = c \cdot x^* \) because \( x^* \) is the solution to the primal. The result follows that \( \mathcal{L}(\lambda^*, \eta^*, x) = \lambda^* b \) for all \( x \) as shown above. The proof of the converse is identical replacing the primal with the dual (the minimization problem) and the dual with the primal (the dual of the dual).

The theorem shows that if \( x^* \) solves the primal, then \( \mu^* \) solves the dual, and that this problem has the same general form as the primal. Hence the dual can be written in the form of the Primal and (applying the transformation one more time) the Dual of the Dual is the Primal. This implies that dual may be solved instead of the primal if there are computational advantages. As a minor point, the property allows a simple way to prove if a solution is optimal.

The most important implication of the result is that duality makes it possible to interpret the “multipliers” economically. Mathematically, the result makes it clear that when you solve a constrained optimization problem you are simultaneously solving another optimization problem for a vector of multipliers.

It is worth to note the **unboundedness property.** Since linear functions are continuous and unbounded, the feasible set cannot be bounded. It turns out that you might be able to make the objective function of the Primal arbitrarily large. As a matter of fact, the primal fails to have a solution only if the feasible set is empty or the objective is unbounded. Exactly, for this
reason, there could be no solution to the profit maximization problem with a firm with linear technologies in a competitive market.

In Example 15-3, ignoring the non-negativity conditions, we have seen that the Lagrangian for the maximization and minimization problems are identical. Moreover, \((\lambda^*, x^*)\) is the maximum point of \(L(\lambda^*, x)\) and is the minimum point of \(L(\lambda, x^*)\). This makes the point \((\lambda^*, x^*)\) a saddle point of the function \(L(\lambda, \eta, x)\). This property is useful especially in nonlinear programming problem where the dual function is the Lagrangian of the primal.

**Theorem 15-6.** If \(x^*\) solves the primal, then \((\lambda^*, \eta^*)\) and \(x^*\) are a saddle point of \(L(\lambda, \eta, x)\). That is, \(L(\lambda, \eta, x^*) \geq L(\lambda^*, \eta^*, x^*) \geq L(\lambda^*, \eta^*, x)\).

**Proof.** By the necessary condition, \(c - A^T\lambda^* + \eta^* = 0\), we have

\[
L(\lambda^*, \eta^*, x) = c \cdot x - \lambda^* \cdot (Ax - b) + \eta^* \cdot x = (c - A^T\lambda^* + \eta^*) \cdot x + \lambda^* b = \lambda^* b
\]

With linear programming problems, \(L(\lambda^*, \eta^*, x)\) constant. One the other hand, note that \(L(\lambda, \eta, x^*) = c \cdot x^* - \lambda (Ax^* - b) + \eta \cdot x^*\). Since \(Ax^* \leq b\) and \(\lambda, \eta, x \geq 0\), we have \(-\lambda (Ax^* - b) \geq 0\) and \(\eta \cdot x^* \geq 0\). Therefore, \(L(\lambda, \eta, x^*) > c \cdot x^* = L(\lambda^*, \eta^*, x^*)\).

**15.3 Karush-Kuhn-Tucker Theorem**

We have shown an equivalence between saddle points and optimality in linear programming problems and that, in general, the saddle point property guarantees optimality. And the converse is also true under a proper constraint qualification condition. Therefore, if \(x^*\) solves \(\text{MPI}\), then there exists multipliers that consist of a saddle point.

The primal and dual problems in nonlinear programming problem have the form:

\[
\begin{align*}
\max_{x \in \mathbb{R}^n} f(x; a) \\
\text{Primal :} & \quad \text{subject to } G(x; a) \leq b \\
& \quad H(x; a) = c
\end{align*}
\]

\[
\begin{align*}
\max_{\mu, \lambda} \theta(\mu, \lambda)
\end{align*}
\]

\textbf{Dual :}
subject to $\mu \leq 0$

where $\theta(\mu, \lambda) = \min_{x \in X} L(\mu, \lambda, x)$, which is called a dual function. As a counterpart of the duality in linear programming problems, the result is known as Lagrangian duality.

In a nonlinear programming problem, the equality constraints are usually assumed affine. However, since the equality constraints have no effect on the shape of the Lagrangian over a feasible set, most of the proofs hold with nonlinear equality constraints. The proofs of this section consider MPI with no equality constraint.

**Theorem 15-4** shows that optimality of MPI implies Karush-Kuhn-Tucker conditions. The following two results establish the equivalence between optimality, Karush-Kuhn-Tucker conditions, and saddle point.

(i) A saddle point of the Lagrange (dual) function implies the optimality of MPI.

(ii) Karush-Kuhn-Tucker conditions for a maximization problem with a concave objective function and convex constraint set implies a saddle point. This result is known as Karush-Kuhn-Tucker theorem.

(iii) A saddle point implies the Karush-Kuhn-Tucker condition.

**Definition 15-4 (Saddle Point)** $(\mu^*, x^*)$ is a saddle point of the Lagrangian of MPI, $L(\mu, x) = f(x) - \mu \cdot (G(x) - b)$, if $\mu^* \geq 0$ and $L(\mu, x)$ is maximized with respect to $x$ and minimized with respect to $\mu$ at $(\mu^*, x^*)$. That is, for all $\mu \geq 0$, $L(\mu^*, x) \leq L(\mu, x^*) \leq L(\mu, x)$.

**Theorem 15-7.** If $(\lambda^*, x^*)$ is a saddle point of $L(\mu, x)$, then $x^*$ solves MPI.

**Proof.** By the definition of the saddle point, for all $\lambda \geq 0$,

$$f(x) - \mu^* \cdot (G(x) - b) \leq f(x^*) - \mu^* \cdot (G(x^*) - b) \leq f(x^*) - \mu \cdot (G(x^*) - b)$$

From the second inequality,

$$(\mu^* - \mu) \cdot (G(x^*) - b) \geq 0$$

Therefore, setting $\mu = 0$ yields

$$\mu^* \cdot (G(x^*) - b) \geq 0$$
On the other hand, since \( G(x^*) - b \leq 0 \) and \( \mu^* \geq 0 \), we should have
\[
\mu^* \cdot (G(x^*) - b) \leq 0
\]
The two inequalities imply
\[
\mu^* \cdot (G(x^*) - b) = 0
\]
This is the complementary slackness condition. On the other hand, substituting the condition in the first inequality in the definition of saddle point yields
\[
f(x) - \mu^* \cdot (G(x) - b) \leq f(x^*)
\]
By the constraints, since we are considering only \( x \) such that \( G(x) - b \leq 0 \),
\[
\mu^* \cdot (G(x) - b) \leq 0 \quad \text{or} \quad -\mu^* \cdot (G(x) - b) \geq 0
\]
Therefore, \( x^* \) solves the maximization problem.  

