Lecture Notes for Mathematical Methods for Economic Dynamics*

Takashi Kunimoto†
SMU School of Economics

First Version: August 2022
This Version: October 2024

Abstract. This manuscript provides the mathematical background for those who wish to acquire the literacy of how academic research is done in Economics. Topics it covers are logic, set theory, and topology in the Euclidean space, linear algebra, multivariate calculus, static optimization, integration, differential equations, calculus of variations, and control theory.

*I am thankful to the students and the teaching assistants at Hitotsubashi, McGill, and SMU for their comments, questions, and suggestions that significantly have improved the quality of this manuscript. Yet, I believe that it still contains many errors, for which I am solely responsible.

†School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, SINGAPORE, tkunimoto@smu.edu.sg
URL: https://sites.google.com/site/tkunimoto73/

Contents

Chapter 1

Preliminaries

1.1 Logic

Theorems provide a compact and precise format for presenting the assumptions and important conclusions of sometimes lengthy arguments, and so help you identify immediately the scope and limitations of the result presented. Theorems must be proved and a proof consists of establishing the validity of the statement in the theorem in a way that is consistent with the rules of logic. I will explain some of the language using the rules of logic.

1.1.1 Necessity and Sufficiency

Consider any two statements, p and q. When I say "q is necessary for p," I mean that q must be true for p to be true. For p to be true requires q to be true, so whenever p is true, we know that q must also be true. So I might have said, instead, that "q is true *if* p is true," or simply that "p *implies* q." I denote this statement by $p \Rightarrow q$).

Suppose we know that "$p \Rightarrow q$" is a true statement. What if q is not true? Because q is necessary for p, when q is not true, then p cannot be true, either. But doesn't this just say that "p not true" is necessary for "q not true"? Or that "not-q" implies "not-p." ($\neg q \Rightarrow \neg p$). This latter form of the original statement is called the *contrapositive* form. Contraposition of the arguments in the statement *reverses* the direction of implication for a true statement.

Let's consider a simple illustration of these ideas. Let q be the statement, "x is an integer less than 10." Let p be the statement that "x is an integer less than 8." Clearly, q is necessary for p ($p \Rightarrow q$). If I form the contrapositive of these two statements, the statement $\neg q$ becomes "x is not an integer less than 10," and $\neg p$ becomes "x is not an integer less than 8." Then, you must observe that $\neg q \Rightarrow \neg p$. However, $\neg p \Rightarrow \neg q$ is *false*. The value of x could well be 9. I must *reverse* the direction of implication to obtain a contrapositive statement that is also true.

The notion of necessity is distinct from that of *sufficiency*. When I say "q is sufficient for p," we mean that whenever q holds, p must hold. I can say, "q is true *only if* p is true," or that "p is implied by q" ($p \Leftarrow q$). Once again, whenever the statement "$p \Leftarrow q$" is true, the contrapositive statement, "$\neg p \Rightarrow \neg q$" is also true.

Two implications, "$p \Leftarrow q$" and "$p \Rightarrow q$," can both be true. When this is so, I say that "p is necessary and sufficient for q," or "p is true if and only if q is true," or "p iff q." When "p is necessary and sufficient for q," I say that the statements p and q are *equivalent* and write "$p \Leftrightarrow q$."

To illustrate briefly, suppose that p and q are the following statements:

- $p \equiv$ "X is yellow,"
- $q \equiv$ "X is a lemon."

Certainly, if X is a lemon, then X is yellow. Here, p is necessary for q. At the same time, just because X is yellow does not mean that it must be a lemon. It could be a banana. So p is *not* sufficient for q.

1.1.2 Theorems and Proofs

Mathematical theorems usually have the form of an implication or an equivalence, where one or more statements are alleged to be related in particular ways. Suppose I have the theorem "$p \Rightarrow q$." Here, p is the *assumption* and q is the *conclusion*. To prove a theorem is to establish the validity of its conclusion given the truth of its assumption, and several methods can be used to do that.

1. In a *constructive proof*, we assume that p is true, deduce various consequences of that, and use them to show that q must also hold. This is also sometimes called a *direct proof*, for obvious reason.

2. In a *contrapositive proof*, we assume that q does *not* hold, then show that p cannot hold. This approach takes advantage of the logical equivalence between the claims "$p \Rightarrow q$" and "$\neg q \Rightarrow \neg p$" noted earlier, and essentially involves a constructive proof of the contrapositive to the original statement.

3. In a *proof by contradiction*, the strategy is to assume that p is true, assume that q is *not* true, and attempt to derive a logical contradiction. This approach relies on the fact that $p \Rightarrow q$ or $\neg q$ is always true and if "$p \Rightarrow \neg q$" is false, then $p \Rightarrow q$ must be true.

4. In a proof by *mathematical induction*, I have a statement $H(k)$ which does depend upon a natural number k. What we want to show is that a statement $H(k)$ is true for each $k = 1, 2, \ldots$. First, we show that $H(1)$ is true. This step is usually easy to establish. Next, we show that $H(k) \Longrightarrow H(k+1)$, i.e.,

whenever $H(k)$ is true, then $H(k+1)$ is also true. These two steps allow us to claim that we complete the proof.

If I assert that p is *necessary and sufficient* for q, or that "$p \Leftrightarrow q$," I must give a proof in "both directions." That is, both "$p \Rightarrow q$" *and* "$q \Rightarrow p$" must be established before a complete proof of the assertion has been achieved.

It is important to keep in mind the old saying that goes, "Proof by example is no proof." Suppose the following two statements are given:

- $p \equiv$ "x is a student,"
- $q \equiv$ "x has red hair."

Assume further that I make the assertion "$p \Rightarrow q$." Then clearly finding one student with red hair and pointing him out to you is not going to convince you of anything. Examples are good for illustrating but typically not for proving.

Finally, a sort of converse to the old saying about examples and proofs should be noted. Whereas citing a hundred examples can never prove that a certain property *always* holds, citing one solitary *counterexample* can disprove that the property always holds. For instance, to disprove the assertion about the color of students' hair, you need simply point out one student with brown hair. A counterexample proves that the claim cannot *always* be true because you have found at least one case where it is not.

1.1.3 Notation

I introduce the few logical symbols which will be used. I first consider five symbols for the five most common sentential connectives. The negation of a formula p is written as $\neg p$. The conjunction of two formulas p and q is written as $p \wedge q$. The disjunction of p and q as $p \vee q$. The equivalence p if and only if q as $p \Leftrightarrow q$. The universal quantifier "for every x" as $(\forall x)$ (the upside down "A"), and the existential quantifier "the exists x" as $(\exists x)$ (the backward "E"). I also use the symbol $(\exists! x)$ for "there is exactly one x."

Thus, the sentence:

> For every x, there is a y such that $x < y$

is symbolized:

$$(\forall x)(\exists y)(x < y).$$

The sentence:

> For every ε, there is a δ such that for every y, if $|x - y| < \delta$ then $|f(x) - f(y)| < \varepsilon$.

is symbolized:

$$(\forall\varepsilon)(\exists\delta)(\forall y)(|x-y|<\delta \Rightarrow |f(x)-f(y)|<\varepsilon).$$

The sentence:

For every x, there is exactly one y such that $x+y=0$

is symbolized:

$$(\forall x)(\exists! y)(x+y=0).$$

A few remarks concerning quantifiers may also be helpful. The *scope* of a quantifier is the quantifier itself together with the smallest formula immediately following the quantifier. What the smallest formula is, is always indicated by parentheses. Thus, in the formula

$$(\exists x)(x<y)\vee y=0$$

the scope of the quantifier "$(\exists x)$" is the formula "$(\exists x)(x<y)$."

1.2 Set Theory

Among the many branches of modern mathematics, set theory occupies a unique place: with a few rare exceptions, the entities which are studied and analyzed in mathematics may be regarded as certain particular sets or classes of objects. This means that the various branches of mathematics may be formally defined within set theory.

A *set* is any collection of elements. Sets of objects will usually be denoted by capital letters, A, S, T for example, while their members by lower case, a, s, t for example (English or Greek). A set S is a *subset* of another set T if every element of S is also an element of T. I write $S \subseteq T$. If $S \subseteq T$, then $x \in S \Rightarrow x \in T$. The set S is a *proper subset* of T if $S \subseteq T$ and $S \neq T$; sometimes one writes $S \subsetneq T$ in this case. Two sets are equal sets if they each contain exactly the same elements. I write $S = T$ if $\forall x$, $x \in S \Rightarrow x \in T$ and $x \in T \Rightarrow x \in S$. The number of elements in a set S, its *cardinality*, is denoted $|S|$.

A set S is *empty* or is an *empty set* if it contains no elements at all. It is a subset of "every" set. For example, if $A = \{x|\ x^2 = 0,\ x > 1\}$, then A is empty. I denote the empty set by the symbol $\emptyset$. The *complement* of a set S in a universal set U is the set of all elements in U that are not in S and is denoted S^c. For any two sets S and T in a universal set U, I define the *set difference* denoted $S\backslash T$, as all elements in the set S that are not elements of T. Thus, I can think $S^c = U\backslash S$. Clearly, for a given set S, what S^c is crucially depends on what the universal set U is. See the example below.

Example 1.2.1 *Let $S = \{x|5 \leq x \leq 10\}$ If $U = \mathbb{R}$, I can define the complement of S as*

$$S^c = \{x \in \mathbb{R}|\ x < 5 \ \text{ or } \ x > 10\}.$$

On the other hand, if $U = \{x|\ x \geq 0\}$ *as the set of nonnegative real numbers, I can define the complement of* S *as*

$$S^c = \{x \in \mathbb{R}|\ 0 \leq x < 5 \ \ or \ \ x > 10\}\,.$$

The *symmetric difference* $S\triangle T = (S\backslash T) \cup (T\backslash S)$ is the set of all elements that belong to exactly one of the sets S and T. Note that if $S = T$, then $S\triangle T = \emptyset$.

For two sets S and T, I define the *union* of S and T as the set $S \cup T \equiv \{x|\ x \in S \text{ or } x \in T\}$. I define the *intersection* of S and T as the set $S \cap T \equiv \{x|\ x \in S \text{ and } x \in T\}$. Let $\Lambda \equiv \{1, 2, 3, \ldots\}$ be an *index set*. In stead of writing $\{S_1, S_2, S_3, \ldots\}$ as a collection of sets, I can write $\{S_\lambda\}_{\lambda\in\Lambda}$. I would denote the union of all sets in the collection by $\bigcup_{\lambda\in\Lambda} S_\lambda$, and the intersection of all sets in the collection as $\bigcap_{\lambda\in\Lambda} S_\lambda$.

The following are some important identities involving the operations defined above.

- $A \cup B = B \cup A,\ (A \cup B) \cup C = A \cup (B \cup C),\ A \cup \emptyset = A$
- $A \cap B = B \cap A,\ (A \cap B) \cap C = A \cap (B \cap C),\ A \cap \emptyset = \emptyset$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C),\ A \cap (B \cup C = (A \cap B) \cup (A \cap C)$ (Distribute laws)
- $A\backslash(B \cup C) = (A\backslash B) \cap (A\backslash C),\ A\backslash(B \cap C) = (A\backslash B) \cup (A\backslash C)$ (De Morgan's laws)
- $A\triangle B = B\triangle A,\ (A\triangle B)\triangle C = A\triangle(B\triangle C),\ A\triangle\emptyset = A$

The collection of all subsets of a set A is also a set, called the *power set* of A and denoted by $\mathcal{P}(A)$. Thus, $B \in \mathcal{P}(A) \Longleftrightarrow B \subseteq A$.

The previous argument reveals its stance that the order of the elements in a set specification does not matter. In particular, $\{a, b\} = \{b, a\}$. However, on many occasions, one is interested in distinguishing between the first and the second elements of a pair. One such example is the coordinates of a point in the $x - y$-plane. These coordinates are given as an *ordered pair* (a, b) of real numbers. The important property of ordered pairs is that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$. The product of two sets S and T is the set of "ordered pairs" in the form (s, t), where the first element in the pair is a member of S and the second is a member of T. The product of S and T is denoted

$$S \times T \equiv \{(s, t)|\ s \in S,\ t \in T\}.$$

The set of real numbers is denoted by the special symbol $\mathbb{R}$ and is defined as

$$\mathbb{R} \equiv \{x|\ -\infty < x < \infty\}.$$

Any n-tuple, or *vector*, is just an n dimensional ordered tuple $(x_1, \ldots, x_n)$ and can be thought of as a "point" in n-dimensional Euclidean space. This space is defined as the set product

$$\mathbb{R}^n \equiv \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}} \equiv \{(x_1, \ldots, x_n) \mid x_i \in \mathbb{R}, \;\; i = 1, \ldots, n\}.$$

Often, I want to restrict my attention to a subset of $\mathbb{R}^n$, called the "nonnegative orthant" and denoted $\mathbb{R}^n_+$, where

$$\mathbb{R}^n_+ \equiv \{(x_1, \ldots, x_n) \mid x_i \geq 0, \; i = 1, \ldots, n\} \subseteq \mathbb{R}^n.$$

Furthermore, I sometimes talk about the strictly "positive orthant" of $\mathbb{R}^n$

$$\mathbb{R}^n_{++} \equiv \{(x_1, \ldots, x_n) \mid x_i > 0, \; i = 1, \ldots, n\} \subseteq \mathbb{R}^n_+.$$

1.3 Relation

Any ordered pair (x, y) associates an element $x \in X$ to an element $y \in Y$. A *relation* between members of X and members of Y can be thought of as a subset of $X \times Y$. Any *collection* of ordered pairs is said to constitute a *binary relation* between the sets X and Y. Many familiar binary relations are contained in the product of one set with itself. For example, let X be the closed unit interval, $X = [0, 1]$. Then the binary relation $\geq$ (i.e., the greater-than-or-equal-to relation) consists of all ordered pairs of numbers in X where the first one in the pair is greater than or equal to the second one. When, as here, a binary relation is a subset of the product of one set X with itself, I say that it is a relation *on* the set X. A binary relation $\mathcal{R}$ on X is represented by the subset of $X \times X$, i.e., $\mathcal{R} \subseteq X \times X$. For a binary relation $\mathcal{R}$ on a set X, that is, $\mathcal{R} \subseteq X \times X$, it is customary to write $x\mathcal{R}y$ rather than $(x, y) \in \mathcal{R}$. A synonym for relation is *correspondence*. A correspondence $\mathcal{R}$ from X to Y is again defined to be a subset of $X \times Y$, but I think of it as associating to each $x \in X$ the subset $\phi(x) = \{y \in Y | \; (x, y) \in \mathcal{R}\}$ of Y.

There are many conditions placed on binary relations in various contexts, and I summarize a number of them here.

Definition 1.3.1 *A relation $\mathcal{R}$ on X is **reflexive** if $x\mathcal{R}x$ for all $x \in X$.*

For example, $\geq$ and $=$ on $\mathbb{R}$ are reflexive, while $>$ on $\mathbb{R}$ is not.

Definition 1.3.2 *A relation $\mathcal{R}$ on X is **irreflexive** if, for any $x \in X$, $\neg(x\mathcal{R}x)$.*

For example, $>$ on $\mathbb{R}$ is irreflexive but $\geq$ and $=$ on $\mathbb{R}$ are not.

Definition 1.3.3 *A relation $\mathcal{R}$ on X is **symmetric** if, for any $x, y \in X$, $x\mathcal{R}y$ implies $y\mathcal{R}x$.*

For example, $=$ on $\mathbb{R}$ is symmetric, while $\geq$ and $>$ on $\mathbb{R}$ are not.

Exercise 1.3.1 *Show that symmetry does not imply reflexivity.*

Definition 1.3.4 *A relation $\mathcal{R}$ on X is **asymmetric** if, for any $x, y \in X$, $x\mathcal{R}y \Rightarrow \neg(y\mathcal{R}x)$.*

For example, $>$ on $\mathbb{R}$ is asymmetric, while $\geq$ and $=$ on $\mathbb{R}$ are not.

Exercise 1.3.2 *Show that if a relation $\mathbb{R}$ on X is asymmetric, it is irreflexive.*

Definition 1.3.5 *A relation $\mathcal{R}$ on X is **anti-symmetric** if, for any $x, y \in X$, $x\mathcal{R}y$ and $y\mathcal{R}x$ implies $x = y$.*

Although $>$ on $\mathbb{R}$ is asymmetric but $\geq$ on $\mathbb{R}$ is not asymmetric, both are anti-symmetric.

Exercise 1.3.3 *Provide an example of a binary relation that is both anti-symmetric and reflexive. In addition, provide an example of a binary relation that is anti-symmetric but not reflexive.*

Definition 1.3.6 *A relation $\mathcal{R}$ on X is **transitive** if, for any three elements x, y, and $z \in X$, $x\mathcal{R}y$ and $y\mathcal{R}z$ implies $x\mathcal{R}z$.*

For instance, all $\geq, =, >$ on $\mathbb{R}$ are transitive.

Definition 1.3.7 *A relation $\mathcal{R}$ on X is **complete** if, for all elements x and y in X, $x\mathcal{R}y$ or $y\mathcal{R}x$.*

For example, $\geq$ on $\mathbb{R}$ is complete, while $>$ and $=$ are not. Note that $\mathcal{R}$ on X is reflexive if it is complete.

A relation $\mathcal{R}$ is said to be a *partial ordering* on X if it is reflexive, transitive, and anti-symmetric. If a partial ordering is complete, it is called a *linear ordering*. For instance, the relation $\geq$ on $\mathbb{R}$ is a linear ordering. A relation $\mathcal{R}$ is *preorder* if it is reflexive and transitive. An anti-symmetric preorder is a partial order. A *chain* in a partially ordered set is a subset on which the order is complete. That is, any two distinct elements of a chain are ranked by the partial order.

1.4 Functions

A *function* is a relation that associates each element of one set with a single, unique element of another set. I say that the function f is a *mapping, map*, or *transformation* from one set D to another set R and write $f : D \to R$. I call the set D the *domain* and the set R the *range* of the mapping. If y is the point in the range mapped into by the point x in the domain, I write $y = f(x)$.

The *image* of f is the set of points in the range into which some point in the domain is mapped, i.e.,

$$f(D) \equiv \{y \in R \mid y = f(x) \text{ for some } x \in D\} \subseteq R.$$

The *inverse image* of a set of points $S \subseteq f(D)$ is defined as

$$f^{-1}(S) \equiv \{x \in D \mid f(x) \in S\}.$$

The *graph* of the function f is the set of ordered pairs

$$\text{graph}(f) \equiv \{(x, y) \mid x \in D,\ y = f(x)\}.$$

If $f(x) = y$, one also writes $x \mapsto y$. The squaring function $s : \mathbb{R} \to \mathbb{R}$, for example, can then be written as $s : x \mapsto x^2$. Thus, $\mapsto$ indicates the effect of the function on an element of the domain. If $f : A \to B$ is a function and $S \subseteq A$, the *restriction* of f to S is the function $f|_S$ defined by $f|_S(x) = f(x)$ for every $x \in S$. I also say that f is an *extension* of $f|_A$.

There is nothing in the definition of functions that prohibit more than one element in the domain from being mapped into the same element in the range. If, however, every point in the range is assigned to "at most" a single point in the domain, the function is said to be *one-to-one* (sometimes called *injective*), that is, for all $x, x' \in D$, whenever $f(x) = f(x')$, then $x = x'$. If the image is equal to the range, i.e., if, for every $y \in R$, there is $x \in D$ such that $f(x) = y$, the function is said to be *onto* (sometimes called *surjective*). If a function is one-to-one and onto (sometimes called *bijective*), then an *inverse function* $f^{-1} : R \to D$ exists and f^{-1} is also one-to-one and onto. The *composition* of a function $f : A \to B$ and a function $g : B \to C$ is the function $g \circ f : A \to C$ given by $(g \circ f)(a) = g(f(a))$ for all $a \in A$.

Exercise 1.4.1 *Show that $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ is not a one-to-one mapping. Show also that the same $f(\cdot)$ is not onto either. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(x) = x^2$. In this case, show that $f(\cdot)$ is bijective.*

1.5 Natural Numbers, Integers, and Rationals

I denote the set of all *natural numbers* by $\mathbb{N}$, that is, $\mathbb{N} \equiv \{1, 2, 3, \ldots\}$. This system satisfies the following property:

[The Principle of Mathematical Induction] If S is a subset of $\mathbb{N}$ such that $1 \in S$, and $i + 1 \in S$ whenever $i \in S$, then $S = \mathbb{N}$.

This property is actually one of the main *axioms* that are commonly used to construct the natural numbers. It is frequently employed when giving a recursive definition, or proving infinitely many propositions by recursion. Suppose $P_1, P_2, \ldots$

are logical statements. If we can prove that P_1 is true, and then show that the validity of P_{i+1} would follow from the validity of P_i (i being arbitrarily fixed in $\mathbb{N}$), then we may invoke the Principle of Mathematical Induction to conclude that each proposition in the string $P_1, P_2, \ldots$ is true.

Adjoining to $\mathbb{N}$ an element to serve as the additive identity, namely, the *zero*, I obtain the set of all *nonnegative integers*, which is denoted as $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. In turn, adjoining to $\mathbb{Z}_+$ the set $\{-1, -2, \ldots\}$ of all *negative integers* (whose construction would mimic that of $\mathbb{N}$), I obtain the set $\mathbb{Z}$ of all *integers*. For instance, consider an equation like $2x = 1$, which makes sense in $\mathbb{Z}$. However, it cannot possibly be solved in $\mathbb{Z}$. Hence, I need to extend $\mathbb{Z}$ to the set $\mathbb{Q}$ of all *rational numbers*, which can be described as

$$\mathbb{Q} = \{m/n |\ m, n \in \mathbb{Z},\ n \neq 0\}\,.$$

1.6 Sets of Real Numbers

There are certainly two rational numbers p and q such that $p^2 > 2 > q^2$, but now I know that there is no $r \in \mathbb{Q}$ with $r^2 = 2$. It is as if there were a "hole" in the set of rational numbers. So, I wish to *complete* $\mathbb{Q}$ by filling up its holes with "new" numbers. And doing this leads us to the set $\mathbb{R}$ of real numbers. Note that any member of the set $\mathbb{R}\backslash\mathbb{Q}$ is said to be an *irrational number*.

A set S of real numbers is *bounded above* if there exists a real number b such that $b \geq x$ for all $x \in S$. This number b is called an *upper bound* for S. A set that is bounded above has many upper bounds. A least upper bound for the set S is a number b^* that is an upper bound for S with the property that $b^* \leq b$ for every upper bound b. The existence of a least upper bound is a basic and non-trivial property of the real number system. This is summarized as the following fact:

Fact 1.6.1 (Least Upper Bound Principle) *Any nonempty set of real numbers that is bounded above has a least upper bound.*

This principle is rather an axiom of real numbers. In what follows, we accept this principle without hesitation. A set S in $\mathbb{R}$ can have at most one least upper bound, because if b_1^* and b_2^* are both least upper bounds for S, then $b_1^* \leq b_2^*$ and $b_2^* \leq b_1^*$, which thus implies that $b_1^* = b_2^*$. The least upper bound b^* of S is often called the *supremum* of S. I write $b^* = \sup S$ or $b^* = \sup_{x \in S} x$.

Example 1.6.1 *The set $S = (0, 5)$, consisting of all x such that $0 < x < 5$, has many upper bounds, some of which are 100, 6.73, and 5. Clearly no number smaller than 5 can be an upper bound, so 5 is the least upper bound. Thus,* $\sup S = 5$.

A set S in $\mathbb{R}$ is bounded below if there exists a $a \in \mathbb{R}$ such that $x \geq a$ for all $x \in S$. The number a is a lower bound for S. A set S in $\mathbb{R}$ that is bounded below has

a greatest lower bound a^* if $a^* \leq x$ for all $x \in S$, and $a^* \geq a$ for every lower bound a. The number a^* is called the *infimum* of S and I write $a^* = \inf S$ or $a^* = \inf_{x \in S} x$. Thus, I summarize

- $\sup S =$ the least number greater than or equal to all numbers in S; and
- $\inf S =$ the greatest number less than or equal to all numbers in S.

Theorem 1.6.1 (Characterization of the Least Upper Bound) *Let S be a set of real numbers and b^* a real number. Then, $\sup S = b^*$ if and only if the following two conditions are satisfied:*

1. *$x \leq b^*$ for all $x \in S$.*
2. *For each $\varepsilon > 0$, there exists an $x \in S$ such that $x > b^* - \varepsilon$.*

Proof: ($\Longrightarrow$) Since b^* is an upper bound for S, by definition, property 1 holds, that is, $x \leq b^*$ for all $x \in S$. Suppose, on the other hand, that there is some $\varepsilon > 0$ such that $x \leq b^* - \varepsilon$ for all $x \in S$. Define $b^{**} = b^* - \varepsilon$. This implies that b^{**} is also an upper bound for S and $b^{**} < b^*$. This contradicts our hypothesis that b^* is a least upper bound for S. ($\Longleftarrow$) Property 1 says that b^* is an upper bound for S. Suppose, on the contrary, that b^* is not a least upper bound. That is, there is some other b such that $x \leq b < b^*$ for all $x \in S$. Define $\varepsilon = b^* - b$. Then, we obtain that $x \leq b^* - \varepsilon$ for all $x \in S$. This contradicts property 2. ■

Chapter 2

Topology in $\mathbb{R}^n$

2.1 Sequences on $\mathbb{R}$

For $x \in \mathbb{R}$, the *absolute value* of x is defined as:

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

$|\cdot|$ satisfies the following properties:

- $|x| \geq 0$; or , $|x| = 0 \Leftrightarrow x = 0$.
- $|x| = |-x| \geq x$.
- $|xy| = |x||y|$.
- $\Big||x| - |y|\Big| \leq |x \pm y| \leq |x| + |y|$.

I omit the proof for the above properties.

Exercise 2.1.1 *Show the above four properties of* $|\cdot|$.

A *sequence* is a function $k \mapsto x(k)$ whose domain is the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of all positive integers and whose range is the set of real numbers $\mathbb{R}$. I denote the *terms* $x(1), x(2), \ldots, x(k), \ldots$ of the sequence by using superscripts: $x^1, x^2, \ldots, x^k, \ldots$. I shall use the notation $\{x^k\}_{k=1}^{\infty}$, or simply $\{x^k\}$, to indicate an arbitrary sequence of real numbers. A sequence $\{x^k\}$ of real numbers is said to be:

1. **nondecreasing** if $x^k \leq x^{k+1}$ for $k = 1, 2, \ldots$
2. **strictly increasing** if $x^k < x^{k+1}$ for $k = 1, 2, \ldots$
3. **nonincreasing** if $x^k \geq x^{k+1}$ for $k = 1, 2, \ldots$
4. **strictly decreasing** if $x^k > x^{k+1}$ for $k = 1, 2, \ldots$

A sequence that is nondecreasing or nonincreasing is called *monotone*. A sequence $\{x^k\}$ is said to *converge* to a number x if x^k becomes arbitrarily close to x for all sufficient large k. I write $\lim_{k\to\infty} x^k = x$ or $x^k \to x$ as $k \to \infty$. The precise definition of convergence is as follows:

Definition 2.1.1 *The sequence $\{x^k\}$ **converges** to x if, for every $\varepsilon > 0$, there exists a natural number $N_\varepsilon \in \mathbb{N}$ such that $|x^k - x| < \varepsilon$ for all $k > N_\varepsilon$. The number x is called the **limit** of the sequence $\{x^k\}$. A **convergent** sequence is one that converges to some number.*

Note that the limit of a convergent sequence is unique. A sequence that does not converge to any real number is said to *diverge*. In some cases, I use the notation $\lim_{k\to\infty} x^k$ even if the sequence $\{x^k\}$ is divergent. For example, I say that $x^k \to \infty$ as $k \to \infty$. A sequence $\{x^k\}$ is *bounded* if there exists a number M such that $|x^k| \leq M$ for all $k \in \mathbb{N}$.