**Theorem 15-8 Optimality under Slater Condition Implies a Saddle Point.**

Suppose that \( f(x) \) is concave, \( g_i(x) \) is convex for all \( i \in K \), and \( h_j(x) \) is affine for all \( j \in M \). Moreover, there exists an interior point \( x_0 \in \) such that \( g_i(x_0) < 0 \) and \( h_j(x_0) = 0 \) for all \( i \) and \( j \). If \( x^* \) is the optimal point for \( \textbf{MPI} \), then there exists \( \lambda^* > 0 \) such that \( (\mu^*, x^*) \) is a saddle point.

By strong duality,
\[
f(x^*) = \theta(\mu^*, x^*), \quad \mu^* \geq 0
\]
By definition,
\[
f(x^*) = \theta(\mu^*, x^*) \leq \mathcal{L}(\mu^*, x) \quad \text{for all} \ x \in X
\]
By optimality condition,
\[
f(x^*) = f(x^*) + \mu^* \cdot (G(x^*) - b) = \mathcal{L}(\mu^*, x^*)
\]
Therefore, \( \mathcal{L}(\mu^*, x^*) \leq \mathcal{L}(\mu^*, x) \). On the other hand,
\[
\mathcal{L}(\mu, x^*) = f(x^*) + \mu \cdot (G(x^*) - b) \leq f(x^*) = \mathcal{L}(\mu^*, x^*)
\]

The first inequality holds because \(G(x^*) - b \leq 0\) and \(\mu \geq 0\).

**Theorem 15-9 (Karush-Kuhn-Tucker Theorem)** Suppose that \(f\) is quasi-concave and \(G\) is quasi-convex. If \(x^*\) is a solution to \text{MPI} and a constraint qualification holds, then there exists \(\mu^* \geq 0\) such that \((\mu^*, x^*)\) is a saddle point of \(\mathcal{L}(\mu, x)\).

The beauty of this result is that it transforms a constrained problem into an unconstrained one with Lagrangian. Finding \(x^*\) to maximize \(\mathcal{L}(\lambda, x)\) for \(\lambda^*\) fixed is relatively easy. Also, note that if that \(f(x)\) and \(g_i(x)\) are differentiable, then the first-order conditions for maximizing \(\mathcal{L}(\lambda, x)\) are the familiar ones:

\[
\nabla f(x^*) = \sum_{i \in M} \lambda_i^* \nabla g_i(x^*)
\]

The first order conditions that characterization the minimization problem are precisely the complementary slackness conditions. Since \(\mathcal{L}(\lambda, x)\) is concave in \(x\) when \(f\) is concave and \(g_i\) is convex, it will turn out that

\[
\nabla f(x^*) = \sum_{i \in M} \lambda_i^* \nabla g_i(x^*), \quad \lambda^* \cdot (G(x^*) - b) = 0, \quad G(x^*) \leq b
\]

are necessary and sufficient for optimization. This will be an immediate consequence of the (ultimate) discussion of the sufficient conditions for optimality.

**Proof.** We assumed concavity and we need to find “multipliers.” It looks like we need a separation theorem. The trick is to find the right set.

Define a set \(C\) such that

\[
C = \{(y, z) \in \mathbb{R}^m \times \mathbb{R} | -G(x) \geq -y \quad \text{and} \quad f(x) \geq z \quad \text{for all} \quad x \in X\}.
\]

I write the constraint of \(G(x) \leq y\) as \(-G(x) \geq -y\) to keep the signs of multipliers non-negative. \(C\) is convex since the upper contour sets of \(f(x)\) and \(-G(x)\) are quasi-concave. \((b, f(x^*)) \in C\) is a boundary point such that \(-y \leq -b\) and \(f(x^*) \leq z \leq f(x^*)\). If not, there exists \(\varepsilon > 0\) such that \(-G(x^*) \geq -b - \varepsilon > -b\) or \(G(x^*) \leq b + \varepsilon < b\), and \(z > f(x) \geq f(x^*) + \varepsilon\).
We have a contradiction that all the constraints are not binding, and $f(x^*)$ is not the maximum.

Therefore, there is a supporting hyperplane passing $(-b, f(x^*))$. There exists $(\bar{\mu}, \xi) \neq 0$ with $\mu \in \mathbb{R}^m$ and $\xi \in \mathbb{R}$ such that, for all $(-y, z) \in K$,

$$(-b, f(x^*)) \cdot (\bar{\mu}, \xi) \geq (-y, z) \cdot (\bar{\mu}, \xi) \quad \text{or} \quad \bar{\mu} \cdot (-b) + \xi f(x^*) \geq \bar{\mu} \cdot (-y) + \xi z.$$ 

Note that $\xi \leq 0$ is not possible. If $\xi < 0$, we have $(-b, z) \in C$ for all arbitrarily small $z < 0$ and the above inequality does not hold. Suppose $\xi = 0$. By the constraint qualification condition, since there exists an $\bar{x} \in \text{int}(X)$ such that $-G(\bar{x}) > -b$, if $\bar{\mu} \cdot (-b) \geq \bar{\mu} \cdot (-y)$ for all $(-y, z) \in C$, we cannot have $-\bar{\mu} \cdot G(\bar{x}) > -\bar{\mu} \cdot b$ if $\bar{\mu} \geq 0$ and $\bar{\mu} \neq 0$. (the sign of $\mu$ is determined shortly independent of $\xi$.) Therefore, without loss of generality, we can normalized them such that $\xi = 1$ and $\mu = \bar{\mu} / \xi$.

Let’s show $(\mu, 1) \geq 0$. If $\mu_j < 0$, for $z = f(x^*)$ and $-y_i = -b_i$, a sufficiently small $-y_j < 0$ for $j \neq i$ violates the inequality,

$$-\mu \cdot b + f(x^*) = -\sum_{i \neq j} \mu_i b_i + \mu_i (-b_j) + f(x^*) \geq -\mu \cdot y + z = -\sum_{i \neq j} \mu_i b_i + \mu_j (-y_j) + z.$$ 

Therefore, we have $\mu \geq 0$, $\mu \neq 0$ and, for all $(y, z) \in K$,

$$\mu \cdot (-b) + f(x^*) \geq \mu \cdot (-y) + z.$$ 

Moreover, since $\mu \geq 0$ and $-G(x^*) \geq -b$ (by feasibility), $-\mu \cdot (G(x^*) - b) \geq 0$.

$$-\mu \cdot (G(x^*) - b) \geq 0 \iff \mathcal{L}(\lambda, x^*) = f(x^*) - \lambda \cdot (G(x^*) - b) \geq f(x^*) \quad (12.1)$$

Now, suppose that $\mu > 0$ (this is the constraint qualification condition and will be shown at the end of this proof.) Since $(-G(x), f(x)) \in C$,

$$\mu^* \cdot (-b) + f(x^*) \geq \mu^* \cdot (-G(x)) + f(x) \quad \text{for all } x.$$ 

$$f(x^*) \geq f(x) - \mu^* \cdot (G(x) - b) = \mathcal{L}(\mu^*, x) \quad (12.2)$$

Similarly, since $(-G(x^*), f(x^*)) \in K$, we have $-\mu^* \cdot (G(x^*) - b) \leq 0$. Together with Eq.(12.1), this implies $-\mu^* \cdot (G(x^*) - b) = 0$ or
\[ \mathcal{L}(\mu^*, x^*) = f(x^*) = f(x^*) - \lambda^* \cdot (G(x^*) - b) \]

Therefore, Eq.(12.1) and Eq.(12.2) implies

\[ \mathcal{L}(\mu^*, x) \leq \mathcal{L}(\mu^*, x^*) \leq \mathcal{L}(\mu, x^*) \]

If \( f(x) \) is concave and \( G(x) \) and \( H(x) \) are convex, the proof becomes significantly simpler.

By properties of the functions,

\[
\begin{align*}
  f(x) &\leq f(x^*) + \nabla f(x^*)(x - x^*) \\
  g_i(x) &\geq g_i(x^*) + \nabla g_i(x^*)(x - x^*) \quad \text{for all } i \in K \\
  h_j(x) &\geq h_j(x^*) + \nabla h_j(x^*)(x - x^*) \quad \text{for all } j \in M
\end{align*}
\]

By multiplying \( \mu_i^* \) and \( \lambda_j^* \) to the inequalities and adding them, we have

\[
\begin{align*}
  \mathcal{L}(\lambda^*, x) &= f(x) - \sum_{i=1}^{k} \mu_i^* g_i(x) - \sum_{j=1}^{m} \lambda_j^* h_j(x) \\
  &\leq f(x^*) - \sum_{i=1}^{k} \mu_i^* g_i(x^*) - \sum_{j=1}^{m} \lambda_j^* h_j(x) \\
  &\quad + \left( \nabla f(x^*) - \sum_{i=1}^{k} \mu_i^* \nabla g_i(x) - \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x) \right) \cdot (x - x^*) \\
  &= f(x^*) - \sum_{i=1}^{k} \mu_i^* g_i(x^*) - \sum_{j=1}^{m} \lambda_j^* h_j(x) = \mathcal{L}(\lambda^*, x^*)
\end{align*}
\]

On the other hand,

\[
\begin{align*}
  \mathcal{L}(\lambda, x^*) &= f(x^*) - \sum_{i=1}^{k} \mu_i \nabla g_i(x) - \sum_{j=1}^{m} \lambda_j \nabla h_j(x) \\
  &\geq f(x^*) = f(x^*) - \sum_{i=1}^{k} \mu_i^* g_i(x^*) - \sum_{j=1}^{m} \lambda_j^* \nabla h_j(x^*) = \mathcal{L}(\mu^*, \lambda^*, x^*)
\end{align*}
\]

Combining two inequalities yields \( \mathcal{L}(\mu^*, x) \leq \mathcal{L}(\mu^*, x^*) \leq \mathcal{L}(\mu, x^*) \).

Saddle point implies Karush-Kuhn-Tucker conditions.
THEOREM 15-10. Suppose that \((\mu^*, x^*)\) with an interior point \(x^*\) and \(\mu^* > 0\) is a saddle point. Then \((\mu^*, x^*)\) satisfy Karush-Kuhn-Tucker conditions.

Proof. Since \(\mathcal{L}(\lambda^*, x^*) \geq \mathcal{L}(\lambda^*, x)\), \(x^*\) maximizes \(\mathcal{L}(\lambda^*, x)\). Since \(x^*\) is an interior point, the first order condition is

\[ D_x \mathcal{L}(\lambda^*, x^*) = 0 \]

or

\[ f(x^*) - \sum_{i=1}^{k} \mu_i^* \nabla g_i(x^*) = 0 \]

Together with the feasibility condition and complementary slackness condition, we have the Karush-Kuhn-Tucker condition.

15.4 SUFFICIENT CONDITIONS

With the second order conditions, we can distinguish local maxima from other critical points. If \(x^*\) is a maximum and \(v\) is a feasible vector, then \(x^* + tv\) is in a feasible set \(X\) for all \(t\) near zero. In unconstrained maximization problem, the second order condition requires that the slope of an objective function changes from positive to negative in the neighborhood of a critical point, or a concave objective function. With constraints, the requirements are essentially equivalent so that an objective function is concave for all feasible directions. As usual, an one variable argument shows that the second order condition is \(F''(0) \leq 0\) where \(F(t) = f(x^* + tv)\), or

\[ F''(0) = v^T D^2 f(x^*) v < 0. \]

Since the feasible vectors are nothing to do with those nonbinding constraints, the second order condition is exactly the same as those for MPE as stated in THEOREM 14-5.

The following theorem provides simple but quite strong conditions for the second order condition.
**Theorem 15.11 (Sufficient Conditions)** Suppose that $f(x)$ and $g(x)$ are continuously differentiable. And $f(x)$ is concave and the feasible set is convex. If there exist $x^*$ and $\mu^* \geq 0$ such that

$$\nabla f(x^*) = \sum_{i \in M} \mu_i^* \nabla g_i(x^*), \quad \mu^* \cdot (G(x^*) - b) = 0,$$

Then $x^*$ solves MPI.

**Proof.** To reach a contradiction, suppose that there is $x_0 \in X$ such that $f(x_0) > f(x^*)$, and let $v = x_0 - x^*$. Note that $x^* + tv$ is feasible because every convex combination of $x_0$ and $x^*$ is feasible.

$$Df(x^*) \cdot v = \lim_{t \to 0} \frac{f(x^* + tv) - f(x^*)}{t} \geq \lim_{t \to 0} \frac{(1-t)f(x^*) + tf(x_0) - f(x^*)}{t} = f(x_0) - f(x^*) > 0,$$

where the inequality follows from the concavity $f(x)$. Similarly, since $x_0$ is feasible,

$$Dg_i(x^*) \cdot v = \lim_{t \to 0} \frac{g_i(x^* + tv) - g_i(x^*)}{t} \leq 0.$$

Then we reach a contradiction,

$$0 < \nabla f(x^*) \cdot v = \sum_{i \in M} \mu_i^* \nabla g_i(x^*) \cdot v \leq 0$$

Note that the result does not involve any constraint qualification which is replaced by the convexity of a feasible set. With nonlinear equality constraint, the convexity is hardly satisfied.