Exercise 2.1.2 *Show that every finite set in $\mathbb{R}$ is bounded. (Hint: Prove this by induction)*

It is easy to see that *every convergent sequence is bounded*: If $x^k \to x$, by the definition of convergence, only finitely many terms of the sequence can lie outside the interval $I = (x-1, x+1)$. The set I is bounded and the finite set of points from the sequence that are not in I is bounded, so $\{x^k\}$ must be bounded. On the other hand, is every bounded sequence convergent? The answer turns out to be "No." For example, the sequence $\{y^k\} = \{(-1)^k\}$ is bounded but not convergent. Do you see why?

Theorem 2.1.1 *Every bounded monotone sequence is convergent.*

Proof: Suppose, without loss of generality, that $\{x^k\}$ is nondecreasing and bounded. Let b^* be the least upper bound of the set $X = \{x^k | k = 1, 2, \ldots\}$, i.e., $b^* = \sup X$, and let $\varepsilon > 0$ be an arbitrary number. Theorem 2.1 already showed that there must be a term x^N of the sequence for which $x^N > b^* - \varepsilon$. Because the sequence is nondecreasing, $b^* - \varepsilon < x^N \leq x^k$ for all $k > N$. But we also know that $x^k \leq b^*$ for all k because b^* is an upper bound, so we have $b^* - \varepsilon < x^k \leq b^*$ for all $k > N$. Thus, for any $\varepsilon > 0$, there exists a number N such that $|x^k - b^*| < \varepsilon$ for all $k > N$. Hence, $\{x^k\}$ converges to b^*. ■

Theorem 2.1.2 *Suppose that the sequences $\{x^k\}$ and $\{y^k\}$ converge to x and y, respectively. Then,*

1. $\lim_{k\to\infty}(x^k \pm y^k) = x \pm y$
2. $\lim_{k\to\infty}(x^k \cdot y^k) = x \cdot y$
3. $\lim_{k\to\infty}(x^k / y^k) = x/y$, *assuming that $y^k \neq 0$ for all k and $y \neq 0$.*

Exercise 2.1.3 *Prove the above theorem.*

2.2 Subsequences

Let $\{x^k\}$ be a sequence. Consider a strictly increasing sequence of natural numbers

$$k_1 < k_2 < k_3 < \cdots$$

and form a new sequence $\{y^j\}_{j=1}^{\infty}$, where $y^j = x^{k_j}$ for $j = 1, 2, \ldots$. The sequence $\{y^j\}_j = \{x^{k_j}\}_j$ is called a *subsequence* of $\{x^k\}$.

Theorem 2.2.1 *Every subsequence of a convergent sequence is itself convergent, and has the same limit as the original sequence.*

Exercise 2.2.1 *Show the above theorem.*

Although not every bounded sequence is convergent, I obtain the following result.

Theorem 2.2.2 *If the sequence $\{x^k\}$ is bounded, then it contains a convergent subsequence.*

Proof: Since $\{x^k\}$ is bounded, we can assume that there exists some $M \in \mathbb{R}$ such that $|x^k| \leq M$ for all $k \in \mathbb{N}$. For each $n \in \mathbb{N}$, consider a subset of $\{x^k\}$ and denote it by $\{x^k | k \geq n\}$. If there is no maximum element in $\{x^k | k \geq 1\}$, one can construct the following subsequence of $\{x^k\}$: Let $x^{k_1} = x^1$, let x^{k_2} be the first term in the sequence $\{x^2, x^3, \ldots\}$ greater than x_1, let x^{k_3} be the first term in the sequence $\{x^{k_2}, x^{k_2+1}, \ldots\}$ greater than x^{k_2}, and so on. Then, by construction, this subsequence $\{x^{k_j}\}_j$ is nondecreasing. Theorem 2.1.1 allows us to conclude that the subsequence$\{x^{k_j}\}$ is convergent. By the same logic, if, for any $n \in \mathbb{N}$, there is no maximum element in $\{x^k | k \geq n\}$, then we are done. Now, assume that the above case does not apply here. For each $n \in \mathbb{N}$, we can define $y^n = \max\{x^k | k \geq n\}$. By construction, $\{y^n\}$ is a nonincreasing sequence because the set $\{x^k | k \geq n\}$ shrinks as n increases. The sequence $\{y^n\}$ is also bounded because $-M \leq y^n \leq M$. Theorem 2.1.1 already showed that the sequence $\{y^n\}$ is convergent. We complete the proof. ■

Example 2.2.1 *Show that a sequence $\{(-1)^k\}$ is bounded and there exists a convergent subsequence.*

2.3 Point Set Topology in $\mathbb{R}^n$

Consider the n-dimensional Euclidean space $\mathbb{R}^n$, whose elements, or points, are n-vectors $x = (x_1, \ldots, x_n)$. The *Euclidean distance (or metric)* $d(x, y)$ between any two points $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ is the norm $\|x - y\|$ of the vector difference between x and y. Thus,

$$d(x, y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

If x, y, and z are points in $\mathbb{R}^n$, then

$$d(x,z) \leq d(x,y) + d(y,z) \quad (\textbf{triangle inequality})$$

I generalize the concept of the Euclidean space to metric spaces.

Definition 2.3.1 *Let X be a nonempty set. A function $d : X \times X \to \mathbb{R}_+$ that satisfies the following properties is called a **metric** on X: for any $x, y, z \in X$,*

1. *$d(x,y) = 0$ if and only if $x = y$.*
2. *(Symmetry): $d(x,y) = d(y,x)$.*
3. *(Triangle inequality): $d(x,z) \leq d(x,y) + d(y,z)$.*

So, the Euclidean distance is a particular metric on $\mathbb{R}^n$. If d is a metric on X, I say that (X,d) is a *metric space*, refer to the elements of X as *points* in (X,d). If d satisfies properties 2 and 3 and $d(x,x) = 0$ for any $x \in X$, I say that d is a *semimetric* and (X,d) is a *semimetric space*. Note that when the metric under consideration is apparent from the context, it is customary to dispense with the notation (X,d) and refer to X as a metric space.

If x^0 is a point in $\mathbb{R}^n$ and r is a positive real number, then the set of all points $x \in \mathbb{R}^n$ whose distance from x^0 is less than r, is called the *open ball* around x^0 with radius r. This open ball is denoted by $B_r(x^0)$. Thus,

$$B_r(x^0) = \{x \in \mathbb{R}^n | d(x^0, x) < r\}$$

Definition 2.3.2 *A set $S \subseteq \mathbb{R}^n$ is **open** if, for all $x \in S$, there exists some $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq S$.*

On the real line $\mathbb{R}$, the simplest type of open set is an open interval. Let S be any subset of $\mathbb{R}^n$. A point $x^0 \in S$ is called an *interior point* of S if there is some $\varepsilon > 0$ such that $B_\varepsilon(x^0) \subseteq S$. The set of all interior points of S is called the *interior* of S, and is denoted $\text{Int}(S)$ or S°. A set S in $\mathbb{R}^n$ is said to be a *neighborhood* of x^0 if x^0 is an interior point of S, that is, if S contains some open ball $B_\varepsilon(x^0)$ (i.e., $B_\varepsilon(x^0) \subseteq S$) for some $\varepsilon > 0$.

Theorem 2.3.1 *The following are the properties of open sets.*

1. *The entire space $\mathbb{R}^n$ and the empty set $\emptyset$ are both open.*
2. ***arbitrary** unions of open sets are open: Let Λ be an arbitrary index set. If U_λ is open for each $\lambda \in \Lambda$, then, $\bigcup_{\lambda\in\Lambda} U_\lambda$ is also open.*
3. *The intersection of **finitely** many open sets is open: Let Λ be a finite set. If U_λ is open for each $\lambda \in \Lambda$, then $\bigcap_{\lambda\in\Lambda} U_\lambda$ is open.*

Proof: (1) It is clear that $B_1(x) \subseteq \mathbb{R}^n$ for all $x \in \mathbb{R}^n$, so $\mathbb{R}^n$ is open. The empty set $\emptyset$ is open because the set has no element, so every member is an interior point.

(2) Let $\{U_\lambda\}_{\lambda\in\Lambda}$ be an arbitrary family of open sets in $\mathbb{R}^n$, and let $U^* = \bigcup_{\lambda\in\Lambda} U_\lambda$ be the union of the whole family. For each $x \in U^*$, there is at least one $\lambda \in \Lambda$ such that $x \in U_\lambda$. Since U_λ is open by our hypothesis, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset U_\lambda \subseteq U^*$. Hence, U^* is open.

(3) Let $\{U_k\}_{k=1}^K$ be *finite* collection of open sets in $\mathbb{R}^n$, and let $U_* = \bigcap_{k=1}^K U_k$ be the intersection of all these sets. Let x be any point in U_*. Since U_k is open for each k, there is $\varepsilon_k > 0$ such that $B_{\varepsilon_k}(x) \subseteq U_k$ for each k. Let $\varepsilon_* = \min\{\varepsilon_1, \ldots, \varepsilon_K\}$. This is well defined because of finiteness of K. Then, $B_{\varepsilon_*}(x) \subseteq U_k$ for each k, which implies that $B_{\varepsilon_*}(x) \subseteq U_*$. Hence, U_* is open. ■

Exercise 2.3.1 *There are two questions. First, draw the graph of* $S = \{(x, y) \in \mathbb{R}^2 | 2x - y < 2$ *and* $x - 3y < 5\}$. *Second, prove that* S *is open in* $\mathbb{R}^2$.

Definition 2.3.3 *A set* S *in* $\mathbb{R}^n$ *is* ***closed*** *if its complement,* $S^c = \mathbb{R}^n \backslash S$ *is open.*

A point $x^0 \in \mathbb{R}^n$ is said to be a *boundary point* of the set $S \subseteq \mathbb{R}^n$ if $B_\varepsilon(x^0) \cap S^c \neq \emptyset$ and $B_\varepsilon(x^0) \cap S \neq \emptyset$ for every $\varepsilon > 0$. In general, a set may include none, some, or all of its boundary points. An open set, for instance, contains none of its boundary points. The set of all boundary points of a set S is said to be the *boundary* of S and is denoted ∂S or $\text{bd}(S)$. A set $S \in \mathbb{R}^n$ is said to be *closed* if it contains all its boundary points. The union of S and its boundary $(S \cup \partial S)$ is called the *closure* of S, denoted by $\bar{S}$.

Exercise 2.3.2 *Show that a point* x *belongs to* $\bar{S}$ *if and only if* $B_\varepsilon(x) \cap S \neq \emptyset$ *for any* $\varepsilon > 0$.

The closure $\bar{S}$ of any set S is indeed closed. In fact, $\bar{S}$ is the smallest closed set containing S. Note that S is closed if and only if $S = \bar{S}$.

Example 2.3.1 *Let* $S = (0, 1)$ *be an open interval on the real line.* $\bar{S} = [0, 1]$ *becomes a closed interval.*

Theorem 2.3.2 *The following are the properties of closed sets.*

1. *The whole space* $\mathbb{R}^n$ *and the empty set* $\emptyset$ *are both closed.*
2. *Arbitrary intersections of closed sets are closed.*
3. *The union of finitely many closed sets is closed.*

Exercise 2.3.3 *Prove the above theorem.*

In topology, any set containing *some* of its boundary points but not all of them, is neither open nor closed. The half-open intervals $[a, b)$ and $(a, b]$, for examples, are neither open nor closed. Hence, openness and closedness are not mutually exclusive.

2.4 Topology and Convergence

I generalize the argument on sequences on the real line into $\mathbb{R}^n$. The basic idea is to apply the previous argument in a coordinate-wise. A *sequence* $\{x^k\}_{k=1}^{\infty}$ in $\mathbb{R}^n$ is a function that for each natural number k yields a corresponding point x^k in $\mathbb{R}^n$.

Definition 2.4.1 *A sequence $\{x^k\}$ in $\mathbb{R}^n$* ***converges*** *to a point $x \in \mathbb{R}^n$ if for each $\varepsilon > 0$, there exists a natural number $N \in \mathbb{N}$ such that $x_k \in B_\varepsilon(x)$ for all $k > N$, or equivalently, if $d(x_k, x) < \varepsilon$ for all $k > N$.*

Theorem 2.4.1 *Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$. Then, $\{x^k\}$ converges to the vector $x \in \mathbb{R}^n$ if and only if for each $j = 1, \ldots, n$, the real number sequence $\{x_j^k\}_{k=1}^{\infty}$, consisting of jth component of each vector x^k, converges to $x_j \in \mathbb{R}$, the jth component of x.*

Proof: ($\Longrightarrow$) For every k and every j, one has $d(x^k, x) = \|x^k - x\| \geq |x_j^k - x_j|$. It follows that if $x^k \to x$, then $x_j^k \to x_j$ for each $j = 1, \ldots, n$. ($\Longleftarrow$) Suppose that $x_j^k \to x_j$ as $k \to \infty$ for $j = 1, \ldots, n$. Then, given any $\varepsilon > 0$, for each $i = 1, \ldots, n$, there exists a number N_j such that $|x_j^k - x_j| < \varepsilon/\sqrt{n}$ for all $k > N_j$. It follows that

$$d(x^k, x) = \sqrt{|x_1^k - x_1|^2 + \cdots + |x_n^k - x_n|^2} < \sqrt{\varepsilon^2/n + \cdots \varepsilon^2/n} = \varepsilon,$$

for all $k > \max\{N_1, \ldots, N_n\}$. This is well defined because of the finiteness of n. Therefore, $x^k \to x$ as $k \to \infty$. ■

2.5 Properties of Sequences in $\mathbb{R}^n$

The notion of closedness (and hence openness) of a set in a metric space (in particular, $\mathbb{R}^n$) can be characterized by means of the sequences that live in that space.

Theorem 2.5.1 *The following two results are obtained.*

1. *For any set $S \subseteq \mathbb{R}^n$, a point $x \in \mathbb{R}^n$ belongs to $\bar{S}$ if and only if there exists a sequence $\{x^k\} \in S$ such that $x^k \to x$ as $k \to \infty$.*
2. *A set $S \subseteq \mathbb{R}^n$ is closed if and only if every convergent sequence of points in S has its limit in S.*

Proof: ($\Longrightarrow$ of Property 1) Let $x \in \bar{S}$. Regardless of whether $x \in \text{int}S$ or ∂S, for each $k \in \mathbb{N}$, we can construct x^k such that $x^k \in B_{1/k}(x) \cap S$ (In particular, take $x^k = x$ for each k). Then $x^k \to x$ as $k \to \infty$. ($\Longleftarrow$ of Property 1) Suppose that $\{x^k\}$ is a convergent sequence for which $x^k \in S$ for each k and $x = \lim_{k\to\infty} x^k$. We claim that $x \in \bar{S}$. For any $\varepsilon > 0$, there is a number $N \in \mathbb{N}$ such that $x^k \in B_\varepsilon(x)$ for all $k > N$. Since $x^k \in S$ for each k, it follows that $B_\varepsilon(x) \cap S \neq \emptyset$ for any $\varepsilon > 0$. Suppose, on the other hand, $x \notin \bar{S}$. Since $\bar{S}$ is closed, i.e., S^c is open, there is some

$\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq S^c$. This implies that $B_\varepsilon(x) \cap S = \emptyset$. This contradicts the conclusion we just drew that $B_\varepsilon(x) \cap S \neq \emptyset$ for any $\varepsilon > 0$. Hence, $x \in \bar{S}$.

($\Longrightarrow$ of Property 2) Assume that S is closed and let $\{x^k\}$ be a convergent sequence such that $x^k \in S$ for each k. Note that $x \in \bar{S}$ by property 1. Since $S = \bar{S}$ if S is closed, it follows that $x \in S$. ($\Longleftarrow$ of Property 2) By property 1, for any point $x \in \bar{S}$, there is some sequence $\{x^k\}$ for which $x^k \in S$ for each k and $\lim_{k\to\infty} x^k = x$. By our hypothesis, $x \in S$. This shows that $x \in \bar{S}$ implies $x \in S$, i.e., $\bar{S} \subseteq S$. By definition, $S \subseteq \bar{S}$ for any S. Hence $S = \bar{S}$, that is, S is closed. ■

Definition 2.5.1 *A set S in $\mathbb{R}^n$ is **bounded** if there exists a number $M \in \mathbb{R}_+$ such that $\|x\| \leq M$ for all $x \in S$. A set that is not bounded is called **unbounded**. Here $\|x\| = d(x, \mathbf{0}) = \sqrt{x_1^2 + \cdots + x_n^2}$, called the **Euclidean norm**.*

Similarly, a sequence $\{x^k\}$ in $\mathbb{R}^n$ is *bounded* if the set $\{x^k | k \in \mathbb{N}\}$ is bounded.

Lemma 2.5.1 *Any convergent sequence $\{x^k\}$ in $\mathbb{R}^n$ is bounded.*

Proof: If $x^k \to x$, then only finitely many terms of sequence can lie outside the ball $B_1(x)$. The ball $B_1(x)$ is bounded and any finite set of points is bounded, so $\{x^k\}$ must be bounded. ■

On the other hand, a bounded sequence $\{x^k\}$ in $\mathbb{R}^n$ is not necessarily convergent. This is the same as sequences in $\mathbb{R}$. The theorem below gives a characterization of boundedness of the set in terms of sequences.

Theorem 2.5.2 *A subset S of $\mathbb{R}^n$ is bounded if and only if every sequence of points in S has a convergent subsequence.*

Proof: ($\Rightarrow$) We complete this part by simply applying Theorem 2.2.2 to every coordinate and thereafter appealing to Theorem 2.4.1. ($\Leftarrow$) We take the contrapositive statement: if S is not bounded, there exists a sequence in S that does not have convergent sequences. We shall prove this. Since S is not bounded, for each $k \in \mathbb{N}$, there exists $x^k \in S$ such that $\|x^k\| > k$. Collecting all such x^k, we construct a sequence $\{x^k\}$. By construction of $\{x^k\}$, we conclude that no subsequence of $\{x^k\}$ is convergent. This completes the proof. ■

The theorem below is a characterization of closed and bounded sets in $\mathbb{R}^n$ by means of sequences.

Theorem 2.5.3 (Bolzano-Weierstrass) *A subset S of $\mathbb{R}^n$ is closed and bounded if and only if every sequence of points in S has a subsequence that converges to a point in S.*

Proof of Bolzano-Weierstrass's theorem: ($\Longrightarrow$) Let $\{x^k\}$ be a sequence such that $x^k \in S$ for each k. Since S is bounded, there is a convergent subsequence

$\{y^n\}_{n=1}^{\infty} = \{x^{k_n}\}_{n=1}^{\infty}$. Furthermore, $\lim_{n\to\infty} y^n = y \in S$ because S is closed. ($\Longleftarrow$) Let $\{x^k\}$ be any sequence for which $x^k \in S$ for each k. By our hypothesis, there exists a subsequence $\{y^n\} = \{x^{k_n}\}$ of the sequence $\{x^k\}$ with $\lim_{n\to\infty} y^n = y \in S$. By the previous proposition, we conclude that S is bounded. To show the closedness of S, let $\{x^k\}$ be any convergent sequence for which $x^k \in S$ for each k and $x = \lim_{k\to\infty} x^k$. Since $\bar{S}$ is closed by definition, it follows that $x \in \bar{S}$. By our hypothesis, $\{x^k\}$ has a subsequence $\{x^{k_j}\}$ that converges to $\lim_{j\to\infty} x^{k_j} = x^{'} \in S$. But $\{x^{k_j}\}_j$ also converges to x. Hence, the limit points must be the same, that is, $x = x^{'} \in S$. ■

I now introduce the notion of compactness in the Euclidean space.

Definition 2.5.2 (Heine-Borel Theorem) *A set S in $\mathbb{R}^n$ is **compact** if it is closed and bounded.*

Exercise 2.5.1 *Let the number of commodities of the competitive market be n. Let $p_i > 0$ be a price for commodity i for each $i = 1, \ldots, n$. Let $w > 0$ be the consumer's wealth. Define the consumer's budget set $B(p, w)$ as*

$$B(p,w) \equiv \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+ \middle| \sum_{i=1}^{n} p_i x_i \leq w \right\}.$$

Show that $B(p, w)$ is nonempty and compact.

2.6 Continuous Functions

Consider first a real-valued function $y = f(x) = f(x_1, \ldots, x_n)$ of n variables. Roughly speaking, $f(\cdot)$ is continuous if small changes in the independent variables cause only small changes in the function value.

Definition 2.6.1 *A function $f : S \to \mathbb{R}$ with domain $S \subseteq \mathbb{R}^n$ is **continuous** at a point $x^0 \in S$ if, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for any $x \in S$,*

$$d(x, x^0) < \delta \Rightarrow |f(x) - f(x^0)| < \varepsilon.$$

If $f(\cdot)$ is continuous at every point in a set S, we simply say that $f(\cdot)$ is continuous on S.

Exercise 2.6.1 *Let $f(x) = \sqrt{x}$ be a function $f : \mathbb{R}_+ \to \mathbb{R}$. Prove that $f(\cdot)$ is continuous.*

Exercise 2.6.2 *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be given below.*

$$f(x) = \begin{cases} 1 & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \end{cases}$$

*Show that $f(\cdot)$ is **not** a continuous function.*

Consider next the general case of vector-valued functions.

Definition 2.6.2 *A function $f = (f_1, \ldots, f_m)$ from a subset S of $\mathbb{R}^n$ to $\mathbb{R}^m$ is said to be **continuous** at x^0 in S if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $x \in S$,*

$$d(x, x^0) < \delta \Rightarrow d(f(x), f(x^0)) < \varepsilon.$$

Equivalently, $f : S \to \mathbb{R}^m$ is continuous at $x^0 \in S$ if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that $f(B_\delta(x^0) \cap S) \subseteq B_\varepsilon(f(x^0))$.

The next theorem shows that the continuity of vector-valued functions can reduce to the continuity of each component (coordinate) functions, and vice versa.

Theorem 2.6.1 *A function $f = (f_1, \ldots, f_m)$ from $S \subseteq \mathbb{R}^n$ to $\mathbb{R}^m$ is continuous at a point x^0 in S if and only if each component function $f_j : S \to \mathbb{R},\ j = 1, \ldots, m$, is continuous at x^0.*

Proof: ($\Longrightarrow$) Suppose $f(\cdot)$ is continuous at x^0. Then, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|f_j(x) - f_j(x^0)| \leq d(f(x), f(x^0)) < \varepsilon$$

for every $x \in S$ with $d(x, x^0) < \delta$. Hence, f_j is continuous at x^0 for $j = 1, \ldots, m$. ($\Longleftarrow$) Suppose that each component f_j is continuous at x^0. Then, for every $\varepsilon > 0$ and every $j = 1, \ldots, m$, there exists $\delta_j > 0$ such that $|f_j(x) - f_j(x^0)| < \varepsilon/\sqrt{m}$ for every point $x \in S$ with $d(x, x^0) < \delta_j$. Let $\delta = \min\{\delta_1, \ldots, \delta_m\}$. Then $x \in B_\delta(x^0) \cap S$ implies that

$$d(f(x), f(x^0)) = \sqrt{|f_1(x) - f_1(x^0)|^2 + \cdots + |f_m(x) - f_m(x^0)|^2} < \sqrt{\frac{\varepsilon^2}{m} + \cdots + \frac{\varepsilon^2}{m}} = \varepsilon.$$

This proves that $f(\cdot)$ is continuous at x_0. ■

The next result is a characterization of the continuity of functions in terms of sequences.

Theorem 2.6.2 *A function f from $S \subseteq \mathbb{R}^n$ into $\mathbb{R}^m$ is continuous at a point x^0 in S if and only if $f(x^k) \to f(x^0)$ for every sequence $\{x^k\}$ of points in S that converges to x^0.*

Proof: ($\Longrightarrow$) Suppose that $f(\cdot)$ is continuous at x^0 and let $\{x^k\}$ be a sequence for which $x^k \in S$ and $\lim_{k\to\infty} x^k = x^0$. Let $\varepsilon > 0$ be given. Since $x^k \to x^0$, for any $\delta > 0$, there exists a number $N \in \mathbb{N}$ such that $d(x^k, x^0) < \delta$ for all $k > N$. Therefore, because of the continuity of $f(\cdot)$, there exists $\delta > 0$ such that $d(f(x), f(x^0)) < \varepsilon$ whenever $x \in B_\delta(x^0) \cap S$. But then $x^k \in B_\delta(x^0) \cap S$ and so $d(f(x^k), f(x^0)) < \varepsilon$ for all $k > N$. This implies that $f(x_k) \to f(x^0)$. ($\Longleftarrow$) Take any $x \in S$ and $\varepsilon > 0$. What we want to show is that one can find $\delta > 0$ such that $f(B_\delta(x)) \subseteq B_\varepsilon(f(x))$.

Suppose, by way of contradiction, that there is no such δ. Then, we can find a sequence $\{y^k\} \in \mathbb{R}^m$ such that for each $k \in \mathbb{N}$,

$$y^k \in f(B_{1/k}(x)) \backslash B_\varepsilon(f(x)).$$

Note that this construction is guaranteed by the Axiom of Choice.[1] Clearly, we have that for each $k \in \mathbb{N}$, there exists $x^k \in B_{1/k}(x)$ such that $y^k = f(x^k)$. By construction of $\{x^k\}$, we must have $x^k \to x$ as $k \to \infty$. By our hypothesis, we also have $y^k \to f(x)$. This implies that there exists $N \in \mathbb{N}$ such that $y^k \in B_\varepsilon(f(x))$ for each $k > N$. This contradicts the choice of $\{y^k\}$. ■

The theorem below shows that continuous mappings preserve the compactness of a set.

Theorem 2.6.3 *Let $S \subseteq \mathbb{R}^n$ and let $f : S \to \mathbb{R}^m$ be continuous. Then $f(K) = \{f(x)|x \in K\}$ is compact for every compact subset K of S.*

Proof: Let $\{y^k\}$ be any sequence in $f(K)$. By definition, for each k, there is a point $x^k \in K$ such that $y_k = f(x^k)$. Because K is compact, by Theorem 2.5.3, the sequence $\{x^k\}$ has a subsequence $\{x^{k_j}\}$ with the property that $x^{k_j} \in K$ for each j and $\lim_{j\to\infty} x^{k_j} = x^0 \in K$. Because $f(\cdot)$ is continuous, by Theorem 2.6.2, $f(x^{k_j}) \to f(x^0)$ as $j \to \infty$, where $f(x^0) \in f(K)$ because $x^0 \in K$. But then $\{y^{k_j}\}$ is a subsequence of $\{y^k\}$ that converges to a point $f(x^0) \in f(K)$. So, we have proved that any sequence in $f(K)$ has a subsequence converging to a point of $f(K)$. ■

Suppose that $f(\cdot)$ is a continuous function from $\mathbb{R}^n$ to $\mathbb{R}^m$. If V is an open set in $\mathbb{R}^n$, the image $f(V) = \{f(x)|x \in V\}$ of V need not be open in $\mathbb{R}^m$. Nor need $f(C)$ be closed if C is closed. Nevertheless, the *inverse image* $f^{-1}(U) = \{x|f(x) \in U\}$ of an open set U under continuous function f is always open. Similarly, the inverse image of any closed set must be closed. I state the result without proof.