### 15.5 Non-Negativity Condition

Most of the economic variables do not have negative values. Since non-negativity conditions appear in almost every optimization problem, the condition is usually suppressed with inequality in the first order conditions.

$$\max_{x \in \mathbb{R}^n} f(x)$$
subject to $G(x) \leq b$
$x \geq 0$

Let $\eta \geq 0$ be the multiplier for the constraint $-x \leq 0$. Kuhn-Tucker condition is

$$\nabla f(x^*) = \sum_{i \in M} \mu_i^* \nabla g_i(x^*) - \eta^*$$

with standard complementary slackness conditions $\mu^* \cdot [G(x^*; a) - b] = 0$ and $-\eta^* \cdot x^* = 0$.

Since $\mu^* \geq 0$, we often write the first order condition as

$$\nabla f(x^*) \leq \sum_{i \in M} \mu_i^* \nabla g_i(x^*).$$

If $\eta_i^* > 0$ and

$$\frac{\partial f(x^*)}{\partial x_i} < \sum_{i \in M} \mu_i^* \frac{\partial g_i(x^*)}{\partial x_i},$$

we must have $x_i^* = 0$. By the same token, if the condition has an inequality such as

$$\frac{\partial f(x^*)}{\partial x_i} > \sum_{i \in M} \lambda_i^* \frac{\partial g_i(x^*)}{\partial x_i},$$

there must be a constraint of $x_i \leq d$ and the optimal value is $x_i^* = d$.

15.6 COMPARATIVE STATICS

In MP, if the associated value function, $V(a) = f(x(a), a)$, has a unique solution and is differentiable with respect to $a$, then

$$\frac{\partial V(a)}{\partial a} = D_a f(x^*; a) - \lambda^* D_a G(x^*, a) - \mu^* D_a H(x^*, a)$$

According to THEOREM 15-4, the multipliers of non-binding constraints are 0. The proof of the equality constraint case can be directly applied.
16 DIFFERENCE EQUATION SYSTEMS

Let time $t = 0, 1, \cdots$ be discrete. A function $x: \mathbb{N} \to \mathbb{R}^n$ that depends on time via a sequence of vectors of $x_0, x_1, x_2, \cdots$.

$$x(t) = x_t = f(x_0, x_1, x_2, \cdots)$$

The dynamic system is important in economics to study the convergence and the stability of an economic system.

**DEFINITION 16-1.** A $k$th order discret system of difference equations is an expression of the form

$$x_t = f(x_{t-1}, \cdots, x_{t-k}, t), \quad t = 0, 1, 2, \cdots$$

Where $x_t = \mathbb{R}^n$ and $f: \mathbb{R}^{n \times k} \times [0, \infty) \to \mathbb{R}^n$. The system is

(i) autonomous if $f$ does not depend on $t$, $x_t = f(x_{t-1}, \cdots, x_{t-k})$.

(ii) linear if $f$ is linear, $x_t = b(t) + A_1 x_{t-1} + A_2 x_{t-2} + \cdots + A_k x_{t-k}$.

(iii) of first order if $k = 1$.

**EXAMPLE 16-1.** Consider a first order difference equation,

$$x_n = \frac{x_{n-1}}{2} + 3$$

The equation can be solved by substitution.

$$x_{n-1} = \frac{x_{n-2}}{2} + 3, \quad x_{n-2} = \frac{x_{n-3}}{2} + 3, \quad \cdots, \quad x_2 = \frac{x_1}{2} + 3, \quad x_1 = \frac{x_0}{2} + 3$$

$$x_1 = \frac{1}{2} x_0 + 3, \quad x_2 = \frac{1}{2} \left( \frac{1}{2} x_0 + 3 \right) + 3, \quad x_3 = \frac{1}{2} \left( \frac{1}{2} \left( \frac{1}{2} x_0 + 3 \right) + 3 \right) + 3, \quad \cdots$$

Therefore,
\[ x_n = \left( \frac{1}{2} \right)^n x_0 + 3 \left( 1 + \cdots + \left( \frac{1}{2} \right)^{n-1} \right) = \left( \frac{1}{2} \right)^n (x_0 - 6) + 6 \]

**DEFINITION 16-2.** A sequence of \( \{x_0, x_1, x_2, \ldots\} \) obtained from the recursion of a difference equations system with the initial value of \( x_0 \) is called a trajectory, orbit or path of the dynamic system from \( x_0 \).

Economic models often take the linear autonomous difference equation form \( x_t = b + A_1 x_{t-1} + A_2 x_{t-2} + \cdots + A_k x_{t-k} \), where \( x_t \) is a vector of economic variables. A linear autonomous system can be written in the form of a first order autonomous linear system.

**THEOREM 16-1.** The \( k \)th order system \( z_t = c + A_1 z_{t-1} + \cdots + A_k z_{t-k} \) may be rewritten as the first order system \( x_t = b + A x_{t-1} \), where

\[
A = \begin{bmatrix}
A_1 & \cdots & A_{k-1} & A_k \\
I_k & \cdots & 0 & 0 \\
0 & \cdots & I_n & 0
\end{bmatrix}, \quad x_t = \begin{bmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-k+1} \end{bmatrix}, \quad b = \begin{bmatrix} c_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}
\]

**Proof.** \( x_t = b + A x_{t-1} \) consists of \( k \) vector equations. The first is simply \( z_t = c + A_1 z_{t-1} + \cdots + A_k z_{t-k} \) itself. The second to \( k \)th are \( z_{t-1} = z_{t-1}, \ldots, z_{t-k+1} = z_{t-k+1} \).

**16.1 SOLUTION AND STABILITY**

An equilibrium of \( x_t = b + A x_{t-1} \) is a value of \( x \) denoted by \( x^* \), which reproduces itself in the sense that \( x^* = b + A x^* \), or \( (I - A)x^* = b \). Thus, nonsingularity of \( (I - A) \) is a necessary and sufficient condition for uniqueness of equilibrium; and, given that condition, the equilibrium is \( x^* = (I - A)^{-1}b \).