Theorem 2.6.4 *Let $f(\cdot)$ be any function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then, f is continuous if and only if either of the following equivalent conditions is satisfied.*

1. *$f^{-1}(U)$ is open for each open set U in $\mathbb{R}^m$.*
2. *$f^{-1}(F)$ is closed for each closed set F in $\mathbb{R}^m$.*

Theorem 2.6.5 *Let S be a compact set in $\mathbb{R}$ and let x^* be the greatest lower bound of S and x_* be the least upper bound of S. Then, $x_* \in S$ and $x^* \in S$.*

Proof: Let $S \subseteq \mathbb{R}$ be closed and bounded and let x^* be the least upper bound of S. Then, by definition of any lower bound, we have $x^* \geq x$ for all $x \in S$. If $x^* = x$ for some $x \in S$, we are done. Suppose, therefore, that x^* is strictly greater

[1] You simply ignore this comment if you do not know what the Axiom of Choice is.

than every point in S. If $x^* > x$ for all $x \in S$, then $x^* \notin S$, so $x^* \in \mathbb{R}\backslash S$. Since S is closed, $\mathbb{R}\backslash S$ is open. Then, by the definition of open sets, there exists some $\varepsilon > 0$ such that $B_\varepsilon(x^*) = (x^* - \varepsilon, x^* + \varepsilon) \subseteq \mathbb{R}\backslash S$. Since $x^* > x$ for all $x \in S$ and $B_\varepsilon(x^*) \subseteq \mathbb{R}\backslash S$, we claim that for any $\tilde{x} \in B_\varepsilon(x^*)$, we must have $\tilde{x} > x$ for all $x \in S$. In particular, $x^* - \varepsilon/2 \in B_\varepsilon(x^*)$ and $x^* - \varepsilon/2 > x$ for all $x \in S$ (Recall Theorem 1.6.1). But then this contradicts our hypothesis that x^* is the least upper bound of S. Thus, we must conclude that $x^* \in S$. The same argument should be constructed for the greatest lower bound of S, i.e., x_*. ■

Theorem 2.6.6 (Weierstrass Theorem) *Let $f : S \to \mathbb{R}$ be a continuous real-valued mapping where S is a nonempty compact subset of $\mathbb{R}^n$. Then there exist two vectors $x^*, x_* \in S$ such that for all $x \in S$,*

$$f(x_*) \leq f(x) \leq f(x^*).$$

Proof: It follows from Theorems 2.6.3 and 2.6.5. ■

Example 2.6.1 *The following examples illustrates why the compactness and the continuity are needed for Weiertrass Theorem to apply. Each example fails to attain a maximum (or minimum) on the given interval.*

1. *Let $S = [0, \infty)$ and $f(x) = x$. Then $f(\cdot)$ cannot attain a maximum because S is not bounded from above. But S is closed and $f(\cdot)$ is continuous.*

2. *Let $S = (0, 1)$ and $f(x) = x$. Then $f(\cdot)$ cannot attain a maximum or minimum because S is not closed. But S is bounded and $f(\cdot)$ is continuous.*

3. *Let $S = [0, 1]$. Define $f : S \to \mathbb{R}$ as follows:*

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

This $f(\cdot)$ fails to attain a maximum because $f(\cdot)$ is not continuous at $x = 0$. But S is compact.

Chapter 3

Linear Algebra

I assume that the reader is familiar with the elementary algebra of real numbers.

3.1 Basic Concepts of Matrix Algebra in $\mathbb{R}^n$

An $m \times n$ *matrix* is a rectangular array with m rows and n columns:

$$A = (a_{ij})_{m\times n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Here $a_{ij} \in \mathbb{R}$ denotes the elements in the ith row and the jth column.

If $A = (a_{ij})_{m\times n},\ B = (b_{ij})_{m\times n}$, and $\alpha \in \mathbb{R}$ is a scalar, I define

- $A + B = (a_{ij} + b_{ij})_{m\times n}$,
- $\alpha A = (\alpha a_{ij})_{m\times n}$,
- $A - B = A + (-1)B = (a_{ij} - b_{ij})_{m\times n}$.

The *zero matrix* $\mathbf{0}_{m\times n}$ is defined as a matrix where all entries are zero:

$$\mathbf{0}_{m\times n} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Let A be an $m \times n$ and B be a $p \times m$ matrix. I would like to define the multiplication of two matrices, such as BA. I define the product $C = BA$ as the $p \times n$ matrix

$C = (c_{ij})_{p\times n}$, whose element in the ith row and the jth column is the inner product of the ith row of A and the jth column of B. That is,

$$c_{ij} = \sum_{r=1}^{n} a_{ir}b_{rj} = \underbrace{a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ik}b_{kj} + \cdots + a_{in}b_{nj}}_{n \text{ terms}}$$

It is important to note that the product BA is well defined only if the number of columns in B is equal to the number of rows in A.

If A, B, and C are matrices whose dimensions are such that the given operations are well defined, then the basic properties of matrix of multiplication are:

- $(AB)C = A(BC)$ (**associative law**)
- $A(B + C) = AB + AC$ (**left distributive law**)
- $(A + B)C = AC + BC$ (**right distributive law**)

Exercise 3.1.1 *Show the above three properties when we consider* 2×2 *matrices.*

However, matrix multiplication is *not* commutative. In fact,

- $AB \neq BA$, except in special cases
- $AB = \mathbf{0}$ does not imply that A or B is $\mathbf{0}$
- $AB = AC$ and $A \neq \mathbf{0}$ do not imply that $B = C$

Exercise 3.1.2 *Confirm the above three points by example.*

By using matrix multiplication, one can write a general system of linear equations in a very concise way. Specifically, the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots &\cdots \cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

can be written as $Ax = b$ if I define

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \; b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

A matrix is *square* if it has an equal number of rows and columns.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

The elements $a_{11}, a_{22}, \ldots, a_{nn}$ form the *principal diagonal* of the matrix A. If A is a square matrix and n is a positive integer, we define the nth power of A in the obvious way:

$$A^n = \underbrace{AA\cdots A}_{n \text{ factors}}$$

For *diagonal matrices* it is particularly easy to compute powers:

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_m \end{pmatrix} \Longrightarrow D^n = \begin{pmatrix} d_1^n & 0 & \cdots & 0 \\ 0 & d_2^n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_m^n \end{pmatrix}$$

The *identity matrix* of order n, denoted by $\mathbf{I}_n$, is the $n \times n$ matrix having ones along the main diagonal and zeros elsewhere:

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \quad (\textbf{identity matrix})$$

If A is any $m \times n$ matrix, then $A\mathbf{I}_n = A = \mathbf{I}_m A$. In particular,

$$A\mathbf{I}_n = \mathbf{I}_n A = A \quad \text{for every } n \times n \text{ matrix } A$$

If $A = (a_{ij})_{m\times n}$ is any matrix, the *transpose* of A is defined as $A^T = (a_{ji})_{n\times m}$. The subscripts i and j are interchanged because every row of A becomes a column of A^T, and every column of A becomes a row of A^T. The following rules apply to matrix transposition:

1. $(A^T)^T = A$
2. $(A + B)^T = A^T + B^T$
3. $(\alpha A)^T = \alpha A^T$
4. $(AB)^T = B^T A^T$

Exercise 3.1.3 *Prove the above four properties when we consider* 2×2 *matrices.*

A square matrix is said to be *symmetric* if $A = A^T$.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a mapping (transformation). A mapping f is said to be *linear* if, for any $x, y \in \mathbb{R}^n$ and any $\alpha \in \mathbb{R}$, the following two conditions are satisfied: (1) $f(x+y) = f(x)+f(y)$ and (2) $f(\alpha x) = \alpha f(x)$. For any linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$, there exists a *unique* $m \times n$ *matrix* A such that $f(x) = Ax$ for all $x \in \mathbb{R}^n$.[1] With this notation, I can express $f(x) = Ax$ as follows:

$$f(x) = \begin{pmatrix} f^{(1)}(x) \\ \vdots \\ f^{(j)}(x) \\ \vdots \\ f^{(m)}(x) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

The next result shows that linear mappings are continuous.

Lemma 3.1.1 *Any linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous.*

Proof: We skip the proof. ■

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ be two linear mappings. Then, I can set $m \times n$ matrix $A = (a_{ij})_{m\times n}$ associated with f and $p \times m$ matrix $B = (b_{ij})_{p\times m}$ associated with g. Consider the composite mapping $(g \circ f)(x) = g(f(x))$. Then, by the definition of multiplication of matrices, I have that $g \circ f \equiv BA$.

3.2 Determinants

Let $n \times n$ matrix A be given. Consider any product of n elements which appear in different rows and different columns of the matrix A, i.e., a product containing *just one element from each row and each column.* Such a product can be written in the form

$$a_{\alpha_1 1} a_{\alpha_2 2} \cdots a_{\alpha_n n}.$$

For the first factor I can always choose the element appearing in the first column of the matrix A; then, if I denote by α_1 the number of the row in which the element appears, the indices will be $\alpha_2, 2$, where α_2 is the number of the row in which the element appears, and so on. Thus, the indices $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the numbers of the rows in which the factors of the product $a_{\alpha_1 1} a_{\alpha_2 2} \cdots a_{\alpha_n n}$ appear, when I agree to write the column indices in increasing order. Since, by hypothesis, the elements $a_{\alpha_1 1}, a_{\alpha_2 2}, \ldots a_{\alpha_n n}$ appear in *different* rows of the matrix A, one from each row, then the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ are all different and represent some permutation of the numbers $1, 2, \ldots, n$.

[1] This is a non-trivial statement. But I take this one-to-one correspondence between linear mapping and matrix representation as a fact with no proof provided.

By an *inversion* in the sequence $\alpha_1, \ldots, \alpha_n$, I mean an arrangement of two indices such that the larger index comes before the smaller index. The total number of inversions will be denoted by $N(\alpha_1, \ldots, \alpha_n)$. For example, in the permutation $2, 1, 4, 3$, there are two inversions (2 before 1, 4 before 3), so that

$$N(2, 1, 4, 3) = 2.$$

In the permutation $4, 3, 1, 2$, there are five inversions (4 before 3, 4 before 1, 4 before 2, 3 before 1, 3 before 2), so that

$$N(4, 3, 1, 2) = 5.$$

Let $\alpha : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be a bijective map, often called a *permutation* over $\{1, \ldots, n\}$. Let $\mathcal{A}$ be the set of all permutations. I define the *determinant* of a matrix.

Definition 3.2.1 *The **determinant** of an $n \times n$ matrix $A = (a_{ij})_{i,j=1,\ldots,n}$, denoted by $|A|$, is the algebraic sum of the $n!$ products of the form $a_{\alpha_1 1} a_{\alpha_2 2} \cdots a_{\alpha_n n}$, each preceded by the sign determined by the rule $(-1)^{N(\alpha_1,\ldots,\alpha_n)}$. That is,*

$$|A| = \sum_{\alpha \in \mathcal{A}} (-1)^{N(\alpha_1,\ldots,\alpha_n)} a_{\alpha_1 1} a_{\alpha_2 2} \cdots a_{\alpha_n n}.$$

Exercise 3.2.1 *Verify that the determinants $|A|$ of 2×2 and 3×3 matrices are defined by*

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}; \quad \textit{and}$$

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}.$$

Consider any column, the j-th say, of the determinant $|A|$. Let a_{ij} be any element of this column. Add up all the terms containing the element a_{ij} appearing in the right hand side of the equation below:

$$|A| = \sum_{(\alpha_1,\ldots,\alpha_n)} (-1)^{N(\alpha_1,\ldots,\alpha_n)} a_{\alpha 1} a_{\alpha_2 2} \cdots a_{\alpha_n n},$$

then factor out the element a_{ij}. The quantity which remains, denoted by A_{ij}, is called the *cofactor of the element a_{ij}* of the determinant $|A|$.

Since every term of the determinant $|A|$ contains an element from the j-th column, the above formula for $|A|$ can be written in the form:

$$|A| = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj}.$$

This is called the *expansion of the determinant* $|A|$ *with respect to the (elements of the)* j*-th column*. Naturally, I can write a similar formula for any *row* of the determinant $|A|$. For example, for the i-th row, I have the formula:

$$|A| = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots a_{ij}A_{ij} + \cdots + a_{in}A_{in}.$$

The *cofactor* A_{ij} is the determinant of $(n-1) \times (n-1)$ matrices given by deleting ith row and jth columns from the matrix A:

$$A_{ij} = (-1)^{i+j} \begin{vmatrix} a_{11} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2,j-1} & a_{2,j+1} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{nn} \end{vmatrix}$$

I now indicate the role of determinants in solving systems of linear equations, by considering the example of a system of two equations in two unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \end{aligned}$$

Eliminating one of the unknowns in the usual way, one can easily obtain the formulas

$$x_1 = \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{21}a_{12}} \text{ and } x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{21}a_{12}}$$

assuming that $a_{11}a_{22} - a_{21}a_{12} \neq 0$. The numerators and denominators of the ratio can be represented by:

$$\begin{aligned} a_{11}a_{22} - a_{12}a_{21} &= \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \\ b_1a_{22} - b_2a_{12} &= \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}, \\ a_{11}b_2 - a_{21}b_1 &= \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}. \end{aligned}$$

Exercise 3.2.2 *Calculate the determinant of the following* 5×5 *matrix* A*:*

$$A = \begin{pmatrix} -2 & 5 & 0 & -1 & 3 \\ 1 & 0 & 3 & 7 & -2 \\ 3 & -1 & 0 & 5 & -5 \\ 2 & 6 & -4 & 1 & 2 \\ 0 & -3 & -1 & 2 & 3 \end{pmatrix}.$$

A square matrix A is said to be *nonsingular* if $|A| \neq 0$ and *singular* if $|A| = 0$.

3.2.1 Some Properties of Determinants

Proposition 3.2.1 *Let A and B be $n \times n$ matrices. Then, $|AB| = |A||B|$*

Exercise 3.2.3 *Prove the above proposition. If it is difficult, please focus on the case when $n = 2$.*

The next lemma shows that the transpose of a determinant has the same value as the original determinant.

Lemma 3.2.1 *Let A be a square matrix. Then, $|A^T| = |A|$.*

Exercise 3.2.4 *Show the above lemma.*

Theorem 3.2.1 (The linear property of determinants) *If all the elements of the j-th column of an $n \times n$ matrix A are linear combinations of two columns, then $|A|$ is equal to a linear combination of two determinants.*

Lemma 3.2.2 *If a column of an $n \times n$ matrix A consists of entirely zeros, then $|A| = 0$.*

Lemma 3.2.3 *Let A be an $n \times n$ matrix. If A contains two identical columns, then $|A| = 0$.*

Proof: Interchanging the columns does not change the determinant $|A|$. On the other hand, the determinant must change its sign. Thus, $|A| = -|A|$, which implies $|A| = 0$. ■

Using the above lemma, I also establish the following important property for the cofactor of a matrix.

Lemma 3.2.4 *Let A be an $n \times n$ matrix. Then, for each $j, k = 1, \ldots, n$ with $j \neq k$,*

$$a_{1k}A_{1j} + a_{2k}A_{2j} + \cdots + a_{nk}A_{nj} = 0,$$

and for each $i, \ell = 1, \ldots, n$ with $i \neq \ell$,

$$a_{\ell 1}A_{i1} + a_{\ell 2}A_{i2} + \cdots + a_{\ell n}A_{in} = 0.$$

Proof: Recall that for each $j = 1, \ldots, n$, we have

$$\sum_{\alpha_1, \ldots, \alpha_n} (-1)^{N(\alpha_1, \ldots, \alpha_n)} a_{\alpha_1 1} a_{\alpha_2 2} \cdots a_{\alpha_n n} = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj}. \quad (*)$$

Equation $(*)$ is an identity in the quantities $a_{1j}, a_{2j}, \ldots, a_{nj}$. Therefore, it remains valid if we replace a_{ij} $(i = 1, \ldots, n)$ by any other quantities. The quantities $A_{1j}, \ldots, A_{nj}$ remain unchanged when such a replacement is made, since they do not depend on the elements a_{ij}. Suppose that in the right and left hand sides of the

equality $(*)$, we replace the elements $a_{1j}, a_{2j}, \ldots, a_{nj}$ by the corresponding elements of any other column, say, $a_{1k}, a_{2k}, \ldots, a_{nk}$. Then, the determinant in the left-hand side of $(*)$ will have two identical columns and will be zero, according to the previous lemma. This completes the proof. ■

3.3 Cramer's Rule

A linear system of n equations and n unknowns is given:

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &=& b_2 \quad (*) \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &=& b_n \end{array}$$

Theorem 3.3.1 (Cramer's Rule) $(*)$ *has a unique solution if and only if* $|A| \neq 0$. *The solution is then*

$$x_j = \frac{|A_j|}{|A|}, \quad j = 1, \ldots, n$$

where the determinant $|A_j|$ *is defined as:*

$$|A_j| = \begin{vmatrix} a_{11} & \cdots & a_{1,j-1} & \boxed{b_1} & a_{1,j+1} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2,j-1} & \boxed{b_2} & a_{2,j+1} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & \boxed{b_n} & a_{n,j+1} & \cdots & a_{nn} \end{vmatrix}.$$

Note that $|A_j|$ *is obtained by replacing the* j*th column of* $|A|$ *by the column whose components are* $b_1, b_2, \ldots, b_n$.

Proof: We begin by assuming that $c_1, c_2, \ldots, c_n$ is a solution to $(*)$ so that

$$\begin{array}{rcl} a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n &=& b_1 \\ a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n &=& b_2 \quad (**) \\ \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots & & \\ a_{n1}c_1 + a_{n2}c_2 + \cdots + a_{nn}c_n &=& b_n \end{array}$$

We multiply the first of the equations in $(**)$ by the cofactor A_{11} of the element a_{11} in the coefficient matrix, then we multiply the second equation by A_{21}, the third by A_{31}, and so on, and finally the last equation by A_{n1}. Then we add all the equations so obtained. The result is

$$(a_{11}A_{11} + a_{21}A_{21} + \cdots + a_{n1}A_{n1})c_1 + (a_{12}A_{11} + a_{22}A_{21} + \cdots + a_{n2}A_{n1})c_2 + \cdots$$
$$\cdots + (a_{1n}A_{11} + a_{2n}A_{21} + \cdots + a_{nn}A_{n1})c_n = b_1A_{11} + b_2A_{21} + \cdots + b_nA_{n1}.$$

By Lemma 3.2.4, the coefficients of all the other $c_j (j \neq 1)$ vanish. So, we have

$$(a_{11}A_{11} + a_{21}A_{21} + \cdots + a_{n1}A_{n1})c_1 = b_1A_{11} + b_2A_{21} + \cdots + b_nA_{n1}.$$

This implies that $|A|c_1 = A_1$, so that $c_1 = A_1/|A|$. In a completely analogous way, we can obtain $c_j = A_j/|A|$ for each $j = 1, \ldots, n$. This completes the proof. ■

If the right-hand side of the equation system $(*)$ consists only of zeros, so that it can be written in matrix form as $Ax = 0$, the system is called *homogeneous*. A homogeneous system will always have the *trivial solution* $x_1 = x_2 = \cdots = x_n = 0$.

Lemma 3.3.1 *$Ax = 0$ has nontrivial solutions if and only if $|A| = 0$.*

Exercise 3.3.1 *Prove Lemma 3.3.1.*

Exercise 3.3.2 *Use Cramer's rule to solve the following system of equations:*

$$\begin{aligned} 2x_1 - 3x_2 &= 2 \\ 4x_1 - 6x_2 + x_3 &= 7 \\ x_1 + 10x_2 &= 1. \end{aligned}$$

3.4 Vectors

An *n-vector* is an ordered n-tuple of numbers. It is often convenient to regard the rows and columns of a matrix as vectors, and an *n-vector* can be understood either as a $1 \times n$ matrix $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ (a *row vector*) or as an $n \times 1$ matrix $\mathbf{a}^T = (a_1, a_2, \ldots, a_n)^T$ (a column vector). The operations of addition, subtraction and multiplication by scalars of vectors are defined in the obvious way.

The *dot product* (or *inner product*) of the n-vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ is defined as

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n = \sum_{i=1}^{n} a_ib_i.$$

The dot product of the n-vectors has the following properties:

Proposition 3.4.1 *If $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are n-vectors and α is a scalar, then*

1. $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$,
2. $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$,
3. $(\alpha\mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot (\alpha\mathbf{b}) = \alpha(\mathbf{a} \cdot \mathbf{b})$.
4. $\mathbf{a} \cdot \mathbf{a} = 0 \Longrightarrow \mathbf{a} = \mathbf{0}$

5. $(\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}+\mathbf{b}) = \mathbf{a}\cdot\mathbf{a} + 2(\mathbf{a}\cdot\mathbf{b}) + \mathbf{b}\cdot\mathbf{b}$.

Proof: We skip the proof. ■

The *Euclidean norm* or *length* of the vector $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ is defined:

$$\|\mathbf{a}\| = \sqrt{\mathbf{a}\cdot\mathbf{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

Note that $\|\alpha\mathbf{a}\| = |\alpha|\|\mathbf{a}\|$ for all scalars and vectors. The following useful inequalities hold.

Lemma 3.4.1 (Cauchy-Schwartz inequality) *For any* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$|\mathbf{a}\cdot\mathbf{b}| \le \|\mathbf{a}\|\cdot\|\mathbf{b}\|.$$

Proof: Define $f(t)$ as

$$f(t) = (t\mathbf{a}+\mathbf{b})\cdot(t\mathbf{a}+\mathbf{b}),$$

where $t \in \mathbb{R}$. Because of the definition of dot products, we have $f(t) \ge 0$ for any $t \in \mathbb{R}$.

$$f(t) = t^2\|\mathbf{a}\|^2 + 2t(\mathbf{a}\cdot\mathbf{b}) + \|\mathbf{b}\|^2.$$

Then, using the formula, we solve the above equation with respect to t:

$$t = \frac{-(\mathbf{a}\cdot\mathbf{b}) \pm \sqrt{(\mathbf{a}\cdot\mathbf{b})^2 - \|\mathbf{a}\|^2\|\mathbf{b}\|^2}}{\|\mathbf{a}\|^2}$$

Since $f(t) \ge 0$ for any $t \in \mathbb{R}$, we must have

$$|\mathbf{a}\cdot\mathbf{b}| \le \|\mathbf{a}\|\|\mathbf{b}\|. \ ■$$

Lemma 3.4.2 (Minkowski inequality) *For any* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\|\mathbf{a}+\mathbf{b}\| \le \|\mathbf{a}\| + \|\mathbf{b}\|.$$

Exercise 3.4.1 *Prove Lemma 3.4.2. (Hint: It suffices to show that* $\|\mathbf{a}+\mathbf{b}\|^2 \le (\|\mathbf{a}\| + \|\mathbf{b}\|)^2$*)*

Cauchy-Schwartz inequality implies that, for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$-1 \le \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \le 1.$$

Thus, the *angle* θ between nonzero vectors $\mathbf{a}$ and $\mathbf{b} \in \mathbb{R}^n$ is defined by

$$\cos\theta = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}, \quad \theta \in [0, \pi]$$

This definition reveals that $\cos\theta = 0$ if and only if $\mathbf{a} \cdot \mathbf{b} = 0$. Then $\theta = \pi/2$. In symbols,

$$\mathbf{a} \perp \mathbf{b} \iff \mathbf{a} \cdot \mathbf{b} = 0$$

The *hyperplane* in $\mathbb{R}^n$ that passes through the point $\mathbf{a} = (a_1, \ldots, a_n)$ and is orthogonal to the nonzero vector $\mathbf{p} = (p_1, \ldots, p_n)$, is the set of all points $\mathbf{x} = (x_1, \ldots, x_n)$ such that

$$\mathbf{p} \cdot (\mathbf{x} - \mathbf{a}) = 0.$$

3.5 Complex Numbers

Simple quadratic equations like $x^2 + 1 = 0$ and $x^2 + 4x + 8 = 0$ have no solution within the real number system. The standard formula for solving the equation $x^2 + 4x + 8 = 0$ yields $x = -2 \pm \sqrt{-4} = -2 \pm 2\sqrt{-1}$. By pretending that $\sqrt{-1}$ is a number i whose square is -1, we make i a solution of the equation $i^2 = -1$.

Mathematical formalism regard complex numbers as 2-vectors (a, b). We usually write this complex number as $a+bi$, where a and b are real numbers. The operations of addition, subtraction, and multiplication are defined by

$$\begin{aligned}(a + bi) + (c + di) &= (a + c) + (b + d)i,\\ (a + bi) - (c + di) &= (a - c) + (b - d)i,\\ (a + bi)(c + di) &= (ac - bd) + (ad + bc)i,\end{aligned}$$

respectively. The division of two complex numbers is

$$\frac{a + bi}{c + di} = \frac{(a + bi)(c - di)}{(c + di)(c - di)} = \frac{(ac + bd) + (bc - ad)i}{c^2 + d^2}.$$

3.5.1 Trigonometric Form of Complex Numbers

Each complex number $z = x + yi = (x, y)$ can be represented by a point in the plane. I could use *polar coordinates*. Let θ be the angle (measured in radians) between the positive real axis and the vector from the origin to the point (x, y), and let r be the distance from the origin to the same point. Then, $x = r\cos\theta$ and $y = r\sin\theta$, so

$$z = x + yi = r(\cos\theta + i\sin\theta).$$

The distance from the origin to the point (x, y) is $r = \sqrt{x^2 + y^2}$. This is called the *modulus* of the complex number, denoted by $|z|$.

If $z = x + iy$, then the *complex conjugate* of z is defined as $\bar{z} = x - iy$. We see that $\bar{z}z = x^2 + y^2 = |z|^2$, where $|z|$ is the modulus of z. Multiplication of complex numbers have a neat geometric interpretation:

$$r_1(\cos\theta_1 + i\sin\theta_1)r_2(\cos\theta_2 + i\sin\theta_2) = r_1 r_2\left[\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)\right]$$

Division of complex numbers have a neat geometric interpretation:

$$\frac{r_1(\cos\theta_1 + i\sin\theta_1)}{r_2(\cos\theta_2 + i\sin\theta_2)} = \frac{r_1}{r_2}\left[\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)\right]$$

3.5.2 Euler's Formula

Euler's formula states that for any real number $x \in \mathbb{R}$,

$$e^{ix} = \cos x + i \sin x.$$

where e is the base of the natural logarithm, i is the imaginary unit, and cos and sin are the trigonometric functions of cosine and sine respectively, with the argument x given in radians. This formula can be interpreted as saying that the function e^{ix} traces out the unit circle in the complex number plane as x ranges through the real numbers. Here, x is the angle that a line connecting the origin with a point on the unit circle makes with the positive real axis, measured counter clockwise and in radians

I will have the following important formula: for any $x, y \in \mathbb{R}$,

$$e^{ix}e^{iy} = e^{i(x+y)} = \cos(x+y) + i\sin(x+y).$$

Euler's formula provides a means of conversion between cartesian coordinates and polar coordinates. Any complex number $z = x + iy$ can be written as

$$\begin{aligned} z &= x + iy = |z|(\cos\theta + i\sin\theta) = re^{i\theta} \\ \bar{z} &= x - iy = |z|(\cos\theta - i\sin\theta) = re^{-i\theta} \end{aligned}$$

where x is the real part of z, y is the imaginary part of z, $r = |z| = \sqrt{x^2+y^2}$, and $\tan\theta = y/x$. Note that θ is called the *argument* of z, i.e., the angle between the x axis and the vector z measured counterclockwise and in radians.

Let $\alpha = r_1 e^{i\theta_1}$ and $\beta = r_2 e^{i\theta_2}$ be two arbitrary complex numbers using Euler's formula.

$$\alpha\beta = r_1 r_2 e^{i\theta_1} e^{i\theta_2} = r_1 r_2 e^{i(\theta_1+\theta_2)} = r_1 r_2(\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2).$$

Let $\alpha = r_1 e^{i\theta_1}$ and $\beta = r_2 e^{i\theta_2}$ be two arbitrary complex numbers using Euler's formula. Assuming that $r_2 \neq 0$,

$$\frac{\alpha}{\beta} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \frac{r_1}{r_2} e^{i(\theta_1-\theta_2)} = \frac{r_1}{r_2}(\cos(\theta_1-\theta_2) + i\sin(\theta_1-\theta_2)).$$

3.6 Number Fields

Linear algebra makes use of number systems (number fields). By a *number field* I mean any K of objects, called "numbers," which, when subjected to the four arithmetic operations again give elements of K. More precisely, operations have the following properties (field axioms):

(a) To every pair of numbers α and β in K there corresponds a unique number $\alpha+\beta$ in K, called the *sum* of α and β, where

(1) $\alpha + \beta = \beta + \alpha$ for every $\alpha, \beta \in K$ (*addition is commutative*);

(2) $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ for every $\alpha, \beta, \gamma \in K$ (*addition is assortative*);

(3) There exists a number 0 (*zero*) in K such that $0 + \alpha = \alpha$ for every $\alpha \in K$;

(4) For every $\alpha \in K$, there exists a number (*negative element*) $\gamma \in K$ such that $\alpha + \gamma = 0$.

The solvability of the equation $\alpha + \gamma = 0$ for every $\alpha \in K$ allows us to carry out the operation of subtraction, by defining the *difference* $\beta - \alpha$ as the sum of the numbers β and teh solution γ of the equation $\alpha + \gamma = 0$.

(b) To every pair of numbers α and β in K there corresponds a unique number $\alpha \cdot \beta$ (or $\alpha\beta$) in K, called the *product* of α and β, where

(5) $\alpha\beta = \beta\alpha$ for every $\alpha, \beta \in K$ (*multiplication is commutative*);

(6) $(\alpha\beta)\gamma = \alpha(\beta\gamma)$ for every $\alpha, \beta, \gamma \in K$ (*multiplication is assortative*);

(7) There exists a number 1 ($\neq 0$) in K such that $1 \cdot \alpha = \alpha$ for every $\alpha \in K$;

(8) For every $\alpha \neq 0$ in K, there exists a number (*reciprocal element*) γ in K such that $\alpha\gamma = 1$.

(c) *Multiplication is distributive over addition*, i.e.,

(9) $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$ for every $\alpha, \beta, \gamma \in K$[2]

I provide the most commonly encountered examples of number fields.

Example 3.6.1 (The Field of Rational Numbers) *The field of rational numbers, i.e., of quotients p/q where p and $q \neq 0$ are the ordinary integers subject to the ordinary operations of arithmetic.*

It should be noted that the integers by themselves *do not form a field*, since they do not satisfy axiom (8).

Example 3.6.2 (The Field of Real Numbers) *The field of real numbers has the set of all points of the real line as its geometric counterpart. An axiomatic treatment of the field of real numbers is achieved by supplementing axioms (1) through (9) with the axioms of order and the least upper bound axiom.*

Example 3.6.3 (The Field of Complex Numbers) *The field of complex numbers of the form $a + ib$, where a and b are real numbers (i is not a real number), equipped with the following operations of addition and multiplication:*

$$\begin{aligned}(a_1 + ib_1) + (a_+ib_2) &= (a_1 + a_2) + i(b_1 + b_2),\\ (a_1 + ib_1)(a_2 + ib_2) &= (a_1a_2 - b_1b_2) + i(a_1b_2 + a_2b_1).\end{aligned}$$

[2]Note that axioms (5) and (9) also imply $(\alpha + \beta)\gamma = \alpha\gamma + \beta\gamma$.

For numbers of the form $a + i0$, these operations reduce to the corresponding operations for real numbers; briefly I write $a + i0 = a$ and call complex numbers of this form ***real****. Thus, it can be said that the field of complex numbers has a subset (subfield) isomorphic to the field of real numbers. Complex numbers of the form $0 + ib$ are said to be* ***(purely) imaginary*** *and are designated briefly by ib. It follows from the multiplication rule that*

$$i^2 = i \cdot i = (0 + i1)(0 + i1) = -1.$$

Henceforth, I will designate the field of real numbers by $\mathbb{R}$ and the field of complex numbers by $\mathbb{C}$. According to the "fundamental theorem of algebra," we can not only carry out the four arithmetic operations in $\mathcal{C}$ but also solve any algebraic equation

$$z^n + a_1 z^{n-1} + \cdots + z_n = 0.$$

The field $\mathbb{R}$ of real numbers does not have this property. For example, the equation $z^2 + 1 = 0$ has no solutions in the field $\mathbb{R}$.

Many of the subsequent considerations are valid for any number field. In what follows, I will use the letter K to denote an *arbitrary* number field. If some property is true for the field K, then it is automatically true for the field $\mathbb{R}$ and the field $\mathbb{C}$, which are special cases of the general field K.

3.7 Linear Spaces

The concept of a linear space generalizes that of the set of all vectors. The generalization consists first in getting away from the concrete nature of the objects involved (directed line segments) without changing the properties of the operations on the objects, and secondly in getting away from the concrete nature of the admissible numerical factors (real numbers). This leads the following definition.

Definition 3.7.1 *A set V is called a linear (or affine) space over a field K if*

1. *Given any two elements $x, y \in V$, there is a rule (the addition rule) leading to a (unique) element $x + y \in V$, called the sum of x and y;*

2. *Given any element $x \in V$ and any number $\lambda \in K$, there is a rule (the scalar multiplication) leading to a (unique) element $\lambda x \in V$, called the product of the element x and the number λ;*

3. *These two rules obey the axioms listed below, VS1 and VS2.*

VS 1: The addition rule has the following properties:

1. $x + y = y + x$ for every $x, y \in V$;

2. $(x + y) + z = x + (y + z)$ for every $x, y, z \in V$;

3. There exists an element $0 \in V$ (the *zero vector*) such that $x + 0 = x$ for every $x \in V$;

4. For every $x \in V$, there exists an element $y \in V$ (the *negative element*) such that $x + y = 0$.

VS 2: The rule for scalar multiplication has the following properties:

1. $1 \cdot x = x$ for every $x \in V$;

2. $\alpha(\beta x) = (\alpha\beta)x$ for every $x \in V$ and $\alpha, \beta \in K$;

3. $(\alpha + \beta)x = \alpha x + \beta x$ for every $x \in V$ and $\alpha, \beta \in K$;

4. $\alpha(x + y) = \alpha x + \alpha y$ for every $x \in V$ and every $\alpha \in K$.

The elements of a linear space will be called *vectors*, regardless of the fact that their concrete nature may be quite unlike the more familiar directed line segments.

Theorem 3.7.1 *The zero vector in a linear space is unique.*

Proof: The existence of at least one zero vector is asserted in Axiom 3 of VS 1. Suppose there are two zero vectors 0_1 and 0_2 in the space $\mathbf{K}$. Setting $x = 0_1$, $0 = 0_2$ in Axiom 3 of VS 1, we obtain

$$0_1 + 0_2 = 0_1.$$

Setting $x = 0_2, 0 = 0_1$ in the same axiom, we obtain

$$0_2 + 0_1 = 0_2.$$

By Axiom 1 of VS 1, the above two equations imply that $0_1 = 0_2$. ■

Theorem 3.7.2 *Every element in a linear space has a unique negative.*

Proof: We skip the proof. ■

Theorem 3.7.3 *The relation*

$$0 \cdot x = \mathbf{0}$$

holds for every element x in a linear space.

Proof: We skip the proof. ■

Theorem 3.7.4 *Given any element x of a linear space, the element*

$$y = (-1) \cdot x$$

serves as the negative of x.

Proof: We skip the proof. ■

The negative of a given element x will now be denoted by $-x$. The presence of a negative allows us to introduce the operation of subtraction, i.e., the *difference* $x-y$ is defined as the sum of x and $-y$.

A linear space over the field $\mathbb{R}$ of real numbers will be called *real*. A linear space over the field $\mathbb{C}$ of complex numbers will be called *complex*. If the nature of the elements $x, y, z, \ldots$ and the rules for operating on them are specified, then I call the linear space *concrete*.

Example 3.7.1 (The Space K^n) *An element of the space K^n is any ordered n-tuple*

$$x = (\xi_1, \xi_2, \ldots, \xi_n)$$

of n numbers from the field K. The numbers $\xi_1, \ldots, \xi_n$ are called the ***component*** *of the element x. The operations of addition and multiplication by a number $\lambda \in K$ are specified by the following rules:*

$$\begin{aligned}(\xi_1, \ldots, \xi_n) + (\eta_1, \ldots, \eta_n) &= (\xi_1 + \eta_1, \ldots, \xi_n + \eta_n)\\ \lambda(\xi_1, \ldots, \xi_n) &= (\lambda\xi_1, \ldots, \lambda\xi_n).\end{aligned}$$

It is easily verified that all axioms in VS 1 and VS 2 are satisfied. In particular, the element $\mathbf{0}$ is the n-tuple consisting of n zeros:

$$\mathbf{0} = (0, \ldots, 0).$$

If K is the field $\mathbb{R}$ of real numbers, I write $\mathbb{R}^n$ instead of K^n, while if K is the field $\mathbb{C}$ of complex numbers, I write $\mathbb{C}^n$ instead of K^n.

I introduce the (inner) dot product in a real linear space as follows:

Definition 3.7.2 (Inner Product (Dot Product)) *An inner product in a real linear space $\mathbb{R}$ is defined as a mapping $\langle \cdot, \cdot \rangle$ from $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with the following property: for any $x, y, z \in R$ and $\alpha \in \mathbb{R}$,*

1. *$\langle x, x \rangle \geq 0$ where $\langle x, x \rangle = 0$ if and only if $x = \mathbf{0}$;*
2. *$\langle x, y \rangle = \langle y, x \rangle$;*
3. *$\langle (\alpha x), y \rangle = \alpha \langle x, y \rangle$;*
4. *$\langle x, (y + z) \rangle = \langle x, y \rangle + \langle x, z \rangle$.*

Although I have discussed $\mathbb{R}^n$ extensively so far and pretended to know exactly what it is, this is the first time I can formally define a Euclidean space.

Definition 3.7.3 *A* ***Euclidean space*** *is a linear space $\mathbb{R}$ equipped with an inner product $\langle \cdot, \cdot \rangle$.*

3.8 Linear Independence

Definition 3.8.1 *The n vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ of the linear space $\mathbf{K}$ over a field K are **linearly dependent** if there exist numbers $c_1, c_2, \ldots, c_n \in K$, not all zero, such that*

$$c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n = \mathbf{0}$$

*If this equation holds only when $c_1 = c_2 = \cdots = c_n = 0$, then the vectors are **linearly independent**.*

Exercise 3.8.1 *$\mathbf{a}_1 = (1, 2), \mathbf{a}_2 = (1, 1)$, and $\mathbf{a}_3 = (5, 1) \in \mathbb{R}^2$. Show that $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are linearly dependent.*

The next result is a characterization of linear independence and dependence in a linear space.

Proposition 3.8.1 *A set of n vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ of the linear space $\mathbf{K}$ is linearly dependent if and only if at least one of them can be written as a linear combination of the others. Or equivalently: A set of vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ of the linear space $\mathbf{K}$ is linearly independent if and only if none of them can be written as a linear combination of the others.*

Proof: Suppose that $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ are linearly dependent. Then the equation $c_1\mathbf{a}_1 + \cdots + c_n\mathbf{a}_n = 0$ holds with at least *one* of the coefficients c_i differently from 0. We can, without loss of generality, assume that $c_1 \neq 0$. Solving the equation for $\mathbf{a}_1$ yields

$$\mathbf{a}_1 = -\frac{c_2}{c_1}\mathbf{a}_2 - \cdots - \frac{c_n}{c_1}\mathbf{a}_n.$$

Thus, $\mathbf{a}_1$ is a linear combination of the other vectors. ■

3.8.1 Linear Dependence between Columns

Suppose I am given n columns of numbers with m numbers in each: $\mathbf{a}_1, \ldots, \mathbf{a}_n \in K^m$ where

$$\mathbf{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix}, \ldots, \mathbf{a}_m = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix}.$$

I multiply every element of the first column by some number λ_1, every element of the second column by λ_2, so on, and finally every element of the last (n-th) column by λ_n. I then add corresponding elements of the columns. As a result, I get a new column of numbers, whose elements I denote by $c_1, c_2, \ldots, c_m$. That is,

$$\lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \cdots + \lambda_n\mathbf{a}_n = \mathbf{c},$$

where

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix}.$$

The column $\mathbf{c}$ is called a *linear combination of the columns* $\mathbf{a}_1, \ldots, \mathbf{a}_n$, and the numbers $\lambda_1, \ldots, \lambda_n$ are called the *coefficients* of the linear combination. As special cases of the linear combination $\mathbf{c}$, I have the *sum* of the columns if $\lambda_1 = \cdots = \lambda_n = 1$ and the *product* of a column by a number if $n = 1$.

Suppose now that these columns make up an $n \times n$ matrix A. Then, I obtain the following.

Theorem 3.8.1 *If one of the columns of an $n \times n$ matrix A is a linear combination of the other columns, then $|A| = 0$.*

Proof: For each $j \in \{1, \ldots, n\}$, define

$$\mathbf{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}.$$

We define an $n \times n$ matrix A as $A = (\mathbf{a}_1 \mathbf{a}_2 \cdots \mathbf{a}_n)$. Since one of the columns of A is a linear combination of the columns, we can assume, without loss of generality, that there exist $c_1, \ldots, c_{n-1} \in \mathbb{N}$, not all zeros such that $\mathbf{a}_n = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_{n-1}\mathbf{a}_{n-1}$. Since the value of $|A|$ remains unchanged if a multiple of one column is added to another column, we obtain the following:

$$\begin{aligned}
|A| &= |\mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_{n-1}(\mathbf{a}_n - c_1\mathbf{a}_1)| \\
&= |\mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_{n-1}(\mathbf{a}_n - c_1\mathbf{a}_1 - c_2\mathbf{a}_2)| \\
&\vdots \qquad\qquad\qquad \vdots \\
&= |\mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_{n-1}(\mathbf{a}_n - (c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_{n-2}\mathbf{a}_{n-2})| \\
&= |\mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_{n-1}(\mathbf{a}_n - (c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_{n-2}\mathbf{a}_{n-2} + c_{n-1}\mathbf{a}_{n-1})| \\
&= |\mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_{n-1}\mathbf{0}| \\
&\quad (\because\ \mathbf{a}_n = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_{n-1}\mathbf{a}_{n-1}) \\
&= 0\ (\because \text{ we expand on the } n\text{-th column}). \blacksquare
\end{aligned}$$

Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in K^m$. Then,

$$A = (\mathbf{a}_1 \cdots \mathbf{a}_n) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

If k columns and k rows of this matrix is held fixed, then the elements appearing at the intersections of these columns and rows from a $k \times k$ square matrix, whose determinant is said to be a *minor of order k* of the original matrix A. This determinant may either be zero or nonzero. If, as I always assume, all $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are nonzero vectors, I can always find an integer r which has the following two properties:

1. The matrix A has a minor of order r which is nonzero; and
2. Every minor of the matrix A of order $r+1$ and higher (if exists) is zero.

The number r which has these properties is called the *rank* of the matrix A.

Definition 3.8.2 *The **rank** of a matrix A, written rank(A), is the maximum number of linearly independent column vectors in A. If A is the* $\mathbf{0}$ *matrix, we put* $rank(A) = 0$.

Henceforth, I shall assume that $r > 0$. The minor of order r which is different from zero is called the *basis minor* of the matrix A. Of course, A can have several basis minors, but they all have the same order r. The columns which contain the basis minor are called the *basis columns*. Concerning the basis columns, I have the following result:

Theorem 3.8.2 (Basis Minor Theorem) *Any column of the matrix A is a linear combination of its basis columns.*

Proof: We omit the proof. ■

Theorem 3.8.3 *Let A be an* $n \times n$ *matrix. Then,* $|A| = 0$ *if and only if there is a linear dependence between its columns of A.*

Proof: ($\Leftarrow$) This follows from Theorem 3.8.1. ($\Rightarrow$) We skip the proof this direction. ■

Exercise 3.8.2 *Prove the only-if ($\Rightarrow$) part of Theorem 3.8.3.*

3.8.2 Linear Dependence and Systems of Linear Equations

Consider the general *system of linear equations*:

$$\begin{array}{rcll} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & b_1 & \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & b_2 & \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & . & . & \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & . & . & (*) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & . & . & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = & b_m & \end{array}$$

can be written as $Ax = b$ where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \; b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

Let

$$A_b = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{pmatrix}$$

be the *augmented* matrix of the system $(*)$.

It turns out that the relationship between the ranks of A and A_b is crucial in determining whether system $(*)$ has a solution. Because all the columns in A occur in A_b, the rank of A_b is certainly greater than or equal to the rank of A. Moreover, because A_b contains only one more column than A, $\text{rank}(A_b) \leq \text{rank}(A) + 1$.

Theorem 3.8.4 *$Ax = b$ has at least one solution if and only if rank(A) = rank(A_b).*

Proof: Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n, \mathbf{b}$ be the column vectors in A_b. Suppose that $Ax = b$ has a solution $x = (x_1, \ldots, x_n)$ such that $x_1\mathbf{a}_1 + \cdots x_n\mathbf{a}_n = \mathbf{b}$. We multiply the first n columns in A_b by $-x_1, \ldots, -x_n$, respectively, and add each of the resulting column vectors to the last column in A_b. These elementary operations make the last column $\not\vdash$.

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} & b_1 - x_1a_{11} - \cdots - x_na_{1n} \\ a_{21} & \cdots & a_{2n} & b_2 - x_1a_{21} - \cdots - x_na_{2n} \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & b_n - x_1a_{n1} - \cdots - x_na_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} & 0 \\ a_{21} & \cdots & a_{2n} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & 0 \end{pmatrix}$$

Because elementary operations preserve the rank, this transformed matrix has the same rank as A, so $\text{rank}(A_b) = \text{rank}(A)$.

Assume that there exists $k \in \mathbb{N}$ such that $\text{rank}(A) = \text{rank}(A_b) = k$. This implies that k of the columns of A are linearly independent. To simply the notation, suppose that the first k columns $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are linearly independent. Because $\text{rank}(A_b) = k$, the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{b}$ are linearly dependent. Thus, there exits numbers $c_1, \ldots, c_k$ and β, not all equal to 0, such that $c_1\mathbf{a}_1 + \cdots + c_k\mathbf{a}_k + \beta\mathbf{b} = \mathbf{0}$. If $\beta = 0$, then $\mathbf{a}_1, \ldots, \mathbf{a}_n$ would not be linearly independent. Hence, $\beta \neq 0$. Then, we have

$$\mathbf{b} = -\frac{c_1}{\beta}\mathbf{a}_1 - \cdots - \frac{c_k}{\beta}\mathbf{a}_k.$$

For each $h \in \{1, \ldots, k\}$, let $x_h^0 = -c_k/\beta$. It follows that $(x_1^0, \ldots, x_k^0, 0, \ldots, 0)$ is a solution of $Ax = b$. This completes the proof.. ■

Remark: When we say $Ax = b$ has at least one solution, there may well be multiple solutions to $Ax = b$.

Theorem 3.8.5 *Suppose that system $(*)$ has solutions with rank(A) = rank(A_b) = k.*

1. *If $k < m$, then $m-k$ equations are **superfluous** in the sense that if we choose any subsystem of equations corresponding to k linearly independent rows, then any solution of these k equations also satisfies the remaining $m-k$ equations.*
2. *If $k < n$, there exist $n-k$ variables that can be chosen freely, whereas the remaining k variables are uniquely determined by the choice of these $n-k$ free variables. The system then has $n-k$ **degrees of freedom**.*

Proof: (Proof of *1.*) By the definition of rank, there exists k row vectors in A_b that are linearly independent, and any other row vector in A_b is a linear combination of those k vectors. To simplify notation, reorder the equations so that the first k row vectors in A_b are linearly independent. The other rows are dependent on these first k rows. Thus, for each $s \in \{k+1, \ldots, m\}$, there exist numbers $\lambda_{s1}, \lambda_{s2}, \ldots, \lambda_{sk}$ such that

$$\text{The } s\text{-th row: } (a_{s1}, a_{s2}, \ldots, a_{sn}, b_s) = \sum_{\ell=1}^{k} \lambda_{s\ell}(a_{\ell 1}, a_{\ell 2}, \ldots, a_{\ell n}, b_\ell) \ \ (*).$$

It follows from $(*)$ that $b_s = \sum_{\ell=1}^{k} \lambda_{s\ell} b_\ell$ for each $s \in \{k+1, \ldots, m\}$ and, for each $j \in \{1, \ldots, n\}$,

$$a_{sj} = \sum_{\ell=1}^{k} \lambda_{s\ell} a_{\ell j}.$$

Suppose that there exist a vector $(x_1^0, \ldots, x_n^0)$ that satisfies the first k equations. This implies

$$\sum_{j=1}^{n} a_{\ell j} x_j^0 = b_\ell, \ \ \forall \ell = 1, \ldots, k.$$

For each $s = k+1, \ldots, m$, we thus have

$$\sum_{j=1}^{n} a_{sj} x_j^0 = \sum_{j=1}^{n} \left(\sum_{\ell=1}^{k} \lambda_{s\ell} a_{\ell j} \right) x_j^0 = \sum_{\ell=1}^{k} \lambda_{s\ell} \left(\sum_{j=1}^{n} a_{\ell j} x_j^0 \right) = \sum_{\ell=1}^{k} \lambda_{s\ell} b_\ell = b_s.$$

This confirms that if the vector $(x_1^0, \ldots, x_n^0)$ satisfies the first k equations, then it automatically satisfies the last $m-k$ equations.

(Proof of *2.*) We omit the proof of this portion. ■

Suppose that $(*)$ has two solutions $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$. Then,

$$u_1\mathbf{a}_1 + \cdots + u_n\mathbf{a}_n = \mathbf{b} \text{ and } v_1\mathbf{a}_1 + \cdots + v_n\mathbf{a}_n = \mathbf{b}$$

Subtracting the second equation from the first yields

$$(u_1 - v_1)\mathbf{a}_1 + \cdots + (u_n - v_n)\mathbf{a}_n = \mathbf{0}.$$

Let $c_1 = u_1 - v_1, \ldots, c_n = u_n - v_n$. The two solutions are different if and only if $c_1, \ldots, c_n$ are not all equal to 0. I conclude that if system $(*)$ has more than one solution, then the column vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are linearly dependent. This follows from Lemma 3.3.1 and Theorem 3.8.3. Equivalently, *if the column vectors* $\mathbf{a}_1, \ldots, \mathbf{a}_n$ *are linearly independent, then system* $(*)$ *has "at most" one solution.*[3]

Theorem 3.8.6 *The* n *column vectors* $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ *of the* $n \times n$ *matrix*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \textit{where} \quad a_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix} \quad j = 1, \ldots, n$$

are linearly independent if and only if $|A| \neq 0$.

Proof: The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are linearly independent iff the vector equation $x_1\mathbf{a}_1 + \cdots x_n\mathbf{a}_n = \mathbf{0}$ has only the trivial solution $x_1 = \cdots = x_n = 0$. This vector equation is equivalent to a homogeneous systems of equations, and therefore, it has only the trivial solution iff $|A| \neq 0$.[4] ■

3.9 Matrix Inverses

An operator B acting in a space X is called a *left inverse* of the operator A acting in the same space X if

$$BA = I$$

where I is the identity operator. The operator A is then called a *right inverse* of the operator B.

It is possible for an operator A to have many left inverses and no right inverses at all, or conversely, many right inverses and no left inverses at all. However, suppose A has both a left inverse P and a right inverse Q, so that

$$P = PI = P(AQ) = (PA)Q = IQ = Q.$$

[3] Is there anything to say about a question of when there is no solution? The answer is yes. I can use Farkas Lemma to check if there is any solution to the system.

[4] Recall that the system of linear equations is homogeneous if it is expressed by $Ax = 0$.

Fixing Q, I see that every left inverse coincides with P and hence is uniquely determined. In just the same way, the right inverse Q is uniquely determined under these circumstances. The uniquely determined operator $P = Q$, which is simultaneously both a left and a right inverse of the operator A, is called the *inverse* of the operator A and is denoted by A^{-1}. The operator A itself, with the inverse A^{-1}, is said to be *invertible* (*nonsingular*).

Let A be an invertible operator acting in an n-dimensional space X, and let $B = A^{-1}$ be its inverse. Choosing a basis $e_1, \dots, e_n$, let $A = (a_{ij})$ and $B = (b_{ij})$ be the matrices of the operators A and B in this basis.

I now find an explicit formula for the elements b_{ij} in terms of the elements a_{ij}. Fixing the row number i, I write down expressions for the elements of the ith row of the matrix $BA = I$:

$$\begin{array}{ccccccccc}
b_{i1}a_{11} & + & b_{i2}a_{21} & + & \cdots & + & b_{in}a_{n1} & = & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
b_{i1}a_{1,i-1} & + & b_{i2}a_{2,i-1} & + & \cdots & + & b_{in}a_{n,i-1} & = & 0 \\
b_{i1}a_{1i} & + & b_{i2}a_{2i} & + & \cdots & + & b_{in}a_{ni} & = & 1 \\
b_{i1}a_{1,i+1} & + & b_{i2}a_{2,i+1} & + & \cdots & + & b_{in}a_{n,i+1} & = & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
b_{i1}a_{1n} & + & b_{i2}a_{2n} & + & \cdots & + & b_{in}a_{nn} & = & 0
\end{array}$$

The unknowns $b_{i1}, \dots, b_{in}$ can be determined from this system of equations by using Cramer's rule, since $|A| \neq 0$ by our hypothesis. Expanding the determinant in the

numerator of the resulting expression for b_{ij} with respect to the jth column, I get

$$
\begin{aligned}
b_{ij} &= \frac{1}{|A|}\begin{vmatrix}
a_{11} & a_{21} & \cdots & a_{j-1,1} & 0 & a_{j+1,1} & \cdots & a_{n1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{1,i-1} & a_{2,i-1} & \cdots & a_{j-1,i-1} & 0 & a_{j+1,i-1} & \cdots & a_{n,i-1} \\
a_{1i} & a_{2i} & \cdots & a_{j-1,i} & 1 & a_{j+1,i} & \cdots & a_{ni} \\
a_{1,i+1} & a_{2,i+1} & \cdots & a_{j-1,i+1} & 0 & a_{j+1,i+1} & \cdots & a_{n,i+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{1n} & a_{2n} & \cdots & a_{j-1,n} & 0 & a_{j+1,n} & \cdots & a_{nn}
\end{vmatrix} \\
&\quad (\because\ |A| = |A^T|) \\
&= \frac{1}{|A|}\begin{vmatrix}
a_{11} & a_{12} & \cdots & a_{1,i-1} & a_{1i} & a_{1,i+1} & \cdots & a_{1n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{j-1,1} & a_{j-1,2} & \cdots & a_{j-1,i-1} & a_{j-1,i} & a_{j-1,i+1} & \cdots & a_{j-1,n} \\
0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\
a_{j+1,1} & a_{j+1,2} & \cdots & a_{j+1,i-1} & a_{j+1,i} & a_{j+1,i+1} & \cdots & a_{j+1,n} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & \cdots & a_{n,i-1} & a_{ni} & a_{n,i+1} & \cdots & a_{nn}
\end{vmatrix} \\
&= \frac{1}{|A|}\{0 \times A_{j1} + \cdots + 0 \times A_{j,i-1} + 1 \times A_{ji} + 0 \times A_{j,i+1} + \cdots + 0 \times A_{jn}\}
\end{aligned}
$$

This implies the following:

$$b_{ij} = \frac{A_{ji}}{|A|},$$

where A_{ji} is the cofactor of the element a_{ji} in the matrix A. In words, the element b_{ij} of the inverse matrix A^{-1} equals the ratio of the cofactor of the element a_{ij} of the original matrix A to the determinant of A. Thus, I proved the following:

Theorem 3.9.1 *Every nonsingular matrix A has a unique inverse matrix B such that*

$$AB = BA = I.$$

If $A = (a_{ij})_{n\times n}$ and $|A| \neq 0$, the unique inverse of A is given by

$$A^{-1} = \frac{1}{|A|}\text{adj}(A), \text{ where } \text{adj}(A) = \begin{pmatrix}
A_{11} & A_{21} & \cdots & A_{n1} \\
A_{12} & A_{22} & \cdots & A_{n2} \\
\vdots & \vdots & \ddots & \vdots \\
A_{1n} & A_{2n} & \cdots & A_{nn}
\end{pmatrix}$$

with A_{ij}, the *cofactor* of the element a_{ij}. Note carefully the order of the indices in the *adjoint matrix*, adj(A) with the column number preceding the row number. The matrix $(A_{ij})_{n\times n}$ is called the *cofactor matrix*, whose transpose is the adjoint matrix.