**THEOREM 16-2.** If \( (I - A) \) is nonsingular, the \( n \times 1 \) system \( x_t = b + A x_{t-1} \) has solution

\[ x_t = (I - A)^{-1}b + A^t(x_0 - (I - A)^{-1}b) \]

**Proof.** If \( (I - A) \) is nonsingular, \( x^* = b + A x^* \) has a solution such that \( x^* = (I - A)^{-1}b \). The substitution yields

\[ (I - A)^{-1}b = b + A(I - A)^{-1}b \]
Subtracting it from $x_t = b + Ax_{t-1}$ gives

$$x_t - (I - A)^{-1}b = A(x_{t-1} - (I - A)^{-1}b)$$

Since this holds for all $t$, by repeating backward substitutions, we have

$$x_t - (I - A)^{-1}b = A^2(x_{t-2} - (I - A)^{-1}b)$$

$$\vdots$$

$$= A^t[x_0 - (I - A)^{-1}b]$$

The solution consists of “equilibrium” and deviation from “equilibrium”, and an equilibrium exists if the deviation part vanishes eventually. That is, an equilibrium is stable if, for any $x_0$, $x_t \to (I - A)^{-1}b$ as $t \to \infty$. This condition holds if and only if $A^t \to 0$, and by Theorem 10-18, which is equivalent to the condition that the eigenvalue of $A$ have moduli less than one.

For the eigenvalues of $A$ to have moduli less than one, each of the following conditions, taken separately, is necessary, or sufficient, or both, as indicated.

**THEOREM 16-3 (Conditions for Convergence)**

(i) Necessary and sufficient when $n = 1$: $|a_{11}| < 1$

(ii) Necessary and sufficient when $n = 2$ and $A$ is real: $1 - \text{tr}(A) + \text{det}(A) > 0$, $1 + \text{tr}(A) + \text{det}(A) > 0$, and $1 - \text{det}(A) > 0$

(iii) Necessary: $|\text{tr}(A)| < n$

(iv) Necessary: $|\text{det}(A)| < 1$

(v) Sufficient: $f(A) < 1$ for any matrix norm $f$

**Proof.**

(i) is apparent.

(ii) follows from direct (and tedious) manipulation of the characteristic equation of $A$.

(iii) and (iv) follow from the facts that the trace and determinant are sum and product of the eigenvalues, respectively.

(v) Theorem 10-16
16.2 Trajectories

The path of $x_t$ depends critically on $A^t$. Consider the case when $A$ is fully diagonalizable. Then, for some $P$,

$$x_t = (I - A)^{-1}b + P^{-1}\text{diag}(\lambda_1^t, \ldots, \lambda_n^t)P[x_0 - (I - A)^{-1}b]$$

Where $\text{diag}(\lambda_1^t, \ldots, \lambda_n^t)$ is a diagonal matrix with $\lambda_i^t$ in position $(i, i)$. The only place where $t$ appears is as a power for the eigenvalues. Thus, a real positive eigenvalue contributes a monotonic component to the path. A real negative eigenvalue contributes to a saw tooth component. And a complex eigenvalue contributes a cyclic component.

Note that, in time, the eigenvalue or eigenvalues of largest modulus dominate the rest. When $A$ is not diagonalizable, the same message applies. The diagonal matrix $\text{diag}(\lambda_1^t, \ldots, \lambda_n^t)$ in the previous equation is then replaced by a triangular matrix with the same diagonals $(\lambda_1^t, \ldots, \lambda_n^t)$ and with above-diagonal which are polynomials in the eigenvalues.

The "most stable" value of $A$ in $x_t = b + Ax_{t-1}$ is perhaps $A = 0$, which has zero eigenvalues. By continuity of $\det(A - \lambda I) = 0$, the eigenvalues depart only slightly from zero when $A$ departs only slightly from $A = 0$. Intuitively, then $x_t = b + Ax_{t-1}$ is likely to be stable if $A$ is not too different from zero. A matrix norm is a scalar measure $f(A)$ of the deviation of $A$ from 0. There are many matrix norms. It turns out that, for any matrix norm $f$, $f(A) < 1$ is a sufficient stability condition. Since norms are often easier to compute than eigenvalues, the sufficient stability condition $f(A) < 1$ is often convenient. Martix norms also have other uses. For example, they are used as measures of approximation error in numerical analysis, and they are used as ingredients in proofs.

Assorted facts Related to THEOREM 16-2:

(i) $x_t = b + Ax_{t-1}$ can be rewritten $z_t = (X^{-1}AX)z_{t-1}$, where $z_t = X^{-1}x_t$.

(ii) $x_t = Ax_{t-1}$ has the solution $x_t = A^tx_0$.

(iii) $x_t = \|x_t\|^t (\cos(t\theta_i) + i \sin(t\theta_i))$, where $\theta_i$ is the angle corresponding to the point in the complex plane (real part vertical, imaginary part horizontal.)

(iv) If $|\lambda| < 1$, $\lambda^t \to 0$ as $t \to \infty$. 
(v) If $T$ is a triangular matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_n$, then $T^t$ is a triangular matrix with diagonal elements, $\lambda_1^t, \lambda_2^t, \ldots, \lambda_n^t$.

(vi) $(P^{-1}AP)^t = P^{-1}A^tP$

(vii) $(I - A)(I + A + A^2 + \cdots + A^t) = I - A^{t+1}$

16.3 NONLINEAR SYSTEM

**THEOREM 16-4.** Let $f(x)$ be a $n \times 1$ vector of differentiable functions of an $n \times 1$ vector $x$, and let $x^*$ satisfy $x^* = f(x^*)$. Consider the system $x_t = f(x_{t-1})$. If all eigenvalues of the Jacobian matrix $Df(x^*)$ have moduli less than one, then the system is locally stable. That is, there exists a $\delta > 0$ such that, if $|x_0 - x^*| < \delta$, then $x_t \to x^*$ as $t \to \infty$.

Let $x_t = b + Ax_{t-1}$ be a linear approximation of the system in the vicinity of $x^*$. A condition for a good approximation would be

$$a_{ij} = \left. \frac{\partial f_i(x)}{\partial x_j} \right|_{x=x^*}, \quad \text{and} \quad b = (I - A)x^*$$

16.4 STOCHASTIC SYSTEM

Many economic variables seem to behave randomly. Adding a disturbance vector $u_t$ to our deterministic system creates a stochastic system $x_t = b + Ax_{t-1} + u_t$ capable of modelling a random $x_t$. Although we cannot apply deterministic methods directly to $x_t$, we can apply deterministic methods to the moments of $x_t$. In particular, consider the first two moments, the mean and covariance matrix of $x_t$:

$$\mu_t = E(x_t) \quad \text{and} \quad \Sigma_t = E[(x_t - \mu_t)(x_t - \mu_t)^T]$$

Under appropriate assumptions on $\mu_t$, it turns out that each of these arrays obeys its own system of a deterministic linear difference equation.