Exercise 3.9.1 *Define A as*

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Then, derive A^{-1}, assuming $|A| \neq 0$.

Lemma 3.9.1 *The following rules for inverses can be established.*

- $(A^{-1})^{-1} = A$,
- $(AB)^{-1} = B^{-1}A^{-1}$,
- $(A^T)^{-1} = (A^{-1})^T$,
- $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$, *where* $\alpha \in \mathbb{R}$.

Exercise 3.9.2 *Prove Lemma 3.9.1. If you find it difficult to prove, focus on the case when $n = 2$.*

Proposition 3.9.1 *Let A be a $n \times n$ matrix. Then, $|A^{-1}| = 1/|A|$.*

Exercise 3.9.3 *Prove Lemma 3.9.1. If you find it difficult to prove, focus on the case when $n = 2$.*

3.10 Quadratic Forms

A *quadratic form* in n variables is a bilinear form, that is, a function Q of the form

$$Q(x_1, \ldots, x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j = a_{11}x_{11}^2 + a_{12}x_1x_2 + \cdots + a_{ij}x_ix_j + \cdots + a_{nn}x_n^2.$$

where the a_{ij} are constants. Suppose we put $x = (x_1, \ldots, x_n)^T$ and $A = (a_{ij})$. Then, it follows from the definition of matrix multiplication that

$$Q(x_1, \ldots, x_n) = Q(x) = x^T Ax = x \cdot Ax.$$

Of course, $x_ix_j = x_jx_i$, so I can write $a_{ij}x_ix_j + a_{ji}x_jx_i = (a_{ij} + a_{ji})x_ix_j$. If I replace a_{ij} and a_{ji} by $(a_{ij} + a_{ji})/2$, then the new numbers a_{ij} and a_{ji} become equal without changing $Q(x)$. Thus, I can assume that $a_{ij} = a_{ji}$ for all i and j, which means that the matrix A is symmetric. Then A is called the *symmetric matrix associated with* Q, and Q is called a *symmetric quadratic form.*

Definition 3.10.1 *A quadratic form $Q(x) = x^T Ax$, as well as its associated symmetric matrix A, are said to be* ***positive definite, positive semidefinite, negative definite****, or* ***negative semidefinite*** *according as*

$$Q(x) > 0,\ Q(x) \geq 0,\ Q(x) < 0,\ Q(x) \leq 0,$$

for all $x \in \mathbb{R}^n \backslash \{\mathbf{0}\}$. The quadratic form $Q(x)$ is ***indefinite*** *if there exist vectors x^* and y^* in $\mathbb{R}^n \backslash \{0\}$ such that $Q(x^*) < 0$ and $Q(y^*) > 0$.*

Let $A = (a_{ij})$ be any $n \times n$ matrix. An arbitrary *principal minor* of A of *order* k is the determinant of the matrix obtained by deleting all but r rows and r columns in A with the same numbers. In particular, a principal minor of order k always includes exactly k elements of the main (principal) diagonal. I call the determinant $|A|$ itself a principal minor (no rows and columns are deleted). A principal minor is said to be a *leading principal minor* of order k $(1 \leq k \leq n)$, if it consists of the first "leading" rows and columns of $|A|$. The leading principal minors of A of order k is

$$D_k = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix}.$$

Exercise 3.10.1 *Consider a 3×3 matrix A:*

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Compute all the principal minors of A.

Theorem 3.10.1 *Consider the quadratic form*

$$Q(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \quad (a_{ij} = a_{ji})$$

with the associated symmetric matrix $A = (a_{ij})_{n \times n}$. Let D_k be the leading principal minor of A of order k and let Δ_k denote an arbitrary principal minor of A of order k. Then we have

1. *Q is positive definite $\iff D_k > 0$ for $k = 1, \ldots, n$*
2. *Q is positive semidefinite $\iff \Delta_k \geq 0$ for all Δ_k and $k = 1, \ldots, n$.*
3. *Q is negative definite $\iff (-1)^k D_k > 0$ for $k = 1, \ldots, n$*
4. *Q is negative semidefinite $\iff (-1)^k \Delta_k \geq 0$ for all Δ_k and $k = 1, \ldots, n$.*

Proof: We only prove this for $n = 2$. Then, the quadratic form is

$$Q(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$$

After some manipulation through *perfect square*, we obtain

$$Q(x_1, x_2) = a_{11}\underbrace{\left(x_1 + \frac{a_{12}}{a_{11}}x_2\right)^2}_{>0} + \left(a_{22} - \frac{a_{12}^2}{a_{11}}\right)\underbrace{x_2^2}_{\geq 0}$$

Thus, we obtain

- $Q(x_1, x_2) > 0 \ \forall x_1, x_2 \Longleftrightarrow a_{11} > 0$ and $a_{11}a_{22} - a_{12}^2 > 0$.
- $Q(x_1, x_2) < 0 \ \forall x_1, x_2 \Longleftrightarrow a_{11} < 0$ and $a_{11}a_{22} - a_{12}^2 > 0$. ■

3.10.1 Quadratic Forms with Linear Constraints

In constrained optimization theory, the second-order conditions involve the signs of quadratic forms subject to homogeneous linear constraints.

Let $Q = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$ and assume that the variables are subject to the homogeneous linear constraint $b_1x_1 + b_2x_2 = 0$, where $b_1 \neq 0$. If $b_1 \neq 0$, we have $x_1 = -b_2x_2/b_1$. Plugging $x_1 = -b_2x_2/b_1$ into Q, we obtain

$$\begin{aligned} Q &= a_{11}\left(-\frac{b_2x_2}{b_1}\right)^2 + 2a_{12}\left(-\frac{b_2x_2}{b_1}\right)x_2 + a_{22}x_2^2 \\ &= \frac{1}{b_1^2}\left(a_{11}b_2^2 - 2a_{12}b_1b_2 + a_{22}b_1^2\right)x_2^2. \end{aligned}$$

It is easy to see

$$a_{11}b_2^2 - 2a_{12}b_1b_2 + a_{22}b_1^2 = -\begin{vmatrix} 0 & b_1 & b_2 \\ b_1 & a_{11} & a_{12} \\ b_2 & a_{12} & a_{22} \end{vmatrix}.$$

Therefore,

$$Q \text{ is positive definite (PD) subject to } b_1x_1 + b_2x_2 = 0 \ \Leftrightarrow \begin{vmatrix} 0 & b_1 & b_2 \\ b_1 & a_{11} & a_{12} \\ b_2 & a_{12} & a_{22} \end{vmatrix} < 0.$$

$$Q \text{ is negative definite (ND) subject to } b_1x_1 + b_2x_2 = 0 \ \Leftrightarrow \begin{vmatrix} 0 & b_1 & b_2 \\ b_1 & a_{11} & a_{12} \\ b_2 & a_{12} & a_{22} \end{vmatrix} > 0.$$

This is also valid when $b_1 = 0$ but $b_2 \neq 0$.

We move on to the general case of Q (i.e., $n \geq 2$) when there are the linear constraints:

$$Q(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \quad (a_{ij} = a_{ji})$$

subject to m linear homogeneous constraints:

$$\begin{array}{rcl} b_{11}x_1 + \cdots + b_{1n}x_n &=& 0 \\ b_{21}x_1 + \cdots + b_{2n}x_n &=& 0 \\ \vdots \qquad\qquad \vdots &=& \vdots \ \Leftrightarrow \mathbf{Bx} = \mathbf{0} \ (*) \\ b_{m1}x_1 + \cdots + b_{mn}x_n &=& 0, \end{array}$$

where $\mathbf{B}$ is an $m \times n$ matrix.

Definition 3.10.2 *The following are the definitions of definiteness subject to linear constraints.*

1. *Q is positive definite (PD) subject to the linear constraints $(*)$ if $Q(x) > 0$ for all $x = (x_1, \ldots, x_n) \neq \mathbf{0}$ that satisfy $(*)$*
2. *Q is negative definite (ND) subject to the linear constraints $(*)$ if $Q(x) < 0$ for all $x = (x_1, \ldots, x_n) \neq \mathbf{0}$ that satisfy $(*)$*

Define the symmetric determinants

$$B_r = \begin{vmatrix} 0 & \cdots & 0 & b_{11} & \cdots & b_{1r} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & b_{m1} & \cdots & b_{mr} \\ b_{11} & \cdots & b_{m1} & a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ b_{1r} & \cdots & b_{mr} & a_{r1} & \cdots & a_{rr} \end{vmatrix}$$

The determinant B_r is the $(m+r)$th leading principal minor of the $(m+n) \times (m+n)$ bordered matrix

$$\begin{pmatrix} \mathbf{0}_{m\times m} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{A} \end{pmatrix}.$$

Theorem 3.10.2 *Assume that rank$(B) = m$. Then, a necessary and sufficient condition for the quadratic form Q to be positive definite (PD) subject to the linear constraints $(*)$ is*

$$(-1)^m B_r > 0, \ \forall r = m+1, \ldots, n.$$

The corresponding necessary and sufficient condition for Q to be negative definite (ND) subject to the linear constraints $(*)$ is

$$(-1)^r B_r > 0, \ \forall r = m+1, \ldots, n.$$

Proof: We omit the proof. ■

Chapter 4

Multivariate Calculus

4.1 Functions of a Single Variable

Roughly speaking, a function $y = f(x)$ is *differentiable* if it is both continuous and "smooth," with no breaks or kinks. The *derivative* of $f(\cdot)$ is a function giving, at each value of x, the slope of change in $f(x)$. I sometimes write

$$\frac{dy}{dx} = f^{'}(x).$$

to indicate that $f^{'}(x)$ gives us the (instantaneous) amount, dy, by which y changes per unit change, dx, in x. If the first derivative is a differentiable function, one can take its derivative which gets the second derivative of the original function

$$\frac{d^2y}{dx^2} = f^{''}(x).$$

If a function possesses a continuous derivatives $f^{'}, f^{''}, \ldots, f^{n}$, it is called n-times continuously differentiable, or a C^n function. Some rules of differentiation is provided below:

- For constants, α: $d/dx(\alpha) = 0$.
- For sums: $d/dx[f(x) \pm g(x)] = f^{'}(x) \pm g^{'}(x)$.
- Power rule: $d/dx(\alpha x^n) = n\alpha x^{n-1}$.
- Product rule: $d/dx[f(x)g(x)] = f(x)g^{'}(x) + f^{'}(x)g(x))$.
- Quotient rule: $d/dx[f(x)/g(x)] = (g(x)f^{'}(x) - f(x)g^{'}(x))/[g(x)]^2$.
- Chain rule: $d/dx[f(g(x))] = f^{'}(g(x))g^{'}(x)$.

Later in this chapter, I am going to discuss some of the above properties in details from a more general perspective. Until then, just keep in mind that you can use them anytime.

Theorem 4.1.1 (Rolle's Theorem) *Suppose that $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and continuously differentiable (C^1) on (a, b). If $f(a) = f(b) = 0$, then there is a point $c \in (a, b)$ such that $f^{'}(c) = 0$.*

Proof: If f is constant on $[a, b]$, then $f^{'}(c) = 0$ for all $c \in (a, b)$. In this case, we are done. We next assume that f is not constant on $[a, b]$. Then, we further assume without loss of generality that f is sometimes positive on (a, b). Since $[a, b]$ is a compact subset of $\mathbb{R}$ and f is continuous on $[a, b]$, by Theorem 2.6.6 (Weierstrass Theorem), f achieves both the global maximum and minimum points over $[a, b]$. By our assumption that f is sometimes positive on (a, b), there exists a global maximum point $c \in (a, b)$ such that $f(c) > 0$. Since c is an interior point in $[a, b]$, by Theorem 5.1.1 (the necessity of extreme points for optimization problem), c is a stationary point of f so that $f^{'}(c) = 0$. ■

Theorem 4.1.2 (Mean Value Theorem for Single-Variable Functions) *Let $I \subseteq \mathbb{R}$ be an interval on the real line and $f : I \to \mathbb{R}$ be a continuously differentiable (or C^1) function. For any $a, b \in I$, there exists a point c between a and b so that*

$$f(b) - f(a) = f^{'}(c)(b - a).$$

Proof: Fix $a, b \in I$ arbitrarily. For any $x \in \mathbb{R}$, we define the following function:

$$g(x) = f(b) - f(x) + \frac{f(b) - f(a)}{b - a}(x - b).$$

We assume without loss of generality that $a < b$. So, we have the closed interval $[a, b]$. We confirm the following:

$$\begin{aligned} g(a) &= f(b) - f(a) + \frac{f(b) - f(a)}{b - a}(a - b) = f(b) - f(a) - (f(b) - f(a)) = 0; \\ g(b) &= f(b) - f(b) + \frac{f(b) - f(a)}{b - a}(b - b) = 0. \end{aligned}$$

Since $f(\cdot)$ is C^1 on I, we have

$$g^{'}(x) = -f^{'}(x) + \frac{f(b) - f(a)}{b - a}.$$

This implies that $g(\cdot)$ is also C^1 on (a, b). Then, by Theorem 4.1.1 (Rolle's Theorem), there exists $c \in (a, b)$ such that $g^{'}(c) = 0$. This implies

$$g^{'}(c) = -f^{'}(c) + \frac{f(b) - f(a)}{b - a} = 0 \Rightarrow \frac{f(b) - f(a)}{b - a} = f^{'}(c).$$

This completes the proof. ■

Theorem 4.1.3 (Taylor Approximation of Order Two: Single-Variable Functions) *Let $I \subseteq \mathbb{R}$ be an interval on the real line and $f : I \to \mathbb{R}$ be a twice continuously differentiable (or C^2) function. For any points $a, a+h \in I$, there exists a point c_2 between a and $a+h$ such that*

$$f(a+h) - f(a) = f^{'}(a)h + \frac{1}{2}f^{''}(c_2)h^2.$$

Proof: For each $t \in I$, define

$$g_1(t) \equiv f(t) - \Big[f(a) + f^{'}(a)(t-a)\Big] - M_1(t-a)^2,$$

where

$$M_1 = \frac{1}{h^2}\Big[f(a+h) - f(a) - f^{'}(a)h\Big].$$

We compute the following:

$$\begin{aligned} g_1(a) &= f(a) - \Big[f(a) + f^{'}(a)(a-a)\Big] - M_1(a-a)^2 = f(a) - f(a) = 0. \\ g_1(a+h) &= f(a+h) - \Big[f(a) + f^{'}(a)h\Big] - M_1h^2 \\ &= f(a+h) - \Big[f(a) + f^{'}(a)h\Big] - \Big[f(a+h) - f(a) - f^{'}(a)h\Big] \\ &= 0. \end{aligned}$$

Since $g_1(t)$ is a continuous function, by Rolle's Theorem, there exists $c_1 \in (a, a+h)$ such that $g_1^{'}(c_1) = 0$.

We compute the following:

$$g_1^{'}(t) = f^{'}(t) - f^{'}(a) - 2M_1(t-a).$$

We next confirm the following:

$$g_1^{'}(a) = f^{'}(a) - f^{'}(a) - 2M_1(a-a) = 0$$

Since $g_1^{'}(c_1) = 0$ and $g_1^{'}(t)$ is a continuous function, by Rolle's Theorem, there exists $c_2 \in (a, c_1)$ such that $g_1^{''}(c_2) = 0$. We compute the following:

$$g_1^{''}(t) = f^{''}(t) - 2M_1.$$

When $t = c_2$, we have

$$g^{''}(c_2) = f^{''}(c_2) - 2M_1 = 0.$$

Plugging $M_1 = [f(a+h) - f(a) - f^{'}(a)h]/h^2$ into the above equation, we have

$$f^{''}(c_2) - \frac{2}{h^2}\Big[f(a+h) - f(a) - f^{'}(a)h\Big] = 0.$$

This further implies

$$f(a+h) - f(a) = f^{'}(a)h + \frac{f^{''}(c_2)}{2}h^2. \blacksquare$$

4.2 Real-Valued Functions of Several Variables

$f : D \to R$ is said to be a *real-valued* function if D is any set in $\mathbb{R}^n$ and $R \subseteq \mathbb{R}$. Define the following: $x \geq y$ if $x_i \geq y_i$ for every $i = 1, \ldots, n$; and $x \gg y$ if $x_i > y_i$ for every $i = 1, \ldots, n$.

Definition 4.2.1 *Let $f : D \to \mathbb{R}$, where D is a subset of $\mathbb{R}^n$. Then, $f(\cdot)$ is non-decreasing if $f(x) \geq f(y)$ whenever $x \geq y$. If, in addition, the inequality is strict whenever $x \gg y$, then we say that f is is increasing. If, instead, $f(x) > f(y)$ whenever $x \geq y$ and $x \neq y$, then we say that f is strongly increasing.*

Rather than having a single slope, a function of n variables can be thought to have n *partial slopes*, each giving only the rate at which y would change if one x_i alone were to change. Each of these partial slopes is called the *partial derivative*.

Definition 4.2.2 *Let $y = f(x_1, \ldots, x_n)$. The **partial** derivative of f with respect to x_i is defined as*

$$\frac{\partial f(x)}{\partial x_i} \equiv \lim_{h \to 0} \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}$$

$\partial y / \partial x_i$ or $f_i^{'}(x)$ are used to denote partial derivatives.

4.3 Gradients

If $z = F(x, y)$ and C is any number, I call the graph of the equation $F(x, y) = C$ a *level curve* for F. The slope of the level curve $F(x, y) = C$ at a point (x, y) is given by the formula:

$$F(x, y) = C \Longrightarrow y^{'} = \frac{dy}{dx} = -\frac{\partial F(x, y)/\partial x}{\partial F(x, y)/\partial y} = -\frac{F_1^{'}(x, y)}{F_2^{'}(x, y)}.$$

If (x_0, y_0) is a particular point on the level curve $F(x, y) = C$, the slope at (x_0, y_0) is $-F_1^{'}(x_0, y_0)/F_2^{'}(x_0, y_0)$. The equation for the *tangent hyperplane* T is

$$y - y_0 = -\left[F_1^{'}(x_0, y_0)/F_2^{'}(x_0, y_0)\right](x - x_0)$$

or, rearranging

$$F_1^{'}(x_0, y_0)(x - x_0) + F_2^{'}(x_0, y_0)(y - y_0) = 0.$$

Recalling the inner product, the equation can be written as

$$\left(F_1^{'}(x_0, y_0), F_2^{'}(x_0, y_0)\right) \cdot (x - x_0, y - y_0) = 0.$$

The vector $(F_1^{'}(x_0, y_0), F_2^{'}(x_0, y_0))$ is said to be the *gradient* of F at (x_0, y_0) and it is often denoted by $\nabla F(x_0, y_0)$ (∇ is pronounced as "nabla"). The vector

$(x-x_0, y-y_0)$ is a vector on the tangent hyperplane T which implies that $\nabla F(x_0, y_0)$ is *orthogonal* to the tangent hyperplane T at (x_0, y_0).

Suppose more generally that $F(x) = F(x_1, \ldots, x_n)$ is a function of n variables defined on an open set A in $\mathbb{R}^n$, and let $x^0 = (x_1^0, \ldots, x_n^0)$ be a point in A. The *gradient* of F at x^0 is the vector

$$\nabla F(x^0) = \left(\frac{\partial F(x^0)}{\partial x_1}, \cdots, \frac{\partial F(x^0)}{\partial x_n} \right)$$

of first-order partial derivatives. Let T_{x^0} be the tangent hyperplane that passes through $x^0 \in \mathbb{R}^n$. Then,

$$T_{x^0} = \left\{ x \in \mathbb{R}^n | \ \nabla F(x^0) \cdot (x - x^0) = 0 \right\}.$$

A function $f : S \to \mathbb{R}$ is *continuously differentiable* (or C^1) on an open set $S \subseteq \mathbb{R}^n$ if, for each $i = 1, \ldots, n$, $(\partial f / \partial x_i)(x)$ exists for all $x \in S$ and is continuous on S. f is *k-times continuously differentiable* or C^k on S if all the derivatives of f of order less than or equal to $k(\geq 1)$ exist and they are continuous on S.

4.4 Convex Sets

Convex sets are basic building blocks in virtually every area of economics. Convexity guarantees that the analysis is mathematically tractable and the results are clear-cut and "well-behaved."

Definition 4.4.1 *$S \subseteq \mathbb{R}^n$ is a **convex** set if for all $x, y \in S$, we have*

$$\alpha x + (1 - \alpha) y \in S,$$

for all $\alpha \in [0, 1]$

I say that z is a *convex combination* of x and y if $z = \alpha x + (1 - \alpha) y$ for some $\alpha \in [0, 1]$. A very simple and intuitive rule defining convex sets is: *A set is convex if and only if we can connect any two points in the set by a straight line that lies entirely within the set.*

Exercise 4.4.1 *Suppose that $p \gg 0$ and $w \geq 0$. Let $B(p, w) = \{x \in \mathbb{R}^n_+ | p \cdot x \leq w\}$ be the budget set of the consumer. Show that $B(p, w)$ is convex.*

Theorem 4.4.1 *Let S and T be convex sets in $\mathbb{R}^n$. Then $S \cap T$ is a convex set.*

Proof of Theorem 4.4.1: Let x and y be any two points in $S \cap T$. Because $x \in S \cap T$, we have $x \in S$ and $x \in T$. Similarly, we have $y \in S$ and $y \in T$. Let $z = \alpha x + (1 - \alpha) y$ for some $\alpha \in [0, 1]$ be any convex combination of x and y. $z \in S$ because S is convex and $z \in T$ because T is convex. Thus, $z \in S \cap T$. ■

Remark: It is easy to construct an example in which two sets S and T are convex but $S \cup T$ is not convex.

Proposition 4.4.1 *Let S and T be convex sets in $\mathbb{R}^n$. Define*

$$U = S + T \equiv \{u \in \mathbb{R}^n | \text{ there exist } s \in S \text{ and } t \in T \text{ s.t. } u = s + t\}.$$

Then, show that $S + T$ is convex.

Proof: We omit the proof. ■

Example 4.4.1 (Upper Contour Sets) *Let $u(\cdot) : \mathbb{R}^n \to \mathbb{R}$ be a utility function. Define $UC(x^0) = \{x \in \mathbb{R}^n_+ | u(x) \geq u(x^0)\}$. This $UC(x^0)$ is called the upper contour set which consists of all commodity vectors x that the individual values at least as good as x^0. In consumer theory, we usually assume that $UC(x^0)$ is convex for every $x^0 \in \mathbb{R}^n_+$.*

4.5 Concave and Convex Functions

A C^2 function of one variable $y = f(x)$ is said to be *concave (convex)* on the interval I if $f''(x) \leq (\geq)\ 0$ for all $x \in I$.

Definition 4.5.1 *A function $f(x) = f(x_1, \ldots, x_n)$ defined on a convex set S is **concave (convex)** on S if*

$$f(\lambda x + (1 - \lambda)x') \geq (\leq)\ \lambda f(x) + (1 - \lambda) f(x')$$

for all $x, x' \in S$ and all $\lambda \in [0, 1]$

Definition 4.5.2 *A function $f(x) = f(x_1, \ldots, x_n)$ defined on a convex set S is **strictly concave (convex)** on S if*

$$f(\lambda x + (1 - \lambda)x') > (<)\ \lambda f(x) + (1 - \lambda) f(x')$$

for all $x, x' \in S$ with $x \neq x'$ and all $\lambda \in (0, 1)$

Exercise 4.5.1 *Show that the sum of two concave functions, defined over nonempty convex subset S of $\mathbb{R}^n$, is concave.*

sub

4.6 Characterizing Concavity/Convexity via Second Derivatives

Suppose that $z = f(x) = f(x_1, \ldots, x_n)$ is a C^2 function in an open convex set S in $\mathbb{R}^n$. The matrix

$$D^2 f(x) = \left(f''_{ij}(x) \right)_{n \times n}$$

is called the *Hessian* (matrix) of f at x, and for each $r = 1, \ldots, n$,

$$D^2_{(r)} f(x) = \begin{vmatrix} f''_{11}(x) & f''_{12}(x) & \cdots & f''_{1r}(x) \\ f''_{21}(x) & f''_{22}(x) & \cdots & f''_{2r}(x) \\ \vdots & \vdots & \ddots & \vdots \\ f''_{r1}(x) & f''_{r2}(x) & \cdots & f''_{rr}(x) \end{vmatrix}$$

is the *leading principal minors* of the Hessian matrix $D^2 f(x)$ of order r. Here $f''_{ij}(x) = \partial^2 f(x)/\partial x_i \partial x_j$ for any $i, j = 1, \ldots, r$. Let $\Delta^2_{(r)} f(x)$ denote a principal minor of $D^2 f(x)$ of order r.

Checking the sign of the second derivative is often a quick way to decide whether a C^2 function of one variable is concave or convex. For functions of two or more variables, there is also an analogous test which is often used.

Theorem 4.6.1 (Second-Order Characterization of Concave (Convex) Functions) *Suppose that $f : S \to \mathbb{R}$ is a C^2 function where $S \subseteq \mathbb{R}^n$ is open and convex. For each $x \in S$, let $\Delta^2_{(r)} f(x)$ denote a generic principal minor of the Hessian matrix $D^2 f(x)$ of order r. Then*

1. *$f(\cdot)$ is convex in $S \iff \Delta^2_{(r)} f(x) \geq 0$ for all $x \in S$ and all $\Delta^2_{(r)} f(x), r = 1, \ldots, n \iff$ the Hessian matrix $D^2 f(x)$ is positive semidefinite (PSD) for any $x \in S$.*
2. *$f(\cdot)$ is concave in $S \iff (-1)^r \Delta^2_{(r)} f(x) \geq 0$ for all $x \in S$ and all $\Delta^2_{(r)} f(x), r = 1, \ldots, n \iff$ the Hessian matrix $D^2 f(x)$ is negative semidefinite (NSD) for any $x \in S$.*

Proof: We omit the proof. ■

When I only focus on functions of two variables, I obtain the following corollary.

Corollary 4.6.1 *Let $f : S \to \mathbb{R}$ be a C^2 function where $S \subseteq \mathbb{R}^2$ is open and convex. Then,*

1. *$f(\cdot)$ is convex $\iff f''_{11} \geq 0,\ f''_{22} \geq 0$, and $f''_{11} f''_{22} - (f''_{12})^2 \geq 0$.*
2. *$f(\cdot)$ is concave $\iff f''_{11} \leq 0,\ f''_{22} \leq 0$, and $f''_{11} f''_{22} - (f''_{12})^2 \geq 0$.*

Proof: We omit the proof. ■

Example 4.6.1 (When Cobb-Douglas Utility Function is Concave) *Define* $u : \mathbb{R}^2_+ \to \mathbb{R}$ *be the utility function of the consumer: for any* $x \in \mathbb{R}^2_+$, $u(x) = x_1^\alpha x_2^\beta$ *where* α *and* β *are positive real numbers.*

$$\begin{aligned} u_1 &= \alpha x_1^{\alpha-1} x_2^\beta \\ u_{11} &= -\alpha(1-\alpha) x_1^{\alpha-2} x_2^\beta \\ u_{12} = u_{21} &= \alpha\beta x_1^{\alpha-1} x_2^{\beta-1} \\ u_2 &= \beta x_1^\alpha x_2^{\beta-1} \\ u_{22} &= -\beta(1-\beta) x_1^\alpha x_2^{\beta-2}. \end{aligned}$$

Then,

$$D^2u(x) = \begin{pmatrix} -\alpha(1-\alpha)x_1^{\alpha-2}x_2^\beta & \alpha\beta x_1^{\alpha-1}x_2^{\beta-1} \\ \alpha\beta x_1^{\alpha-1}x_2^{\beta-1} & -\beta(1-\beta)x_1^\alpha x_2^{\beta-2} \end{pmatrix}.$$

$$\begin{aligned} u_{11}u_{22} - u_{12}u_{21} &= \alpha\beta(1-\alpha)(1-\beta)x_1^{2\alpha-2}x_2^{2\beta-2} - \alpha^2\beta^2 x_1^{2\alpha-2}x_2^{2\beta-2} \\ &= \alpha\beta x_1^{2\alpha-2}x_2^{2\beta-2}\{(1-\alpha)(1-\beta) - \alpha\beta\} \\ &= \alpha\beta x_1^{2\alpha-2}x_2^{2\beta-2}(1-\alpha-\beta). \end{aligned}$$

Then, $u(\cdot)$ *is concave if and only if* $0 < \alpha + \beta \leq 1$.

When I am concerned with "strict" concavity (convexity), I obtain a similar result to the case of concavity (convexity).

Theorem 4.6.2 (Second-Order (Partial) Characterization of Strict Concavity) *Suppose that* $f : S \to \mathbb{R}$ *is a* C^2 *function where* $S \subseteq \mathbb{R}^n$ *is open and convex. For* $x \in S$, *let* $D^2_{(r)}f(x)$ *be the Hessian matrix defined above. Then*

1. *the Hessian matrix* $D^2f(x)$ *is positive definite for any* $x \in S \iff D^2_{(r)}f(x) > 0$ *for all* $x \in S$ *and all* $r = 1, \ldots, n \implies f(\cdot)$ *is strictly convex.*
2. *the Hessian matrix* $D^2f(x)$ *is negative definite for any* $x \in S \iff (-1)^r D^2_{(r)}f(x) > 0$ *for all* $x \in S$ *and all* $r = 1, \ldots, n \implies f(\cdot)$ *is strictly concave.*

Remark: Note that this result is not a full characterization of strict concavity (convexity). Because I only obtain one direction of the result. That is, if the Hessian matrix is negative definite, the function is strictly concave but the converse is not necessarily true.

Proof: We omit the proof. ■

Once again, when I only focus on functions of two variables, I obtain the following corollary.

Corollary 4.6.2 *Let $f : S \to \mathbb{R}$ be a C^2 function where $S \subseteq \mathbb{R}^2$ is open and convex. Then,*

1. $f''_{11} > 0$ *and* $f''_{11}f''_{22} - (f''_{12})^2 > 0 \Longrightarrow f$ *is* ***strictly*** *convex.*
2. $f''_{11} < 0$ *and* $f''_{11}f''_{22} - (f''_{12})^2 > 0 \Longrightarrow f$ *is* ***strictly*** *concave.*

Proof: We omit the proof. ■

Example 4.6.2 (When Cobb-Douglas Utility Function is Strictly Concave) *We revisit the same example. Define $u : \mathbb{R}^2_+ \to \mathbb{R}$ as the utility function of a consumer such that for any $x \in \mathbb{R}^2_+$, $u(x) = x_1^\alpha x_2^\beta$ where α and β are positive real numbers.*

Then, $u(\cdot)$ is strictly concave if $0 < \alpha + \beta < 1$.

I continue the discussion of concave/convex functions. First, I provide a geometric interpretation of concave/convex functions. The following result is extremely important in both static and dynamic optimization.

Theorem 4.6.3 (First-Order Characterization of Concavity) *Suppose that $f : S \to \mathbb{R}$ is a C^1 function where $S \subseteq \mathbb{R}^n$ is open and convex. Then*

(1) $f(\cdot)$ is concave in S if and only if

$$f(x) - f(x^0) \leq \nabla f(x^0) \cdot (x - x^0) = \sum_{i=1}^{n} \frac{\partial f(x^0)}{\partial x_i}(x_i - x_i^0)$$

for all $x, x^0 \in S$.

(2) $f(\cdot)$ is strictly concave if and only if the above inequality is always strict when $x \neq x^0$.

(3) The corresponding result for convex (strictly convex) functions is obtained by changing $\leq$ to $\geq$ ($<$ to $>$) in the above inequality.

Remark: Geometrically, this result says that the tangent at any point on the graph will lie above the graph.

Proof: (1) ($\Longrightarrow$) Let $x, x^0 \in S$. Since $f(\cdot)$ is concave,

$$\lambda f(x) + (1 - \lambda) f(x^0) \leq f(\lambda x + (1 - \lambda)x^0)$$

for all $\lambda \in (0, 1)$. Rearranging the above inequality, for all $\lambda \in (0, 1)$, we obtain

$$f(x) - f(x^0) \leq \frac{f(x^0 + \lambda(x - x^0)) - f(x^0)}{\lambda} \quad (*)$$

Let $\lambda \to 0$. Note that $(x - x^0)$ is the direction. The right hand side of $(*)$ then approaches $\nabla f(x^0) \cdot (x - x^0)$. $(\Longleftarrow)$ Let $x, x^0 \in S$ and $\lambda \in (0,1)$. Define $z = \lambda x + (1-\lambda)x^0$. Notice that $z \in S$ because S is convex. By our hypothesis, we have

$$\begin{aligned} f(x) - f(z) &\leq \nabla f(z) \cdot (x - z) \quad \text{(i)} \\ f(x^0) - f(z) &\leq \nabla f(z) \cdot (x^0 - z) \quad \text{(ii)} \end{aligned}$$

Multiplying the inequality in (i) by $\lambda > 0$ and the inequality in (ii) by $1 - \lambda > 0$, we obtain

$$\lambda \left(f(x) - f(z)\right) + (1-\lambda)\left(f(x^0) - f(z)\right) \leq \nabla f(z) \cdot \left[\lambda(x - z) + (1-\lambda)(x^0 - z)\right] \quad \text{(iii)}$$

Here $\lambda(x - z) + (1-\lambda)(x^0 - z) = \lambda x + (1-\lambda)x^0 - z = 0$, so the right hand side of (iii) is 0. Thus, rearranging (iii) gives

$$\lambda f(x) + (1-\lambda) f(x^0) \leq f(z) = f(\lambda x + (1-\lambda)x^0)$$

because $z = \lambda x + (1-\lambda)x^0$. This shows that $f(\cdot)$ is concave. (2) $(\Longrightarrow)$ Suppose that $f(\cdot)$ is strictly concave in S. Then, inequality $(*)$ is strict for $x \neq x^0$. $(\Longleftarrow)$ Setting $z = x^0 + \lambda(x - x^0)$, we have

$$f(x) - f(x^0) < \frac{f(z) - f(x^0)}{\lambda} \leq \frac{\nabla f(x^0) \cdot (z - x^0)}{\lambda} = \nabla f(x^0) \cdot (x - x^0).$$

where we used the inequality in (1), which we have already proved, and the fact that $z - x^0 = \lambda(x - x^0)$. This shows that the inequality in the first property holds with strict inequality. (3) This part is trivial. So, we omit the proof. ■

4.7 Quasiconcave and Quasiconvex Functions

I provide the definition of quasiconcavity and quasiconvexity of functions.

Definition 4.7.1 *A function f, defined over a convex set $S \subseteq \mathbb{R}^n$, is **quasiconcave** if the upper level set $P_\alpha = \{x \in S | f(x) \geq \alpha\}$ is convex for each $\alpha \in \mathbb{R}$. We say that f is **quasiconvex** if $-f$ is quasiconcave. So, f is quasiconvex iff the lower level set $P^\alpha = \{x \in S | f(x) \leq \alpha\}$ is convex for each $\alpha \in \mathbb{R}$.*

There are equivalent definitions of quasiconcavity.

Theorem 4.7.1 *Let $f(\cdot)$ be a function of n variables defined on a convex set S in $\mathbb{R}^n$. Then, f is quasiconcave if and only if either of the following conditions is satisfied for all $x, x' \in S$ and all $\lambda \in [0,1]$,*

(1) $f(\lambda x + (1-\lambda)x') \geq \min\{f(x), f(x')\}$

(2) $f(x') \geq f(x) \Longrightarrow f(\lambda x + (1-\lambda)x') \geq f(x)$

Remark: If I replace $f(\cdot)$ with $-f(\cdot)$, this theorem provides an equivalent characterization of quasiconvexity.

Proof: (1) ($\Longrightarrow$) Suppose that $f(\cdot)$ is quasiconcave. Let $x, x^{'} \in S$ and $\lambda \in [0,1]$, and define $a = \min\{f(x), f(x^{'})\}$. Then,

$$x, x^{'} \in P_a = \{x \in S | f(x) \geq a\}$$

Since P_a is convex by our hypothesis, $\lambda x + (1-\lambda)x^{'} \in P_a$ for any $\lambda \in [0,1]$. This implies that $f(\lambda x + (1-\lambda)x^{'}) \geq a = \min\{f(x), f(x^{'})\}$. ($\Longleftarrow$) Suppose that the first inequality is valid and let a be an arbitrary number. We must show that P_a is convex. Take any arbitrary points $x, x^{'} \in P_a$. Then, $f(x) \geq a$ and $f(x^{'}) \geq a$. Also, for all $\lambda \in (0,1)$, the first inequality implies that

$$f\left(\lambda x + (1-\lambda)x^{'}\right) \geq \min\{f(x), f(x^{'})\}$$

Thus, $\lambda x + (1-\lambda)x^{'} \in P_a$. This proves that P_a is convex. We omit the proof for the second property. ■

The next proposition shows that quasiconcavity is weaker than concavity.

Proposition 4.7.1 *If $f(\cdot)$ is concave, then it is quasiconcave. Similarly, if $f(\cdot)$ is convex, then it is quasiconvex.*

Proof: We omit the proof. ■

Recall that a function $F : \mathbb{R} \to \mathbb{R}$ is said to be *strictly increasing* if $F(x) > F(y)$ whenever $x > y$. Quasiconcavity has the following implication: quasiconcavity preserves ordinal preferences.

Theorem 4.7.2 (Quasiconcavity is preserved under positive monotone transformation) *Let $f(\cdot)$ be defined on a convex set S in $\mathbb{R}^n$ and let F be a function of one variable whose domain includes $f(S)$. If $f(\cdot)$ is quasiconcave (quasiconvex) and F is strictly increasing, then $F(f(\cdot))$ is quasiconcave (quasiconvex).*

Proof: Suppose $f(\cdot)$ is quasiconcave. Using Theorem 4.7.1, we must have

$$f\left(\lambda x + (1-\lambda)x^{'}\right) \geq \min\{f(x), f(x^{'})\}.$$

Since $F(\cdot)$ is strictly increasing,

$$F\left(f\left(\lambda x + (1-\lambda)x^{'}\right)\right) \geq F\left(\min\{f(x), f(x^{'})\}\right) = \min\{F(f(x)), F(f(x^{'}))\}.$$

It follows that $F \circ f$ is quasiconcave. The argument in the quasiconvex case is entirely similar, replacing $\geq$ with $\leq$ and min with max. ■

Example 4.7.1 (When Cobb-Douglas Utility Function is Quasiconcave) *Let* $f(x_1,\dots,x_n) = Ax_1^{a_1}\cdots x_n^{a_n}$, *where* $x_1,\dots,x_n > 0$, $A > 0$, *and* $a_1,\dots,a_n > 0$. *Set* $a = a_1 + \cdots + a_n$. *We claim that* $f(\cdot)$ *is quasiconcave for all* $a_1,\dots,a_n$.

Let $F(x) = \ln x$. *We know that* $F(x) = \ln x$ *is strictly increasing in* x *and it is strictly concave. So, we consider*

$$F(f(x)) = \ln\left(Ax_1^{a_1}\cdots x_n^{a_n}\right) = \ln A + a_1\ln x_1 + \cdots + a_n \ln x_n.$$

Since $\ln A, a_1\ln x_1,\dots,a_n\ln x_n$ *all are concave functions,* $F(f(x))$ *is a sum of concave functions. It is easy to verify that a sum of concave functions is a concave function. It follows from Proposition 4.7.1 that* $F(f(x))$ *is quasiconcave. Since* $F(x) = \ln x$ *is strictly increasing, there exists its inverse function* $F^{-1}(y) = e^y$ *that is also strictly increasing. By Theorem 4.7.2, we conclude that* $F^{-1}(F(f(x))$ *is quasiconcave. Since* $F^{-1}F(f(x)) = f(x)$, *we show that* $f(x) = Ax_1^{a_1}\cdots x_n^{a_n}$ *is quasiconcave.*

Definition 4.7.2 *A function* $f : S \to \mathbb{R}$ *defined on a convex set* $S \subseteq \mathbb{R}^n$ *is said to be* ***strictly quasiconcave*** *if*

$$f(\lambda x + (1-\lambda)x^{'}) > \min\{f(x), f(x^{'})\}$$

for all $x, x^{'} \in S$ *with* $x \neq x^{'}$ *and all* $\lambda \in (0,1)$. *The function* $f(\cdot)$ *is* ***strictly quasiconvex*** *if* $-f(\cdot)$ *is strictly quasiconcave.*

Theorem 4.7.3 (First-Order Characterization of Quasiconcavity) *Let* $f(\cdot)$ *be a* C^1 *function of* n *variables defined on an open convex set* S *in* $\mathbb{R}^n$. *Then* $f(\cdot)$ *is quasiconcave on* S *if and only if for all* $x, x^0 \in S$,

$$f(x) \geq f(x^0) \Longrightarrow \nabla f(x^0)\cdot(x - x^0) = \sum_{i=1}^{n} \frac{\partial f(x^0)}{\partial x_i}(x_i - x_i^0) \geq 0.$$

Proof: ($\Longrightarrow$) Suppose $f(\cdot)$ is quasiconcave. Let $x, x^0 \in S$ and define the function $g(\cdot)$ on $[0,1]$ by

$$g(t) = f\left((1-t)x^0 + tx\right) = f\left(x^0 + t(x - x^0)\right).$$

Then, using the chain rule, we have

$$g^{'}(t) = \nabla f(x^0 + t(x - x^0))\cdot(x - x^0).$$

Suppose $f(x) \geq f(x^0)$. By Theorem 4.7.1, $g(t) \geq g(0)$ for all $t \in [0,1]$. For any $t \in (0,1]$, we have

$$\frac{g(t) - g(0)}{t} \geq 0.$$

Letting $t \to 0$, we obtain

$$\lim_{t\to 0}\frac{g(t) - g(0)}{t} = g^{'}(0) \geq 0.$$

This implies

$$g^{'}(0) = \nabla f(x^0) \cdot (x - x^0) \geq 0$$

($\Longleftarrow$) We will be satisfied with the figure for this part. ■

The content of the above theorem is that for any quasiconcave function $f(\cdot)$ and any pair of points x and x^0 with $f(x) \geq f(x^0)$, the gradient vector $\nabla f(x^0)$ and the vector $(x - x^0)$ must form an acute angle.

4.7.1 Characterizations of Quasiconcavity via Bordered Hessian

Theorem 4.7.4 *Let $S \subseteq \mathbb{R}^2$ be an open, convex set and $f : S \to \mathbb{R}$ be a C^2 function. Define the* ***bordered Hessian determinant***

$$B_2(x,y) = \begin{vmatrix} 0 & f_1^{'}(x,y) & f_2^{'}(x,y) \\ f_1^{'}(x,y) & f_{11}^{''}(x,y) & f_{12}^{''}(x,y) \\ f_2^{'}(x,y) & f_{21}^{''}(x,y) & f_{22}^{''}(x,y) \end{vmatrix}.$$

1. *A necessary condition for f to be* ***quasiconcave*** *in S is that $B_2(x,y) \geq 0$ for all $(x,y) \in S$.*

2. *A sufficient condition for f to be* ***strictly quasiconcave*** *in S is that $f_1^{'}(x,y) \neq 0$ and $B_2(x,y) > 0$ for all $(x,y) \in S$.*

Proof: We omit the proof. ■

We move on to the general case. Define the bordered Hessian determinants

$$B_r(\mathbf{x}) = \begin{vmatrix} 0 & f_1^{'}(\mathbf{x}) & \cdots & f_r^{'}(\mathbf{x}) \\ f_1^{'}(\mathbf{x}) & f_{11}^{''}(\mathbf{x}) & \cdots & f_{1r}^{''}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f_r^{'}(\mathbf{x}) & f_{r1}^{''}(\mathbf{x}) & \cdots & f_{rr}^{''}(\mathbf{x}) \end{vmatrix}.$$

for $r = 1, \ldots, n$.

Theorem 4.7.5 *Let $S \subseteq \mathbb{R}^n$ be an open, convex set and $f : S \to \mathbb{R}$ be a C^2 function. Then,*

1. *A necessary condition for f to be* ***quasiconcave*** *is that $(-1)^r B_r(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in S$ and all $r = 1, \ldots, n$.*

2. *A sufficient condition for f to be* ***strictly quasiconcave*** *is that $(-1)^r B_r(\mathbf{x}) > 0$ for all $\mathbf{x} \in S$ and all $r = 1, \ldots, n$.*

Proof: We omit the proof. ■

Chapter 5

Static Optimization

Static optimization plays an important role in Economics. I first discuss unconstrained optimization as an approach to tackle a class of static optimiation problems in which there are no explicit constrains given.

5.1 Unconstrained Optimization

Extreme Points

Let $f(\cdot)$ be a real-valued function of n variables $x_1, \ldots, x_n$ defined on a set S in $\mathbb{R}^n$. Suppose that the point $x^* = (x_1^*, \ldots, x_n^*)$ belongs to S and that the value of f at x^* is greater than or equal to the values attained by f at all other points $x = (x_1, \ldots, x_n) \in S$. Thus,

$$f(x^*) \geq f(x) \quad \text{for all } x \in S \quad (*)$$

Here x^* is called a (global) *maximal point* for f in S and $f(x^*)$ is called the *maximum value*. If the inequality $(*)$ is strict for all $x \neq x^*$, then x^* is a *strict maximum point* for $f(\cdot)$ in S. I define *(strict) minimum point* and *minimum value* by reversing the inequality sign in $(*)$. As collective names, we use *extreme points* and *extreme values* to indicate both maxima or minima.

Theorem 5.1.1 *Let $f(\cdot)$ be defined on a set S in $\mathbb{R}^n$ and let $x^* = (x_1^*, \ldots, x_n^*)$ be an interior point in S at which $f(\cdot)$ has partial derivatives. A necessary condition for x^* to be an extreme point for f is that x^* is a **stationary point** for $f(\cdot)$ – that is, it satisfies the equations*

$$\nabla f(x) = \mathbf{0} \Longleftrightarrow \frac{\partial f(x)}{\partial x_i} = 0, \quad \textit{for } i = 1, \ldots, n$$

Proof: Suppose, on the contrary, that x^* is a maximum point but not a stationary point for $f(\cdot)$. Then, there is no loss of generality to assume that there exists at least i such that $\partial f(x^*)/\partial x_i > 0$. Define $x^{**} = (x_1^*, \ldots, x_i^* + \varepsilon, \ldots, x_n^*)$. Since x^* is

an interior point in S, one can make sure that $x^{**} \in S$ by choosing $\varepsilon > 0$ sufficiently small. Then,

$$f(x^{**}) \approx f(x^*) + \nabla f(x) \cdot (0, \ldots, 0, \underbrace{\varepsilon}_{i}, 0, \ldots, 0) > f(x^*).$$

However, this contradicts the hypothesis that x^* is a maximum point for $f(\cdot)$. ■

The next theorem clarifies under what conditions, the converse of the previous theorem is established.

Theorem 5.1.2 *Suppose that the function $f(\cdot)$ is defined in a convex set $S \subseteq \mathbb{R}^n$ and let x^* be an interior point of S. Assume that $f(\cdot)$ is C^1 in a ball around x^*.*

1. *If $f(\cdot)$ is concave in S, then x^* is a (global) maximum point for $f(\cdot)$ in S if and only if x^* is a stationary point for $f(\cdot)$.*
2. *$f(\cdot)$ is convex in S, then x^* is a (global) minimum point for $f(\cdot)$ in S if and only if x^* is a stationary point for $f(\cdot)$.*

Proof: We focus on the first part of the theorem. The second part follows once we take into account that $-f$ is concave. ($\Longrightarrow$) This follows from Theorem 5.1.1. ($\Longleftarrow$) Suppose that x^* is a stationary point for $f(\cdot)$ and that $f(\cdot)$ is concave. Recall the inequality in the first-order characterization of concave functions: for any $x \in S$,

$$f(x) - f(x^*) \leq \nabla f(x^*) \cdot (x - x^*) = 0 \ \ (\because \nabla f(x^*) = 0)$$

Thus, we have $f(x) \leq f(x^*)$ for any $x \in S$ as desired. ■

Local Extreme Points

The point x^* is a *local maximum point* of $f(\cdot)$ in S if there exists an $\varepsilon > 0$ such that $f(x) \leq f(x^*)$ for all $x \in B_\varepsilon(x^*) \cap S$. If x^* is the unique local maximum point for $f(\cdot)$, then it is a *strict local maximum point* for $f(\cdot)$ in S. A *(strict) local minimum point* is defined in the obvious way, and it should be clear what I mean by *local maximum and minimum values, local extreme points*, and *local extreme values*. A stationary point x^* of $f(\cdot)$ that is neither a local maximum point nor a local minimum point is called a *saddle point* of $f(\cdot)$.

Before stating the next result, recall the *n leading principal minors* of the Hessian matrix $D^2 f(x)$:

$$\left| D^2_{(k)} f(x) \right| = \begin{vmatrix} f''_{11}(x) & f''_{12}(x) & \cdots & f''_{1k}(x) \\ f''_{21}(x) & f''_{22}(x) & \cdots & f''_{2n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ f''_{k1}(x) & f''_{k2}(x) & \cdots & f''_{kk}(x) \end{vmatrix}, \ \ k = 1, \ldots, n$$

Theorem 5.1.3 (Sufficient Conditions for Local Extreme Points) *Suppose that $f(x) = f(x_1, \ldots, x_n)$ is defined on a set $S \subseteq \mathbb{R}^n$ and that x^* is an interior stationary point. Assume also that $f(\cdot)$ is C^2 in an open ball around x^*. Then,*

1. *$D^2 f(x^*)$ is positive definite $\Longrightarrow$ x^* is a local minimum point.*
2. *$D^2 f(x^*)$ is negative definite $\Longrightarrow$ x^* is a local maximum point.*

Proof: We only focus on the first part of the theorem. We should be able to prove the second part of the proof by replacing $f(\cdot)$ with $-f(\cdot)$. Since each $f_{ij}(x)$ is continuous in x (because $f(\cdot)$ is C^2), the determinant is a continuous function of x. Therefore, if $|D^2_{(k)} f(x^*)| > 0$ for all k, it is possible to find a ball $B_\varepsilon(x^*)$ with $\varepsilon > 0$ small enough that $|D^2_{(k)} f(x)| > 0$ for all $x \in B_\varepsilon(x^*)$ and all $k = 1, \ldots, n$. The corresponding quadratic form is positive definite for all $x \in B_\varepsilon(x^*)$. It follows that $f(\cdot)$ is strictly convex in $B_\varepsilon(x^*)$. Then, the stationary point x^* is a maximum point for f in $B_\varepsilon(x^*)$. Hence, x^* is a local minimum point for $f(\cdot)$. ■

Lemma 5.1.1 *If x^* is an interior stationary point of $f(\cdot)$ such that $|D^2 f(x^*)| \neq 0$ and $D^2 f(x^*)$ is neither positive definite nor negative definite, then x^* is a saddle point.*

Necessary Conditions for Local Extreme Points

To study the behavior of $f(\cdot)$ in an arbitrary fixed vector in $\mathbb{R}^n$ with length 1, so $\|h\| = 1$. The function $g(\cdot)$ describes the behavior of $f(\cdot)$ along the straight line through x^* parallel to the vector $h \in \mathbb{R}^n$.

$$g(t) = f(x^* + th) = f(x_1^* + th_1, \ldots, x_n^* + th_n)$$

I have the following characterization of local extreme points.

Theorem 5.1.4 (Necessary Conditions for Local Extreme Points) *Suppose that $f(x) = f(x_1, \ldots, x_n)$ is defined on a set $S \subseteq \mathbb{R}^n$, and x^* is an interior stationary point in S. Assume that f is C^2 in a ball around x^*. Then,*

1. *x^* is a local minimum point $\Longrightarrow$ $D^2 f(x^*)$ is positive semidefinite.*
2. *x^* is a local maximum point $\Longrightarrow$ $D^2 f(x^*)$ is negative semidefinite.*

Proof: Suppose that x^* is an interior local maximum point for $f(\cdot)$. Then, if $\varepsilon > 0$ is small enough, $B_\varepsilon(x^*) \subseteq S$, and $f(x) \leq f(x^*)$ for all $x \in B_\varepsilon(x^*)$. If $t \in (-\varepsilon, \varepsilon)$, then $x^* + th \in B_\varepsilon(x^*)$ because $\|(x^* + th) - x^*\| = \|th\| = |t| < \varepsilon$. Then, for all $t \in (-\varepsilon, \varepsilon)$, we have

$$f(x^* + th) \leq f(x^*) \Longleftrightarrow g(t) \leq g(0).$$

Thus, the function $g(\cdot)$ has an interior maximum at $t = 0$. Using the chain rule, we obtain

$$\begin{aligned} g^{'}(t) &= \sum_{i=1}^{n} f_i^{'}(x^* + th)h_i = \nabla f(x^* + th) \cdot h \\ g^{''}(t) &= \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}^{''}(x^* + th)h_i h_j = h \cdot D^2 f(x^* + th)h \end{aligned}$$

The condition $g^{''}(0) \leq 0$ yields

$$\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}^{''}(x^* + th)h_i h_j \leq 0 \ \forall h = (h_1, \ldots, h_n) \text{ with } \|h\| = 1$$

This implies that the Hessian matrix $D^2 f(x^*)$ is negative semidefinite. This is equivalent to checking all principal minors. The same argument can be used to establish the necessary condition for x^* to be a local minimum point for $f(\cdot)$. ■

Exercise 5.1.1 *Find the local extreme values and classify the stationary points as maxima, minima, or neither.*

1. $f(x_1, x_2) = 2x_1 - x_1^2 - x_2^2$.
2. $f(x_1, x_2) = x_1^2 + 2x_2^2 - 4x_2$.
3. $f(x_1, x_2) = x_1^3 - x2^2 + 2x_2$.
4. $f(x_1, x_2) = 4x_1 + 2x_2 - x_1^2 + x_1x_2 - x_2^2$.
5. $f(x_1, x_2) = x_1^3 - 6x_1x_2 + x_2^3$

5.2 Optimization with Equality Constraints

I disucss optimization with equality constraints as an approach to tackle a class of static optimization problems in which equality constraints are explicitly given.

Tangent Hyperplane

A set of equality constraints in $\mathbb{R}^n$

$$g(x) = \mathbf{0} \Leftrightarrow \begin{aligned} g^1(x) &= 0 \\ g^2(x) &= 0 \\ \vdots &\quad \vdots \\ g^m(x) &= 0 \end{aligned}$$

defines a subset of $\mathbb{R}^n$ which is best viewed as a hypersurface. If, as I assume in this section, the functions $g^j,\ j = 1, \ldots, m$ belong to C^1, the surface defined by them is said to be *smooth*. I introduce the tangent hyperplane M below:

$$M = \{y \in \mathbb{R}^n | Dg(x^*)y = \mathbf{0}\}$$

Note that the tangent hyperplane is a subspace of $\mathbb{R}^n$.