**THEOREM 16-5.** Consider the $n$-equation real stochastic system
\[ \mathbf{x}_t = \mathbf{b} + A\mathbf{x}_{t-1} + \mathbf{u}_t \]

where \( \mathbf{u}_t \) is a zero mean disturbance term, uncorrelated over time, with constant covariance matrix \( U \).

(i) Let \( \mu_t \) denote the mean of \( x_t \); that is, \( \mu_t = E(\mathbf{x}_t) \). If \( I - A \) is nonsingular, then

\[ \mu_t = (I - A)^{-1}\mathbf{b} + A^t[\mu_0 - (I - A)^{-1}\mathbf{b}] \]

Further, \( \mu_t \) converges to \( x_t \to (I - A)^{-1}\mathbf{b} \) as \( t \to \infty \), for any \( \mu_0 \), if and only if the eigenvalues of \( A \) have moduli less than one.

(ii) Let \( \Sigma_t \) denote the covariance matrix of \( x_t \); that is, \( \Sigma_t = E[(\mathbf{x}_t - \mu_t)^T (\mathbf{x}_t - \mu_t)] \). If \( (I - A \otimes A) \) is nonsingular, then

\[ \text{vec}(\Sigma_t) = (I - A \otimes A)^{-1}\text{vec}(U) + (A^T \otimes A^T)[\text{vec}(\Sigma_0) - (I - A \otimes A)^{-1}\text{vec}(U)] \]

Further, \( \text{vec}(\Sigma_t) \) converges to \( (I - A \otimes A)^{-1}\text{vec}(U) \) as \( t \to \infty \), for any \( \Sigma_0 \), if and only if the eigenvalues of \( A \) have moduli less than one.

**Sketch of proof.** (i) Assume \( E(\mathbf{u}_t) = 0 \). Then, taking the expectation of both sides of \( \mathbf{x}_t = \mathbf{b} + A\mathbf{x}_{t-1} + \mathbf{u}_t \) yields

\[ \mu_t = \mathbf{b} + A\mu_{t-1} \]

This is a perfect parallel to the original deterministic system \( \mathbf{x}_t = \mathbf{b} + A\mathbf{x}_{t-1} + \mathbf{u}_t \). That is, the mean vector \( \mu_t \) has the same dynamics as \( \mathbf{x}_t \) did before we added the disturbance.

(ii) Assume that the \( \mathbf{u}_t \) have zero means, are uncorrelated over time, and have common covariance matrix \( U \). Taking the variance-covariance matrix of both sides of \( \mathbf{x}_t = \mathbf{b} + A\mathbf{x}_{t-1} + \mathbf{u}_t \) yields

\[ \Sigma_t = A\Sigma_{t-1}A^T \quad \text{or} \quad \text{vec}(\Sigma_t) = \text{vec}(U) + (A \otimes A)\text{vec}(\Sigma_{t-1}) \]
17 INTEGRATION

Integration is an operation:

(i) to find antiderivatives as an inverse to differentiation, which is essential to solving differential equations.

(ii) to assign a number to functions to find the area, volume, and others.

(iii) to calculate averages of functions (distribution function) in statistics.

17.1 INDEFINITE INTEGRAL

DEFINITION 17-1. A differentiable function $F$ on $(a,b)$ with $F'(x) = f(x)$ is called an antiderivative or primitive or indefinite integral of $f(x)$, and write

$$F(x) = \int f(x) \, dx$$

Since the derivative of a constant is zero, adding an arbitrary constant to indefinite integral gives another antiderivative. The constant is called constant of integration or integral constant.

THEOREM 17-1. If $F$ and $G$ are both antiderivatives of $f$, then $G(x) = F(x) + c$.

EXAMPLE 17-1. Since $Dx^3 = 3x^2$, 

$$\int x^2 \, dx = \frac{1}{3} x^3 + c$$

THEOREM 17-2 (Integral of Elementary Functions)

(i) $\int x^n \, dx = \frac{x^{n+1}}{n+1} + c, \quad n \neq -1$

(ii) $\int \frac{1}{x} \, dx = \ln|x| + c$

(iii) $\int e^x \, dx = e^x + c$

Math for Economists, Sogang University
(iv) \[ \int \lambda f(x)\,dx = \lambda \int f(x)\,dx \]

(v) \[ \int (f \pm g)(x)\,dx = \int f(x)\,dx \pm \int g(x)\,dx \]

Integration by substitution or change of variables is a counterpart of the chain rule. Consider \( h(x) = f(g(x)) \). Since \( h'(x) = f'(g(x)) \cdot g'(x) \) or \( dh(x) = f'(g(x))\,dg(x) \), we can treat the inner function of \( g(x) \) as a variable. The key idea is to reverse the process of chain rule by identifying an inner function. If one can find \( dg(x) \) correctly, the problem is simplified to recover \( h(u) \) from \( dh(u) = f'(u)\,du \).

**Example 17-2.**

\[ \int (x + 2)^4\,dx = \int u^4\,du = \int \frac{u^5}{5}\,du = \frac{(x + 2)^5}{5} + c \]

**Theorem 17-3 (Integration by parts)** If \( f \) and \( g \) are continuous with antiderivatives \( F \) and \( G \), then

\[ \int F(x)g(x) \, dx = F(x)G(x) - \int f(x)G(x) \, dx \]

**Proof.** Consider \( H(x) = F(x)G(x) \).

\[ H'(x) = f(x)G(x) + F(x)g(x) \]

\[ H(x) = F(x)G(x) = \int [f(x)G(x) - F(x)g(x)]\,dx \]

Integration by part is an inverse operation of differentiating a product of functions. Like the change of variables, it is key to separate the argument function into two.

**Example 17-3.**

\[ \int xe^x\,dx \]

Let \( F(x) = x \) and \( g(x) = e^x \). Then \( f(x) = 1 \) and \( G(x) = e^x \), and

\[ \int xe^x\,dx = xe^x - \int e^x\,dx = (x - 1)e^x + c \]

If you choose \( F(x) = e^x \) and \( g(x) = x \), you have \( f(x) = e^x \) and \( G(x) = x^2/2 \).
17.2 DEFINITE INTEGRAL

To find the area under a curve from \( a \) to \( b \), approximate the curve by rectangles, from above and below to get upper and lower bounds. Then to get better approximations, subdivide the interval and repeat the process.