Definition 5.2.1 *A point x^* satisfying the constraint $g(x^*) = \mathbf{0}$ is said to be a **regular point** of the constraint if the gradient vectors $\nabla g^1(x^*), \ldots, \nabla g^m(x^*)$ are linearly independent. That is, Rank$(Dg(x^*)) = m$.*

Equality Constraints: The Lagrange Problem

A general maximization problem with equality constraints is of the form

$$\max_{x=(x_1,\ldots,x_n)} f(x_1, \ldots, x_n) \ \text{ subject to } \ g^j(x) = 0 \ \forall j = 1, \ldots, m \ (m < n) \ \ (*)$$

Define the *Lagrangian*,

$$\mathcal{L}(x) = f(x) - \lambda_1 g^1(x) - \cdots - \lambda_m g^m(x)$$

where $\lambda_1, \ldots, \lambda_m$ are called *Lagrange multipliers*. The necessary first-order conditions for optimality are then:

$$\nabla \mathcal{L}(x) = \nabla f(x) - \lambda Dg(x) = \mathbf{0} \Longleftrightarrow \frac{\partial \mathcal{L}(x)}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g^j(x)}{\partial x_i} = 0, \ \forall i = 1, \ldots, n \quad (**)$$

Theorem 5.2.1 (N&S Conditions for Extreme Points with Equality Constraints) *The following establishes the necessary and sufficient conditions for the Lagrangian method*

1. *(**Necessity**) Suppose that the functions f and $g^1, \ldots, g^m$ are defined on a set S in $\mathbb{R}^n$ and $x^* = (x_1^*, \ldots, x_n^*)$ is an interior point of S that solves the maximization problem $(*)$. Assume further that f and $g^1, \ldots, g^m$ are C^1 in a ball around x^*, and that x^* is a regular point of the constraint g. Then, there exist unique numbers $\lambda_1, \ldots, \lambda_m$ such that the first-order conditions $(**)$ are valid.*

2. *(**Sufficiency**) If there exist numbers $\lambda_1, \ldots, \lambda_m$ and a feasible x^* which together satisfy the first-order conditions $(**)$, and if the Lagrangian $\mathcal{L}(x)$ is concave in x, then x^* solves the maximization problem $(*)$.*

Proof: (Necessity) We rather provide a heuristic argument based on the simplest formulation. Consider

$$\max_{(x,y)\in\mathbb{R}^2} f(x, y) \ \text{ subject to } \ g(x, y) = c.$$

Let (x^*, y^*) be a local maximum point of f to the above constrained optimization problem. So, we must have $g(x^*, y^*) = c$.

We consider a pair of "small" numbers $(\Delta x, \Delta y) \in \mathbb{R}^2$ such that $g(x^* + \Delta x, y^* + \Delta y) = g(x^*, y^*)$. Then, we have

$$\begin{array}{rcl} \Delta g & = & g(x^* + \Delta x, y^* + \Delta y) - g(x^*, y^*) \\ & \underbrace{\approx}_{\text{linear approx}} & g_1'(x^*, y^*)\Delta x + g_2'(x^*, y^*)\Delta y = 0. \end{array}$$

Assuming that $g_1'(x^*, y^*) \neq 0$ (i.e., $\text{rank}(Dg(x^*, y^*)) = m$), we derive

$$\Delta x = -\frac{g_2'(x^*, y^*)}{g_1'(x^*, y^*)}\Delta y. \qquad (*)$$

Since (x^*, y^*) is a local maximum point to the constrained optimization problem,

$$\begin{array}{rcl} 0 & \geq & f(x^* + \Delta x, y^* + \Delta y) - f(x^*, y^*) \\ & \underbrace{\approx}_{\text{linear approx}} & f_1'(x^*, y^*)\Delta x + f_2'(x^*, y^*)\Delta y \\ & = & \left(-\frac{f_1'(x^*, y^*)}{g_1'(x^*, y^*)} g_2'(x^*, y^*) + f_2'(x^*, y^*) \right) \Delta y. \quad (\because\ (*)) \end{array}$$

Since Δy could be positive or negative, we must have

$$-\frac{f_1'(x^*, y^*)}{g_1'(x^*, y^*)} g_2'(x^*, y^*) + f_2'(x^*, y^*) = 0. \qquad (**)$$

Define

$$\lambda^* \equiv \frac{f_1'(x^*, y^*)}{g_1'(x^*, y^*)}.$$

Then, $(**)$ can be translated into:

$$\begin{array}{rcl} f_1'(x^*, y^*) & = & \lambda^* g_1'(x^*, y^*), \\ f_2'(x^*, y^*) & = & \lambda^* g_2'(x^*, y^*). \end{array}$$

This is exactly the first-order condition of the Lagrangian. If $g_1'(x^*, y^*) = 0$ but $g_2'(x^*, y^*) \neq 0$, we can obtain the same condition. The problem occurs when $g_1'(x^*, y^*) = g_2'(x^*, y^*) = 0$, which is explicitly excluded by the assumption that $rank(Dg(x^*)) = m$.

(Sufficiency) Suppose that the Lagrangian $\mathcal{L}(x)$ is concave. The first-order necessary conditions imply that the Lagrangian is stationary at x^*. Then by Theorem 5.1.2 (sufficiency for unconstrained optimization),

$$\mathcal{L}(x^*) = f(x^*) - \sum_{j=1}^{m} \lambda_j g^j(x^*) \geq f(x) - \sum_{j=1}^{m} \lambda_j g^j(x) = \mathcal{L}(x) \quad \forall x \in S$$

But for all feasible x, we have $g^j(x) = 0$ and of course, $g^j(x^*) = 0$ for all $j = 1, \ldots, m$. This implies that $f(x^*) \geq f(x)$. Thus, x^* solves the maximization problem $(*)$. ■

Lagrange Multipliers as Shadow Prices

The optimal values of $x_1, \ldots, x_n$ in the maximization problem $(*)$ will depend upon the parameter vector $r = (r_1, \ldots, r_k) \in \mathbb{R}^k$, in general. If $x^*(r) = (x_1^*(r), \ldots, x_n^*(r))$ denotes the vector of optimal values of the choice variables, then the corresponding value

$$f^*(r) = f(x_1^*(r), \ldots, x_n^*(r))$$

of $f(\cdot)$ is called the *(optimal) value function* for the maximization problem $(*)$. The values of the Lagrange multipliers will also depend on r; we write $\lambda_j = \lambda_j(r)$ for $j = 1, \ldots, m$. Let $\mathcal{L}(x, r) = f(x, r) - \sum_{j=1}^m \lambda_j g^j(x, r)$ be the Lagrangian. Under certain conditions, I have

$$\frac{\partial f^*(r)}{\partial r_i} = \left(\frac{\partial \mathcal{L}(x, r)}{\partial r_i} \right)_{x = x^*(r)} \qquad \forall i = 1, \ldots, k$$

First-Order Necessary Conditions for Local Extreme Points

Lemma 5.2.1 *Let x^* be a regular point of the constraint $g(x) = \mathbf{0}$ and a local extreme point of $f(\cdot)$ subject to these constraints. Then, for any $y \in \mathbb{R}^n$,*

$$Dg(x^*)y = \mathbf{0} \Rightarrow \nabla f(x^*) \cdot y = 0.$$

Remark: This lemma says that $\nabla f(x^*)$ is orthogonal to the tangent hyperplane.

Proof: Let $y = (y_1, \ldots, y_n)$ with $\|y\| = 1$. Let $x(t) = x^* + ty$ be any smooth curve on the constraint surface $g(x(t)) = \mathbf{0}$ passing through x^* with derivative $x'(t)|_{t=0} = y$ at $x(0) = x^*$. There exists some $\varepsilon > 0$ such that $g(x(t)) = \mathbf{0}$ for any $t \in (-\varepsilon, \varepsilon)$.

Since x^* is a regular point, the tangent hyperplane is identical with the set of y's satisfying $\nabla g(x^*)y = \mathbf{0}$. Then, since x^* is a constrained local extreme point of $f(\cdot)$, we have

$$\frac{d}{dt} f(x(t)) \bigg|_{t=0} = 0 \Longrightarrow \nabla f(x^*) x'(0) = 0,$$

equivalently, $\nabla f(x^*)y = 0$. ■

Second-Order Necessary and Sufficient Conditions for Local Extreme Points

Theorem 5.2.2 (Necessity for Local Maximum) *Suppose that x^* is a local maximum of $f(\cdot)$ subject to $g(x) = \mathbf{0}$ and that x^* is a regular point of the constraint $g(x) = \mathbf{0}$. Then, there is a $\lambda \in \mathbb{R}^n$ such that*

$$\nabla f(x^*) - [Dg(x^*)]^T \lambda = \mathbf{0} \Leftrightarrow \frac{\partial f(x^*)}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g^j(x^*)}{\partial x_i} = 0 \ \ \forall i = 1, \ldots, n.$$

If we denote by M the tangent hyperplane $M = \{h \in \mathbb{R}^n | Dg(x^*)h = \mathbf{0}\}$, then, the matrix

$$D^2\mathcal{L}(x^*) = D^2 f(x^*) - \lambda D^2 g(x^*)$$

is negative semidefinite on M, that is,

$$h^T D^2\mathcal{L}(x^*)h \leq 0 \quad \forall h \in M$$

Proof: The first part follows from Theorem 5.2.1. We only focus on the second part. Let $h = (h_1, \ldots, h_n) \in M$ with $\|h\| = 1$. Let $x(t) = x^* + th$ be any smooth curve on the constant surface $g(x(t)) = \mathbf{0}$ passing through x^* with derivative $x^{'}(0) = h$ at $x(0) = x^*$. Suppose that x^* is an interior local maximum point for f subject to $g(x) = \mathbf{0}$. Then, if $\varepsilon > 0$ is small enough,

$$\mathcal{L}(x^* + th) \leq \mathcal{L}(x^*) \Longleftrightarrow f(x^* + th) - \lambda g(x^* + th) \leq f(x^*) - \lambda g(x^*)$$

for all $t \in (-\varepsilon, \varepsilon)$ because $\|(x^* + th) - x^*\| = \|th\| = |t| < \varepsilon$. Define the function $\varphi(t) = \mathcal{L}((x^* + th)$. Then, for all $t \in (-\varepsilon, \varepsilon)$, we have

$$\mathcal{L}(x^* + th) \leq \mathcal{L}(x^*) \Longleftrightarrow \varphi(t) \leq \varphi(0).$$

Thus, the function φ has an interior maximum at $t = 0$. Using the chain rule, we obtain

$$\begin{aligned} \varphi^{'}(t) &= \nabla\mathcal{L}(x^* + th)h = \nabla f(x^* + th)h - \lambda Dg(x^* + th)h \\ \varphi^{'}(0) &= \nabla\mathcal{L}(x^*)h = \nabla f(x^*)h - \lambda Dg(x^*)h = 0 \end{aligned}$$

because $h \in M$ so that $\nabla f(x^*)h = 0$ and $Dg(x^*)h = \mathbf{0}$. Furthermore,

$$\varphi^{''}(t) = h^T D^2\mathcal{L}(x^* + th)h$$

The hypothesis that φ has an interior local maximum at $t = 0$ means $\varphi^{''}(0) \leq 0$. Thus,

$$h^T D^2\mathcal{L}(x^*)h \leq 0 \Longleftrightarrow \sum_{i=1}^{n} \mathcal{L}_{ij}(x^*)h_i h_j \leq 0$$

This implies that the Hessian matrix $D^2\mathcal{L}(x^*)$ is negative semidefinite on M. ■

Theorem 5.2.3 (Sufficiency for Local Maximum) *Suppose there is a point $x^* \in \mathbb{R}^n$ satisfying $g(x^*) = \mathbf{0}$, and a $\lambda \in \mathbb{R}^m$ such that*

$$\nabla f(x^*) - [Dg(x^*)]^T\lambda = \underbrace{\mathbf{0}}_{n\times 1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Suppose also that the matrix $D^2\mathcal{L}(x^*) = D^2 f(x^*) + \lambda D^2 g(x^*)$ is negative definite on $M = \{y \in \mathbb{R}^n | Dg(x^*)y = \mathbf{0}\}$, that is, for $y \in M$ with $y \neq \mathbf{0}$, $y^T D^2\mathcal{L}(x^*)y < 0$. Then, x^* is a strict local maximum of $f(\cdot)$ subject to $g(x) = \mathbf{0}$.

Proof: The first part follows from Theorem 5.2.2. Define the Lagrangian as follows.

$$\mathcal{L}(x) = f(x) - \lambda g(x)$$

Differentiating this with respect to x, and evaluating it at x^*, we obtain

$$\nabla \mathcal{L}(x^*) = \nabla f(x^*) - \lambda Dg(x^*) = \mathbf{0}$$

This implies that $\nabla \mathcal{L}(x^*)y = 0$ for any $y \in \mathbb{R}^n$. By our hypothesis, $D^2\mathcal{L}(x^*)$ is negative definite on M, and therefore, x^* is a local maximum point of $\mathcal{L}(x)$ from Theorem 5.1.3. This implies that x^* is a local maximum of $f(\cdot)$ subject to $g(x) = \mathbf{0}$. ■

In general, I define the following *bordered Hessian* determinants, for $r = m + 1, \ldots, n$:

$$B_r(x^*) = \begin{vmatrix} 0 & \cdots & 0 & \frac{\partial g^1(x^*)}{\partial x_1} & \cdots & \frac{\partial g^1(x^*)}{\partial x_r} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{\partial g^m(x^*)}{\partial x_1} & \cdots & \frac{\partial g^m(x^*)}{\partial x_r} \\ \frac{\partial g^1(x^*)}{\partial x_1} & \cdots & \frac{\partial g^m(x^*)}{\partial x_1} & \mathcal{L}''_{11}(x^*) & \cdots & \mathcal{L}''_{1r}(x^*) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g^1(x^*)}{\partial x_r} & \cdots & \frac{\partial g^m(x^*)}{\partial x_r} & \mathcal{L}''_{r1}(x^*) & \cdots & \mathcal{L}''_{rr}(x^*) \end{vmatrix}$$

The determinant B_r is the $(m+r)$th leading principal minor of $(m+n)\times(m+n)$ bordered matrix

$$\begin{pmatrix} \mathbf{0}_{m\times m} & \underbrace{Dg(x^*)}_{m\times n} \\ \underbrace{(Dg(x^*))^T}_{n\times m} & \underbrace{D^2\mathcal{L}(x^*)}_{n\times n} \end{pmatrix} = \begin{pmatrix} 0 & \cdots & 0 & \frac{\partial g^1(x^*)}{\partial x_1} & \cdots & \frac{\partial g^1(x^*)}{\partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{\partial g^m(x^*)}{\partial x_1} & \cdots & \frac{\partial g^m(x^*)}{\partial x_n} \\ \frac{\partial g^1(x^*)}{\partial x_1} & \cdots & \frac{\partial g^m(x^*)}{\partial x_1} & \mathcal{L}''_{11}(x^*) & \cdots & \mathcal{L}''_{1n}(x^*) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g^1(x^*)}{\partial x_n} & \cdots & \frac{\partial g^m(x^*)}{\partial x_n} & \mathcal{L}''_{n1}(x^*) & \cdots & \mathcal{L}''_{nn}(x^*) \end{pmatrix}.$$

Theorem 5.2.4 (Sufficiency for Local Maximum in terms of Bordered Hessian) *Suppose there is a point $x^* \in \mathbb{R}^n$ satisfying $g(x^*) = \mathbf{0}$, and a $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ such that*

$$\nabla f(x^*) - \sum_{j=1}^{m} \lambda_j \nabla g^j(x^*) = \underbrace{\mathbf{0}}_{n\times 1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

$(-1)^r B_r(x^*) > 0$ for $r = m+1, \ldots, n \Rightarrow x^*$ is the local maximum to $(*)$.

$(-1)^m B_r(x^*) > 0$ for $r = m+1, \ldots, n \Rightarrow x^*$ is the local minimum to $(*)$.

Proof: We skip the proof. ■

Example 5.2.1

$$\max_{x,y,z} f(x,y,z) = x^2+y^2+z^2 \ \textit{ subject to } \ \begin{cases} g^1(x,y,z) = x+2y+z = 30 \\ g^2(x,y,z) = 2x-y-3z = 10 \end{cases}$$

We setup the Lagrangian:

$$\mathcal{L}(x,y,z) = x^2+y^2+z^2 - \lambda_1(x+2y+z-30) - \lambda_2(2x-y-3z-10).$$

The first-order conditions are:

$$\begin{aligned} \mathcal{L}_1^{'} &= 2x-\lambda_1-2\lambda_2 = 0 \\ \mathcal{L}_2^{'} &= 2y-2\lambda_1+\lambda_2 = 0 \\ \mathcal{L}_3^{'} &= 2z-\lambda_1+3\lambda_2 = 0. \end{aligned}$$

Considering two equality constraints, the unique solution to the first-order conditions is $(x,y,z) = (10,10,0)$. *The associated Lagrange multipliers are* $\lambda_1 = 12$ *and* $\lambda_2 = 4$.

We compute the Bordered Hessian determinant B_3*:*

$$B_3(x,y,z) = \begin{vmatrix} 0 & 0 & \partial g^1/\partial x & \partial g^1/\partial y & \partial g^1/\partial z \\ 0 & 0 & \partial g^2/\partial x & \partial g^2/\partial y & \partial g^2/\partial z \\ \partial g^1/\partial x & \partial g^2/\partial x & \mathcal{L}''_{xx} & \mathcal{L}''_{xy} & \mathcal{L}''_{xz} \\ \partial g^1/\partial y & \partial g^2/\partial y & \mathcal{L}''_{yx} & \mathcal{L}''_{yy} & \mathcal{L}''_{yz} \\ \partial g^1/\partial z & \partial g^2/\partial z & \mathcal{L}''_{zx} & \mathcal{L}''_{zy} & \mathcal{L}''_{zz} \end{vmatrix}$$

$$B_3(10,10,0) = \begin{vmatrix} 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 2 & -1 & -3 \\ 1 & 2 & 2 & 0 & 0 \\ 2 & -1 & 0 & 2 & 0 \\ 1 & -3 & 0 & 0 & 2 \end{vmatrix} = 150 > 0$$

This implies $(-1)^2 B_3(10,10,0) > 0$. *Hence,* $(10,10,0)$ *is the local minimum point.*

Exercise 5.2.1 *Solve the problem*

$$\max\{x+4y+z\} \ \textit{ subject to } \ x^2+y^2+z^2 = 216 \ \textit{ and } \ x+2y+3z = 0$$

Exercise 5.2.2 *Consider the problem (assuming* $m \geq 4$*).*

$$\max U(x_1,x_2) = \frac{1}{2}\ln(1+x_1) + \frac{1}{4}\ln(1+x_2) \ \textit{ subject to } \ 2x_1+3x_2 = m$$

Answer the following questions.

1. Let $x_1^*(m)$ and $x_2^*(m)$ denote the values of x_1 and x_2 that solve the above maximization problem. Find these functions and the corresponding Lagrangian multiplier.

2. The optimal value U^* of $U(x_1,x_2)$ is a function of m. Find an explicit expression for $U^*(m)$, and show that $dU^*/dm = \lambda$.

5.3 Optimization with Inequality Constraints

I generalize static optimization with equality constraints by incorporating inequality constraints. In particular, we can treat an equality constraint $g(x) = 0$ as the set of the following two inequality constraints: $g(x) \leq 0$ and $g(x) \geq 0$. I consider the following problem:

$$\max_{x \in S} f(x) \text{ subject to } \begin{cases} g^1(x_1, \ldots, x_n) \leq 0 \\ g^2(x_1, \ldots, x_n) \leq 0 \\ \vdots \\ g^m(x_1, \ldots, x_n) \leq 0 \end{cases}$$

A vector $x = (x_1, \ldots, x_n)$ that satisfies all the constraints is called *feasible*. The set of all feasible vectors is said to be the *feasible set*. We assume that $f(\cdot)$ and all the g^j functions are C^1. In the case of equality constraint, the number of constraints were assumed to be strictly less than the number of variables. This is not necessary for the case of inequality constraints. An inequality constraint $g^j(x) \leq 0$ is said to be *active (binding)* at x if $g^j(x) = 0$ and *inactive (non-binding)* at x if $g^j(x) < 0$.

Note that minimizing $f(x)$ is equivalent to maximizing $-f(x)$. Moreover, an inequality constraint of the form $g^j(x) \geq 0$ can be rewritten as $-g^j(x) \leq 0$. In this way, most constrained optimization problem can be expressed as the above form.

We define the Lagrangian exactly as before.

$$\mathcal{L}(x) = f(x) - \lambda \cdot g(x) = f(x) - \sum_{j=1}^{m} \lambda_j g^j(x),$$

where $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ are the Lagrangian multipliers. Again the first-order partial derivatives of the Lagrangian are equated to 0:

$$\nabla \mathcal{L}(x) = \nabla f(x) - \lambda Dg(x) = \mathbf{0} \iff \frac{\partial \mathcal{L}(x)}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g^j(x)}{\partial x_i} = 0, \ \forall i = 1, \ldots, n \quad (*)$$

In addition, we introduce the *complementary slackness conditions*. For all $j = 1, \ldots, m$,

$$\lambda_j \geq 0 \text{ and } \lambda_j = 0 \text{ if } g^j(x) < 0 \quad (**)$$

An alternative formulation of this condition is that for any $j = 1, \ldots, m$,

$$\lambda_j \geq 0 \text{ and } \lambda_j g^j(x) = 0$$

In particular, if $\lambda_j > 0$, we must have $g^j(x) = 0$. However, it is perfectly possible to have both $\lambda_j = 0$ and $g^j(x) = 0$.

Conditions $(*)$ and $(**)$ are often called the *Kuhn-Tucker conditions*. They are (essentially but not quite) *necessary* conditions for a feasible vector to solve the maximization problem. In general, they are definitely not sufficient on their own. Suppose one can find a point x^* at which $f(\cdot)$ is stationary and $g^j(x^*) < 0$ for all $j = 1, \ldots, m$. Then, the Kuhn-Tucker conditions will automatically be satisfied by x^* together with all the Lagrangian multipliers $\lambda_j = 0$ for all $j = 1, \ldots, m$.

Theorem 5.3.1 (Sufficiency of the Kuhn-Tucker Conditions I) *Consider the maximization problem and suppose that x^* is feasible and satisfies conditions $(*)$ and $(**)$. If the Lagrangian $\mathcal{L}(x) = f(x) - \lambda \cdot g(x)$ (with the λ values obtained from the recipe) is concave, then x^* is optimal.*

Proof: Since $\mathcal{L}(x)$ is concave by assumption and $\nabla\mathcal{L}(x^*) = \mathbf{0}$ from $(*)$, by Theorem 5.2.1, x^* is a global maximum point of $\mathcal{L}(x)$. Hence, for all $x \in S$,

$$f(x^*) - \sum_{j=1}^{m} \lambda_j g^j(x^*) \geq f(x) - \sum_{j=1}^{m} \lambda_j g_j(x)$$

Rearranging gives the equivalent inequality

$$f(x^*) - f(x) \geq \sum_{j=1}^{m} \lambda_j \left(g^j(x^*) - g^j(x)\right).$$

Thus, it suffices to show

$$\sum_{j=1}^{m} \lambda_j \left(g^j(x^*) - g^j(x)\right) \geq 0$$

for all feasible x, because this will imply that x^* solves the maximization problem. Suppose that $g^j(x^*) < 0$. Then $(**)$ shows that $\lambda_j = 0$. Suppose that $g^j(x^*) = 0$, we have $\lambda_j(g^j(x^*) - g^j(x)) = -\lambda_j g^j(x) \geq 0$ because x is feasible, i.e., $g^j(x) \leq 0$ and $\lambda_j \geq 0$. Then, we have $\sum_{j=1}^{m} \lambda_j \left(g^j(x^*) - g^j(x)\right) \geq 0$ as desired. ■

Theorem 5.3.2 (Sufficiency for the Kuhn-Tucker Conditions II) *Consider the maximization problem and suppose that x^* is feasible and satisfies conditions $(*)$ and $(**)$. If $f(\cdot)$ is concave and each $\lambda_j g^j(x)$ (with the λ values obtained from the recipe) is quasiconvex, then x^* is optimal.*

Proof: We want to show that $f(x) - f(x^*) \leq 0$ for all feasible x. Since $f(\cdot)$ is concave, then according to Theorem 4.6.3 (First-order characterization of concavity of $f(\cdot)$),

$$f(x) - f(x^*) \leq \nabla f(x^*) \cdot (x - x^*) \underbrace{=}_{(*)} \sum_{j=1}^{m} \lambda_j \nabla g^j(x^*) \cdot (x - x^*)$$

where we use the first order condition $(*)$. It therefore suffices to show that for all $j = 1, \ldots, m$, and all feasible x,

$$\lambda_j \nabla g^j(x^*) \cdot (x - x^*) \leq 0.$$

The above inequality is satisfied for those j such that $g^j(x^*) < 0$, because then $\lambda_j = 0$ from the complementary slackness condition $(**)$. For those j such that $g^j(x^*) = 0$, we have $g^j(x) \leq g^j(x^*)$ (because x is feasible), and hence $-\lambda_j g^j(x) \geq -\lambda_j g^j(x^*)$ because $\lambda_j \geq 0$. Since the function $-\lambda_j g^j(x)$ is quasiconcave (because $\lambda_j g^j(x)$ is quasiconvex), it follows from Theorem 4.7.3 (a characterization of quasiconcavity) that $\nabla(-\lambda_j g^j(x^*)) \cdot (x - x^*) \geq 0$, and thus, $\lambda_j \nabla g^j(x^*) \cdot (x - x^*) \leq 0$. ■

Exercise 5.3.1 *Reformulate the problem*

$$\min_{(x,y)\in\mathbb{R}^2} 4\ln(x^2 + 2) + y^2 \ \text{ subject to } \ x^2 + y \geq 2, \ x \geq 1$$

as a standard Kuhn-Tucker maximization problem and write down the necessary Kuhn-Tucker conditions. Moreover, find the solution of the problem (Take it for granted that there is a solution).

5.3.1 Constraint Qualifications

Consider the maximization problem below.

$$\max f(x) \ \text{ subject to } \ g^j(x) \leq 0, \ j = 1, \ldots, m$$

Definition 5.3.1 *The constrained maximization problem satisfies the* ***constraint qualification*** *if the gradient vectors $\nabla g^j(x^*)$ $(1 \leq j \leq m)$ corresponding those constraints that are active (binding) at x^*, are linearly independent.*

An alternative formulation of this condition is: Delete all rows in the Jacobian matrix $Dg(x^*)$ that correspond to constraints that are inactive (not binding) at x^*. Then, the remaining matrix should have rank equal to the number of rows.

Theorem 5.3.3 (Kuhn-Tucker Necessary Conditions) *Suppose that $x^* = (x_1^*, \ldots, x_n^*)$ solves the constrained maximization problem where $f(\cdot)$ and $g^1(\cdot), \ldots, g^m(\cdot)$ are C^1 functions. Suppose furthermore that the maximization problem satisfies the constraint qualification. Then, there exist unique numbers $\lambda_1, \ldots, \lambda_m$ such that the Kuhn-Tucker conditions $(*)$ and $(**)$ hold at $x = x^*$.*

Proof: Before we provide the formal proof, we first provide a heuristic argument based on the simplest formulation. Introducing *slack variable* s, we translate the original problem into the one with an equality constraint:

$$\max_{x,y} f(x, y) \ \text{ subject to } \ g(x, y) + s = c, \ s \geq 0.$$

Let (x^*, y^*) be the solution to the original question and s^* be the slack variable associated with (x^*, y^*). We consider a small perturbation $(\Delta x, \Delta y, \Delta s)$ such that $g(x^* + \Delta x, y^* + \Delta y) + s^* + \Delta s = c$.