**Definition 17-2 (Riemann Integral)** Consider a real-valued function \( f: [a, b] \to \mathbb{R} \) and a sequence of a partition of \([a, b]\),

\[
a = x_0(r) \leq x_1(r) \leq \cdots \leq x_{n_r}(r) = b
\]

such that as \( r \to \infty \), \( n_r \to \infty \) and \( \Delta_i(r) = x_i - x_{i-1} \to 0 \) for all \( i = 1, \cdots, n_r \). This partitions the interval into sub-intervals \([x_{i-1}, x_i]\).

The lower bound of the area “under” the graph of \( f: \mathbb{R} \to \mathbb{R} \) is

\[
\lim_{r \to \infty} L(r), \quad \text{where} \quad L(r) = \sum_{i=1}^{n} h_i \Delta_i(r), \quad h_i = \inf_{x \in [x_{i-1}, x_i]} f(x).
\]

And the upper bound is

\[
\lim_{r \to \infty} U(r), \quad \text{where} \quad U(r) = \sum_{i=1}^{n} H_i \Delta_i(r), \quad H_i = \sup_{x \in [x_{i-1}, x_i]} f(x).
\]

If the two sums converge and are equal, we say that \( f \) is integrable on the interval \([a, b]\) and we denote the limit \( \int_a^b f(x) \, dx \) and call it (definite) integral of \( f(x) \).

For \( a < b \) and \( f(x) \geq 0 \), we interpret \( \int_a^b f(x) \, dx \) as area under the graph of \( f(x) \). In general, we interpret \( \frac{1}{b-a} \int_a^b f(x) \, dx \) as an average value of \( f(x) \).

**Theorem 17-4.** If \( f(x) \) is continuous over an closed interval, then \( f \) is integrable.

**Theorem 17-5 (The First Fundamental Theorem of Calculus)** If \( f(x) \) is continuous in the interval \([a, b]\), then for \( x \in [a, b] \) and \( F(x) \) such that

\[
F(x) = \int_a^x f(x) \, dx, \quad \text{for all} \ x \in [a, b].
\]
\( F(x) \) is differentiable and \( F'(x) = f(x), \quad \text{for all } x \in (a, b). \)

**Proof.** By the definition of the integral, the additional increase in “area” is
\[
\Delta \times \sup_{y \in [x, x+h]} f(y) \geq F(x + \Delta) - F(x) = \int_{x}^{x+\Delta} f(x) dx \geq \Delta \times \inf_{y \in [x, x+h]} f(y)
\]
the result follows from dividing through by \( \Delta \) and taking limits. \( \square \)

**Theorem 17-6 (The Second Fundamental theorem of calculus)** If \( F(x) \) is an antiderivative of \( f: [a, b] \to \mathbb{R} \), then
\[
\int_{a}^{b} f(x) \, dx = \int_{a}^{b} F'(x) \, dx = F(b) - F(a) = [F(x)]_{a}^{b}
\]

**Proof.** By Mean Value Theorem, for some \( c \in [a, b] \),
\[
F(y) - F(x) = f(c)(y - x).
\]
For each interval of the partition, since
\[
F(x_i) - F(x_{i-1}) = f(c_i)(x_i - x_{i-1}) \quad \text{and} \quad \inf_{x \in [x_{i-1}, x_i]} f(x) \leq f(c_i) \leq \sup_{x \in [x_{i-1}, x_i]} f(x),
\]
\[
L(r) \leq \sum_{i=1}^{n_r} [F(x_i) - F(x_{i-1})] \leq U(r)
\]
Since
\[
\lim_{r \to \infty} \sum_{i=1}^{n_r} [F(x_i) - F(x_{i-1})] = F(b) - F(a),
\]
the result follows by taking limits of \( r \). \( \square \)

Note that the second theorem does not require continuity of \( f(x) \).

If \( f \) is integrable, say on \([a, b]\), and \( x \in [a, b] \), a function can be defined as \( F(x) = \int_{a}^{x} f(t) \, dt \).
The first part shows that \( F'(x) = f(x) \). The second part of fundamental theorem of calculus is computationally useful.
THEOREM 17-7 (Properties of Definite Integral)

(i) \( \int_{a}^{b} c \ dx = c(b - a) \), where \( c \) is a constant.

(ii) \( \int_{a}^{b} f(x) \ dx \geq (b - a) \inf_{x \in [a,b]} f(x) \)

(iii) \( \int_{a}^{b} f(x) \ dx \leq (b - a) \sup_{x \in [a,b]} f(x) \)

(iv) \( \int_{a}^{b} [\lambda f(s) + \mu g(x)] \ dx = \lambda \int_{a}^{b} f(x)\ dx + \mu \int_{a}^{b} g(x) \ dx \), where \( \lambda, \mu \) are constants.

(v) If \( f \) and \( g \) are continuous on \([a, b]\) with antiderivatives \( F \) and \( G \), then

\[
\int_{a}^{b} f(x)G(x) \ dx = [F(x)G(x)]_{a}^{b} - \int_{a}^{b} F(x)g(x) \ dx
\]

= \( F(b)G(b) - F(a)G(a) - \int_{a}^{b} f(x)G(x) \ dx \)

(vi) Suppose \( \rho(\cdot) \) is strictly monontone on \([a, b]\) and \( f \) is continuous on an open interval \( \rho([a,b]) \). Then

\[
\int_{\rho(a)}^{\rho(b)} f(x) \ dx = \int_{a}^{b} f(\rho(s))\rho'(s)ds
\]

Proof. Let \( x = \rho(s) \). Since \( \rho \) is invertible, we have \( s = \rho^{-1}(x) \) and \( dx = \rho'(s)ds \). When we apply the change of variables, the limits should be taken care of too.

(vii) If \( f: [a, b] \rightarrow \mathbb{R} \) is continuous, then there exists \( c \in (a, b) \) such that

\[
\int_{a}^{b} f(x) \ dx = f(c)(b - a)
\]

Some useful results

(i) \( \int_{1}^{\infty} \frac{1}{t} \ dt = \ln x \), \( D \ln x = \frac{1}{x} \)

(ii) For a function \( F: [a, b] \rightarrow [0,1] \) with \( F(a) = 0 \), \( F(b) = 1 \), and \( F'(x) = f(x) \geq 0 \),
\[
\int_a^b x f(x) \, dx = b - \int_a^b F(x) \, dx
\]

(iii) \[
\int_0^\infty x^n e^{-x} \, dx = n \int_0^\infty x^{n-1} e^{-x} \, dx = n! \quad \text{where} \quad \lim_{x \to \infty} x^n e^{-x} = 0
\]

This is a continuous version of Jensen’s inequality.

**THEOREM 17-8 (Jensen’s Inequality)** \( f \) is concave if and only if for any \( X \) such that \( \int_X dx = 1 \),

\[
f \left( \int_X x \, dx \right) \geq \int_X f(x) \, dx
\]

(\( \Rightarrow \)) Since \( f \) is concave, \( \frac{f(x) - f(x_0)}{x - x_0} \) is decreasing for any \( x_0 \). Then there exists \( z \) such that

\[
\sup_{x < x_0} \frac{f(x) - f(x_0)}{x - x_0} \leq z \leq \inf_{x > x_0} \frac{f(x) - f(x_0)}{x - x_0}.
\]

Or

\[
z(x - x_0) \leq f(x) - f(x_0)
\]

Setting \( x_0 = \int_X x \, dx \) gives

\[
z \left( x - \int_X x \, dx \right) \leq f(x) - f \left( \int_X x \, dx \right)
\]

Integrating both sides yields

\[
0 \leq \int_X f(x) \, dx - f \left( \int_X x \, dx \right)
\]

(\( \Leftarrow \)) Suppose \( f(x) \) is not concave. Then there exists a set \( Y \subset X \) for which \( f(x) \) is strictly convex, or \( \frac{f(x) - f(x_0)}{x - x_0} \) is increasing for any \( x_0 \in Y \). Using the similar argument in the proof of only if part, we can show \( \int_Y f(x) \, dx - f \left( \int_Y x \, dx \right) > 0 \), a contradiction.
A1 LOGICAL IMPLICATIONS

A1.1 LOGICAL IMPLICATIONS

Logical implication

Logical negation: \( \neg P \) means “\( P \) is false.”

Logical equality: \( P \Leftrightarrow Q \) or \( P \equiv Q \)

Logical conjunction (AND): \( P \land Q \) means “\( P \) is true and \( Q \) is true.”

Logical disjunction (OR): \( P \lor Q \) means “\( P \) is true or \( Q \) is true or both are true.”

Order of operation: \( \neg P \land Q \) means \((\neg P) \land Q\), \( \neg P \lor Q \) means \((\neg P) \lor Q\)

Logical implication: \( P \Rightarrow Q \) means “whenever \( P \) is satisfied, \( Q \) is also satisfied.

Logical equivalence: \( P \Leftrightarrow Q \Leftrightarrow \neg P \lor Q \Leftrightarrow \neg Q \lor \neg P \Rightarrow P \lor Q \Leftrightarrow \neg(P \Rightarrow \neg Q)\),

Contrapositive: \( \neg Q \Rightarrow \neg P \) is the contrapositive of \( P \Rightarrow Q \)

Note that a statement “\( P \Rightarrow Q \)” is false if and only if \( P \) is true and \( Q \) is false. Otherwise, the sentence is true. The interpretation of \( \Rightarrow \) does not capture the connotations of “implies.” Two statements do not have to be related.

Example. Consider a statement “If 4 is even, then there is no sun during the night.” Since “there is no sun during the night” is true, the sentence is true.

Example. “If it rains, I carry an umbrella” is true unless I get wet.

A1.2 CONDITIONAL RELATIONSHIP

There is a different form of implication. If two statements can be related by a proposition or a conditional statement, we can put some restrictions on the logical value of each statement based on the other’s. Unlike the logical implication, a set must be a subset of the other. That is, the truth value is evaluated assuming the implication is true such that we exclude the case that \( P \) is true and \( Q \) is false.
Suppose that the proposition \((P \Rightarrow Q)\) is true. Then that implies that

If \(P\) is true, then \(Q\) is true. If \(P\) is false, then \(Q\) could be either true or false.

(i) Proving \(P\) true is sufficient to show that \(Q\) is true. We say that \(P\) is \textbf{sufficient} condition for \(B\), or \(A\) is a \textbf{sufficient} condition of \(Q\).

(ii) On the other hand, if we know that \(Q\) is true, it has no implication on the logical value of \(P\). However, if \(Q\) is false, we know that \(P\) cannot be true. In order for \(P\) to be true, \(Q\) must be true. We say that \(Q\) is \textbf{necessary} for \(P\), or \(Q\) is a \textbf{necessary} condition of \(P\).

The Venn diagram below depicts the set of conditions under which each statement is true.

Let \(P\) be \((x\text{ is an integer})\) and \(Q\) be \((x\text{ is a real number})\). The following are equivalent.

If \((x\text{ is an integer})\), then \((x\text{ is a real number})\)

\((x\text{ is an integer}) \Rightarrow (x\text{ is a real number})\)

\((x\text{ is an integer})\) is sufficient for \((x\text{ is a real number})\)

\((x\text{ is an integer})\) is never true except when \((x\text{ is a real number})\) is true.

\((x\text{ is an integer})\) only if \((x\text{ is a real number})\)

\((x\text{ is a real number})\) is necessary for \((x\text{ is an integer})\)
A2 Method of Proof

Mathematicians often collect information and make observations about particular cases or phenomena in an attempt to form a theory (a model) that describes patterns or relationships among quantities and structures. This approach to the development of a theory uses inductive reasoning.

However, the characteristic thinking of the mathematician is deductive reasoning, in which one uses logic to develop and extend a theory by drawing conclusions based on statements accepted as true. Proofs are essential in mathematical reasoning because they demonstrate that the conclusions are true. Generally speaking, a mathematical explanation for a conclusion has no value if the explanation cannot be backed up by an acceptable proof. (Douglas Smith, Maurice Eggen, Richard Andre, 2014, A Transition to Advanced Mathematics)

To prove the statement $P \implies Q$, there are several approaches.

A2.1 Direct Proof (Proof by Deduction)

This approach proves the statement directly. The conclusion is established by logically combining the axioms, definitions, and earlier theorems.

Assume that $P$ is true, and use $P$ to show that $Q$ must be true.

A2.2 Proof by Contradiction

This approach is based on the fact that any proposition cannot be both true and false at the same time. Assuming the statement $P \implies Q$ is true, demonstrate $P \implies \neg Q$ to drive a contradiction.

Assume that $P$ and $\neg Q$ are true, and use $P$ and $\neg Q$ to show a contradiction
A2.3 Proof by (Mathematical) Induction

$P(n_0)$ is true for some basis case.

Show that $P(n + 1)$ is true whenever $P(n)$ is true for $n > n_0$.

By the principle of induction, $P(n)$ is true for all $n \geq n_0$.

A2.4 Proof by Contraposition

This is not a method of proof per se, and use the equivalence of a statement and its contrapositive.