Case 1: $g(x^*, y^*) < c \Leftrightarrow s^* > 0 \Leftrightarrow$ **the constraint is not binding**

Since $s^* > 0$, g is continuous, and $(\Delta x, \Delta y, \Delta s)$ is very small, we have $s^* + \Delta s > 0$ so that $g(x^* + \Delta x, y^* + \Delta y) < c$. This is the unconstrained optimization problem! Since (x^*, y^*) is the solution to the problem, we have

$$\begin{array}{ccc} 0 & \geq & f(x^* + \Delta x, y^* + \Delta y) - f(x^*, y^*) \\ & \underbrace{\approx}_{\text{linear approx}} & f_x'(x^*, y^*)\Delta x + f_y'(x^*, y^*)\Delta y \end{array}$$

Since Δx and Δy can chosen independently and can be positive or negative, we have

$$f_x'(x^*, y^*) = 0 \quad \text{and} \quad f_y'(x^*, y^*) = 0$$

Define $f^*(c) = f(x^*, y^*)$. Then, we have the following standard interpretation of λ:

$$\frac{df^*(c)}{dc} = \lambda.$$

Given this interpretation of λ, it makes sense to set $\lambda = 0$ in the Lagrangian in the case where $s^* > 0$. This is because if $\lambda < 0$, we can find a point (x, y) such that $g(x, y) < c$ and $f(x, y) > f(x^*, y^*)$, which contradicts the hypothesis that (x^*, y^*) is the solution. Thus,

$$\mathcal{L}_x'(x^*, y^*) = f_x'(x^*, y^*) \quad \text{and} \quad \mathcal{L}_y'(x^*, y^*) = f_y'(x^*, y^*).$$

Case 2: $g(x^*, y^*) = c \Leftrightarrow s^* = 0 \Leftrightarrow$ **the constraint is binding**

By construction of the perturbation $(\Delta x, \Delta y, \Delta s)$, we have

$$\begin{array}{ccc} 0 & = & \underbrace{[g(x^* + \Delta x, y^* + \Delta y) + s^* + \Delta s]}_{=c} - \underbrace{[g(x^*, y^*) + s^*]}_{=c} \\ & \underbrace{\approx}_{\text{linear approx}} & g_x'(x^*, y^*)\Delta x + g_y'(x^*, y^*)\Delta y + \Delta s. \end{array}$$

Assume $g_x'(x^*, y^*) \neq 0$. Then,

$$\Delta x = -\frac{g_y'(x^*, y^*)}{g_x'(x^*, y^*)}\Delta y - \frac{1}{g_x'(x^*, y^*)}\Delta s.$$

Since (x^*, y^*) is the solution to the problem,

$$\begin{array}{ccc} 0 & \geq & f(x^* + \Delta x, y^* + \Delta y) - f(x^*, y^*) \\ & \underbrace{\approx} & f_x^{'}(x^*, y^*)\Delta x + f_y^{'}(x^*, y^*)\Delta y \\ & \text{linear approx} & \\ & = & \left[-\frac{f_x^{'}(x^*, y^*)}{g_x^{'}(x^*, y^*)} g_y^{'}(x^*, y^*) + f_y^{'}(x^*, y^*) \right] \Delta y - \frac{f_x^{'}(x^*, y^*)}{g_x^{'}(x^*, y^*)} \Delta s. \end{array}$$

Assume that $\Delta s = 0$ and $\Delta y \neq 0$. Since Δy could be positive or negative, we must have

$$-\frac{f_x^{'}(x^*, y^*)}{g_x^{'}(x^*, y^*)} g_y^{'}(x^*, y^*) + f_y^{'}(x^*, y^*) = 0.$$

Setting

$$\lambda^* = \frac{f_x^{'}(x^*, y^*)}{g_x^{'}(x^*, y^*)},$$

we obtain

$$\begin{array}{rcl} \mathcal{L}_x^{'}(x^*, y^*) & = & f_x^{'}(x^*, y^*) - \lambda^* g_x^{'}(x^*, y^*) = 0, \\ \mathcal{L}_y^{'}(x^*, y^*) & = & f_y^{'}(x^*, y^*) - \lambda^* g_y^{'}(x^*, y^*) = 0. \end{array}$$

Assume $\Delta s \neq 0$ and $\Delta y = 0$. In this case, we must have

$$0 \geq -\frac{f_x^{'}(x^*, y^*)}{g_x^{'}(x^*, y^*)} \Delta s$$

Since $s^* = 0$, we have $\Delta s > 0$. Hence, the above inequality implies

$$\frac{f_x^{'}(x^*, y^*)}{g_x^{'}(x^*, y^*)} = \lambda^* \geq 0.$$

We obtain the same condition if we assume $g_y^{'}(x^*, y^*) \neq 0$. Combining the conclusions of Cases 1 and 2, we obtain

- $\mathcal{L}_x^{'}(x^*, y^*) = f_x^{'}(x^*, y^*) - \lambda^* g_x^{'}(x^*, y^*) = 0,$
- $\mathcal{L}_y^{'}(x^*, y^*) = f_y^{'}(x^*, y^*) - \lambda^* g_y^{'}(x^*, y^*) = 0,$
- $\lambda^* \geq 0,\ s^* \geq 0$, and $\lambda^* s^* = 0$.

This completes the heuristic argument.

Next, we move on to the formal proof. We assume the following.

1. $x^* \in \mathbb{R}^n$ maximizes f on the constraint set $g^j(x) \leq 0$ for all $j = 1, \ldots, m$
2. only $g^1, \ldots, g^k$ are binding at x^*, where $k \leq m$.

3. the $k \times n$ Jacobian matrix $Dg_k(x^*)$ has maximal rank k. That is,

$$k = \text{Rank}\,[Dg_k(x^*)] = \text{Rank}\begin{pmatrix} \partial g^1(x^*)/\partial x_1 & \cdots & \partial g^1(x^*)/\partial x_n \\ \vdots & \ddots & \vdots \\ \partial g^k(x^*)/\partial x_1 & \cdots & \partial g^k(x^*)/\partial x_n \end{pmatrix}.$$

The proof consists of two steps.

Step 1: $\nabla\mathcal{L}(x,\lambda) = 0$ and $\lambda g(x) = \mathbf{0}$

Since each $g^j(\cdot)$ is a continuous function, there is a open ball $B_\varepsilon(x^*)$ such that $g^j(x) < 0$ for all $x \in B_\varepsilon(x^*)$ and for $j = k+1, \ldots, m$. We will work in the open ball $B_\varepsilon(x^*)$ for the rest of proof.

Note that x^* maximizes $f(\cdot)$ in $B_\varepsilon(x^*)$ over the constraint set that $g^j(x) = 0$ for $j = 1, \ldots, k$. By assumption, Theorem 5.2.1 (Necessity for Optimization with Equality Constraints) applies and therefore, there exist $\mu_1^*, \ldots, \mu_k^*$ such that

$$\nabla\hat{\mathcal{L}}(x^*, \mu^*) = \mathbf{0} \quad \text{and} \quad g^j(x^*) = 0 \ \forall j = 1, \ldots, k$$

where $\hat{\mathcal{L}}(x,\mu) \equiv f(x) - \sum_{j=1}^k \mu_j g^j(x)$ as the restricted Lagrangian.

Consider the usual Lagrangian

$$\mathcal{L}(x, \lambda_1, \ldots, \lambda_m) \equiv f(x) - \sum_{j=1}^m \lambda_k g^j(x).$$

Let $\lambda_i^* = \mu_i^*$ for $i = 1, \ldots, k$ and $\lambda_i^* = 0$ for $j = k+1, \ldots, m$. Then, we see that (x^*, λ^*) is a solution of the $n+m$ equations in $n+m$ unknowns:

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial x_i}(x^*, \lambda^*) &= 0 \ \forall i = 1, \ldots, n \\ \lambda_j^* g^j(x^*) &= 0 \ \forall j = 1, \ldots, m \end{aligned}$$

Step 2: $\lambda_j \geq 0$ for all j

There is a C^1 curve $x(t)$ defined for $t \in [0, \varepsilon)$ such that $x(0) = x^*$ and, for all $t \in [0, \varepsilon)$,

$$g^1(x(t)) = -t \ \text{ and } \ g^j(x(t)) = 0 \ \text{ for } j = 2, \ldots, k$$

By the implicit function theorem, we can still solve the constrained optimization problem in $B_\varepsilon(x^*)$ even if we slightly perturb the constraint set. Let $h = x'(0)$. Using the chain rule, we conclude that

$$\nabla g^1(x^*)h = -1, \ \ \nabla g^j(x^*)h = 0 \ \ \forall j = 2, \ldots, k$$

Since $x(t)$ lies in the constraint set for all t and x^* maximizes $f(\cdot)$ in the constraint set, $f(\cdot)$ must be *nonincreasing* along $x(t)$. Therefore,

$$\frac{d}{dt}f(x(t))\bigg|_{t=0} = \nabla f(x^*)h \leq 0$$

By our first-order conditions, we execute a series of computations.

$$\begin{aligned}
\mathbf{0} &= \nabla \mathcal{L}(x^*)y \\
&= \nabla f(x^*)h - \sum_{j=1}^{k} \lambda_j \nabla g^j(x^*)y \\
&= \nabla f(x^*)h - \lambda_1 \nabla g^1(x^*)h \\
&= \nabla f(x^*)h + \lambda_1
\end{aligned}$$

Since $\nabla f(x^*)h \leq 0$, we conclude that $\lambda_1 \geq 0$. A similar argument shows that $\lambda_j \geq 0$ for $j = 1, \ldots, k$. This completes the proof. ■

Theorem 5.3.4 (Kuhn-Tucker N & S Conditions) *Assume that a feasible vector x^* and a set of multipliers $\lambda_1, \ldots, \lambda_m$ satisfy the Kuhn-Tucker necessary conditions $(*)$ and $(**)$ for the constrained maximization problem. Define $J = \{j | g^j(x^*) = 0\}$, the set of active (binding) constraints, and assume that $\lambda_j > 0$ for all $j \in J$ Consider the Lagrangian problem*

$$\max f(x) \ \textit{subject to} \ g^j(x) = 0 \ \forall j \in J$$

Then, x^* satisfies

$$\nabla \tilde{\mathcal{L}}(x^*) = \nabla f(x^*) - \sum_{j \in J} \lambda_j \nabla g^j(x^*) = \mathbf{0}$$

for the given multipliers λ_j for $j \in J$. If $D^2\tilde{\mathcal{L}}(x^*)$ is negative definite on M, then x^* is a strict local maximum point for the original constrained maximization problem. Here

$$M = \left\{ h \in \mathbb{R}^n \middle| \ \nabla g^j(x^*)h = 0 \ \forall j \in J \right\}$$

Proof: Suppose, on the contrary, that x^* is *not* a local maximum point of the constrained optimization problem. Then, we can consider $\{y_k\}$ as a sequence of *feasible* points converging to x^* such that $f(y_k) \geq f(x^*)$ for each k. More specifically, for each k, define $y_k = x^* + \varepsilon_k h_k$ with $\|h_k\| = 1$ and $\varepsilon_k > 0$. We may assume that $\varepsilon_k \to 0$ and $h_k \to h^*$ as $k \to \infty$. Using the linear approximation through differentiability,

$$f(y_k) \approx f(x^*) + \nabla f(x^*) \cdot (y_k - y^*) = f(x^*) + \varepsilon_k \nabla f(x^*) \cdot h_k$$

for k large enough. Letting $k \to \infty$, because of linearity of $\nabla f(x^*)h_k$ in h_k (continuity follows), we must have $\nabla f(x^*)h^* \geq 0$ from $f(y_k) \geq f(x^*)$. Also for each binding (active) constraint g^j, we have

$$g^j(y_k) \leq g^j(x^*)$$

Again, using the linear approximation through differentiability,

$$g^j(y_k) \approx g^j(x^*) + \nabla g^j(x^*) \cdot (y_k - x^*) = g^j(x^*) + \varepsilon_k \nabla g^j(x^*) \cdot h_k$$

for k large enough. Then, we must have $Dg^j(x^*)h^* \leq 0$ because $Dg^j(x^*)h_k$ is a linear continuous in h_k and $g^j(y_k) \leq g^j(y^*)$ for each k.

If $\nabla g^j(x^*)h^* = 0$ for all $j \in J$, then the proof goes through just as in the case of equality constraints (Theorem 5.2.3). Therefore, to complete the proof, we need to assume that there exists at least one $j \in J$ such that $\nabla g^j(x^*)h^* < 0$. Then, we obtain

$$\begin{aligned} \nabla f(x^*)h^* - \sum_{j \in J} \lambda_j Dg^j(x^*)h^* &> 0 \quad \text{because } \lambda_j > 0 \text{ for all } j \in J \\ \underbrace{\left[\nabla f(x^*) - \sum_{j \in J} \lambda_j Dg^j(x^*)\right]}_{=\mathbf{0}} h^* &> 0 \end{aligned}$$

This, however, contradicts our fulfilled condition that $\nabla f(x^*) - \sum_{j \in J} \lambda_j \nabla g^j(x^*) = \mathbf{0}$. We complete the proof. ■

Exercise 5.3.2 *Consider the following constrained maximization problem.*

$$\max f(x, y) = x \quad \textit{subject to} \quad g(x, y) = x^3 + y^2 = 0.$$

Show that this problem does not satisfy the constraint qualification.

5.3.2 Nonnegativity Constraints

Often the variables involved in economic problems are inherently nonnegative. Thus, we frequently encounter the optimization problem with nonnegativity constraints. Consider the nonlinear programming problem with *nonnegativity constraints*:

$$\max f(x) \text{ subject to } g^j(x) \leq 0 \ \forall j = 1, \ldots, m \text{ and } x_i \geq 0 \text{ for all } i = 1, \ldots, n$$

I introduce n new constraints in addition to the m original ones:

$$\begin{aligned} g^{m+1}(x) &= -x_1 \leq 0 \\ g^{m+2}(x) &= -x_2 \leq 0 \\ &\vdots \quad \vdots \quad \vdots \\ g^{m+n}(x) &= -x_n \leq 0 \end{aligned}$$

I introduce the Lagrangian multipliers $\mu_1, \ldots, \mu_n$ to go with the new constraints and form the extended Lagrangian.

$$\mathcal{L}_1(x) = f(x) - \sum_{j=1}^{m} \lambda_j g^j(x) - \sum_{i=1}^{n} \mu_i(-x_i)$$

The necessary conditions for x^* to solve the problem are

$$\begin{aligned} \frac{\partial f(x^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g^j(x^*)}{\partial x_i} + \mu_i &= 0, \ \forall i = 1, \ldots, n \\ \lambda_i \geq 0 \text{ and } \lambda_j = 0 \text{ if } g^j(x^*) &< 0, \ \forall j = 1, \ldots, m \\ \mu_i \geq 0 \text{ and } \mu_i = 0 \text{ if } x_i &> 0, \ \forall i = 1, \ldots, n \end{aligned}$$

To reduce this collection of $m + n$ constraints and $m + n$ Lagrangian multipliers, the necessary conditions for the optimization problem are sometime formulated slightly differently below.

$$\frac{\partial f(x^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g^j(x^*)}{\partial x_i} \leq 0 \ (= 0 \text{ if } x_i^* > 0), \ \forall i = 1, \ldots, n$$

This formulation follows from the first order condition.

$$\frac{\partial f(x^*)}{\partial x_i} - \sum_{j=1}^{m} \lambda_j \frac{\partial g^j(x^*)}{\partial x_i} = -\mu_i, \ \forall i = 1, \ldots, n$$

Note also that $\mu_i \geq 0$ and $-\mu_i = 0$ if $x_i > 0$

5.4 Concave Programming Problems

The constrained maximization problem is said to be a *concave programming program* in the case when $f(\cdot)$ is concave and each g^j is a convex function. In this case, the set of feasible vectors satisfying the m constraints is convex. I write the concave program as follows:

$$\max f(x) \text{ subject to } g(x) \leq \mathbf{0}$$

where $g(x) = (g^1(x), \ldots, g^m(x))$ and $\mathbf{0} = (0, \ldots, 0)$.

Even if the concave function f is not C^1 in the first-order characterization of concave function (Theorem 4.6.3), I still obtain the following result:

Theorem 5.4.1 (Existence of a Supergradient) *Let $f(\cdot)$ be concave on convex set $S \subset \mathbb{R}^n$, and let x^0 be an interior point in S. Then, there exists a vector $p \in \mathbb{R}^n$ such that for all $x \in S$,*

$$f(x) - f(x^0) \leq p \cdot (x - x^0).$$

*A vector p that satisfies the above inequality is called a **supergradient** for f at x^0.*

Proof: We skip the proof. ■

In the following results, no differentiability requirements are imposed at all. Instead, we make use of the following constraint qualification:

Definition 5.4.1 *The nonlinear programming problem satisfies the* ***Slater qualification*** *if there exists a vector $z \in \mathbb{R}^n$ such that $g(z) \ll \mathbf{0}$, i.e., $g^j(z) < 0$ for all j.*

Theorem 5.4.2 (Necessary Conditions for Concave Programming) *Suppose that the nonlinear programming is a concave programming satisfying the Slater constraint qualification. Then, the optimal value function $f^*(c)$ is defined for (at least) all $c \geq g(z)$, and has a super gradient at $\mathbf{0}$. Furthermore, if λ is any supergradient of f^* at $\mathbf{0}$, then $\lambda \geq \mathbf{0}$, and any solution x^* of the concave programming problem is an unconstrained maximum point of the Lagrangian $\mathcal{L}(x, \lambda) = f(x) - \lambda \cdot g(x)$ which also satisfies $\lambda \cdot g(x^*) = 0$ (the complementary slackness condition).*

Proof: We consider only the special but usual case where, for all $c \in \mathbb{R}^m$, the feasible set of points x that satisfy $g(x) \leq c$ is bounded, so compact because of the assumption that the functions g^j are C^1, i.e., continuous. In this case, $f^*(c)$ is defined as a maximum value whenever there exists at least one x satisfying $g(x) \leq c$, which is certainly true when $c \geq g(z)$. Then, f^* is defined for all $c \geq g(z)$. ■

Theorem 5.4.3 (Sufficient Conditions for Concave Programming) *Consider the nonlinear programming problem with $f(\cdot)$ concave and $g(\cdot)$ convex, and assume that there exists a vector $\lambda \geq \mathbf{0}$ and a feasible vector x^* which together have the property that x^* maximizes $f(x) - \lambda \cdot g(x)$ among all $x \in \mathbb{R}^n$, and $\lambda \cdot g(x^*) = 0$. Then, x^* solves the original concave problem and λ is a supergradient for f^* at $\mathbf{0}$.*

Proof: Since S is assumed to be open, x^* is an interior point in S. By Theorem 5.3.1, it only remains to prove that $\mathcal{L}(x, \lambda)$ is concave. Recall that

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{j=1}^{m} (-\lambda_j g^j(x)).$$

Since each $g^j(\cdot)$ is convex and each $\lambda_j \geq 0$, we have that $-\lambda_j g^j(x)$ is a concave function.

We also know that $f(\cdot)$ is a concave function. The rest of the proof is completed by establishing that the sum of concave functions is a concave function:

Claim 5.4.1 *Let $h^1 : S \to \mathbb{R}$ and $h^2 : S \to \mathbb{R}$ be two concave functions. Let $h(x) = h^1(x) + h^2(x)$ for each $x \in S$. Then, $h(\cdot)$ is a concave function.*

Proof: Fix $x, x^{'} \in S$ and $\alpha \in [0,1]$. Then,

$$\begin{aligned} h(\alpha x + (1-\alpha)x^{'}) &= h^1(\alpha x + (1-\alpha)x^{'}) + h^2(\alpha x + (1-\alpha)x^{'}) \\ &\geq \alpha h^1(x) + (1-\alpha)h^1(x^{'}) + \alpha h^2(x) + (1-\alpha)h^2(x^{'}) \\ &= \alpha(h^1(x) + h^2(x)) + (1-\alpha)(h^1(x^{'}) + h^2(x^{'})) \\ &= \alpha h(x) + (1-\alpha)h(x^{'}). \end{aligned}$$

Thus, $h(\cdot)$ is also concave. This completes the proof of the claim. ■

This completes the proof of the theorem. ■

5.5 Quasiconcave Programming

The following theorem is important for economists, because in many economic optimization problems, the objective function is assumed to be quasiconcave, rather than concave.

Theorem 5.5.1 (Arrow and Enthoven (1961) in *Econometrica*) ***(Sufficient Conditions for Quasiconcave Programming)****: Consider the constrained optimization problem where the objective function $f(\cdot)$ is C^1 and quasiconcave. Assume that there exist numbers $\lambda_1, \ldots, \lambda_m$ and a vector x^* such that*

1. *x^* is feasible and satisfies the Kuhn-Tucker conditions.*
2. *$\nabla f(x^*) \neq \mathbf{0}$.*
3. *$\lambda_j g_j(\mathbf{x})$ is quasiconvex for each $j = 1, \ldots, m$.*

Then, x^* is optimal.

Proof: We first prove that for all x,

$$f(x) > f(x^*) \Rightarrow \nabla f(x^*) \cdot (x - x^*) > 0. \quad (*)$$

Assume $f(x) > f(x^*)$. Due to the continuity of $f(\cdot)$, we choose $\alpha > 0$ small enough so that $f(x - \alpha \nabla f(x^*)) \geq f(x^*)$. Using the characterization of quasiconcave functions via first derivatives (Theorem 4.6.3), we have

$$\nabla f(x^*) \cdot (x - \alpha \nabla f(x^*) - x^*) \geq 0$$

This is equivalent to

$$\nabla f(x^*) \cdot (x - x^*) \geq \alpha (\nabla f(x^*))^2 > 0 \ (\because \nabla f(x^*) \neq \mathbf{0}).$$

This establishes $(*)$.

Let x be any feasible vector, i.e., $g^j(x) \leq 0$ for each $j = 1, \ldots, m$. Let $J = \{j \in \{1, \ldots, m\} |\ g^j(x^*) = 0\}$.

$$j \in J \Rightarrow \lambda_j g^j(x) \leq \lambda_j g^j(x^*) \Leftrightarrow -\lambda_j g^j(x) \geq -\lambda_j g^j(x^*).$$

$$j \notin J \Rightarrow \lambda_j = 0 \Rightarrow -\lambda_j g^j(x) \geq -\lambda_j g^j(x^*).$$

Since each $-\lambda_j g^j(x)$ is quasiconcave (because $\lambda_j g^j(x)$ is quasiconvex), we use the characterization of quasiconcave functions via first derivatives (Theorem 4.6.3),

$$-\lambda_j \nabla g^j(x^*) \cdot (x - x^*) \geq 0 \Leftrightarrow \lambda_j \nabla g^j(x^*) \cdot (x - x^*) \leq 0$$

for each $j \in \{1, \ldots, m\}$. So,

$$0 \geq \sum_{j=1}^{m} \lambda_j \nabla g^j(x^*) \cdot (x - x^*) \underbrace{=}_{KT-1} \nabla f(x^*) \cdot (x - x^*)$$

By $(*)$, we conclude $f(x) \leq f(x^*)$. Hence, x^* is optimal. ■

Chapter 6

Integration

An *integral* assigns numbers to functions in a way that can describe displacement, area, volume, and other concepts that arise by combining infinitesimal data. Integration is one of the two main operations of calculus, with *differentiation* being the other. Roughly speaking, the operation of integration is the reverse of differentiation.

6.1 What's integral (integration)?

Consider two functions $f(x)$ and $F(x)$. Suppose that we can find $f(x)$ using the differentiation of $F(x)$:

$$f(x) = F'(x).$$

Then (reversely) we can find $F(x)$ using the integration of $f(x)$:

$$F(x) = \int f(x)dx.$$

For example, if we have $f(x) = 2x$, then we have $F(x) = x^2 + c$.

6.2 Indefinite integration

Let I be an interval on $\mathbb{R}$ and $f : I \to \mathbb{R}$ be a *continuous* function. $F(\cdot)$ is called an *indefinite integral* of $f(\cdot)$ if, for all $x \in I$,

$$\int f(x)dx = F(x) + C$$

where $F^{'}(x) = f(x)$ and $C \in \mathbb{R}$ is a constant.

Remark: The symbol $\int$ is the *integral sign*, and the function $f(x)$ is the *integrand*, C is a *constant of integration*.

I provide the following examples of indefinite integration:

$$\begin{aligned}
\int x dx &= \frac{x^2}{2} + C, \\
\int \frac{1}{x^3} dx &= \int x^{-3} dx = \frac{x^{-2}}{-2} + C = -\frac{x^{-2}}{2} + C, \\
\int \sqrt{x} dx &= \int x^{1/2} dx = \frac{2}{3} x^{3/2} + C.
\end{aligned}$$

I provide some more formulas for indefinite integrals:

$$\begin{aligned}
\int x^a dx &= \frac{1}{a+1} x^{a+1} + C \ \ (a \neq -1); \\
\int \frac{1}{x} dx &= \ln|x| + C; \\
\int e^{ax} dx &= \frac{1}{a} e^{ax} + C \ \ (a \neq 0); \\
\int a^x dx &= \frac{a^x}{\ln a} + C; \\
\int \ln(x) dx &= x \ln x - x + C.
\end{aligned}$$

Lemma 6.2.1

$$\int \frac{1}{x} dx = \ln|x| + C :$$

Proof: If $x \geq 0$ then $\ln|x| = \ln x$. Thus $(\ln x)' = 1/x$. If $x < 0$, then $\ln|x| = \ln(-x)$. Thus $(\ln(-x))' = (-1)/(-x) = 1/x$. ■

Lemma 6.2.2

$$\int a^x dx = \frac{a^x}{\ln a} + C :$$

Proof: Let $f(x) = a^x$. Then $\ln f(x) = x \ln a$. $f'(x)/f(x) = \ln a$. $f'(x) = f(x) \ln a = a^x \ln a$. ■

6.2.1 Some General Rules of Indefinite Integral

$$\begin{aligned}
\int \alpha f(x) dx &= \alpha \int f(x) dx, \\
\int [f(x) + g(x)] dx &= \int f(x) dx + \int g(x) dx,
\end{aligned}$$

where $\alpha \in \mathbb{R}$ is a constant.

Example 6.2.1

$$\begin{aligned}\int (3x^4+5x^2+2)dx &= 3\int x^4dx + 5\int x^2dx + 2\int dx \\ &= 3\left(\frac{x^5}{5}+C_1\right)+5\left(\frac{x^3}{3}+C_2\right)+2\left(x+C_3\right) \\ &= \frac{3}{5}x^5+\frac{5}{3}x^3+2x+3C_1+5C_2+2C_3 \\ &= \frac{3}{5}x^5+\frac{5}{3}x^3+2x+C.\end{aligned}$$

where $C = 3C_1 + 5C_2 + 2C_3$.

Example 6.2.2

$$\begin{aligned}\int \left(\frac{3}{x}-8e^{-4x}\right)dx &= 3\int \frac{1}{x}dx - 8\int e^{-4x}dx \\ &= 3\ln|x| + 2e^{-4x} + C.\end{aligned}$$

6.3 Definite Integral and Measure of Area

We are primarily motivated by the following question: How do we compute the area A under the graph of a continuous and nonnegative function f over the interval $[a, b]$?

Let t be an arbitrary point in $[a, b]$, and let $A(t)$ denote the area under the curve $y = f(x)$ over the interval $[a, t]$. Clearly, we have $A(a) = 0$, while $A(b) = A$. Suppose we increase t by a positive amount Δt, then $A(t + \Delta t)$ is the area under the curve $y = f(x)$ over the interval $[a, t + \Delta t]$. Hence, the difference $A(t + \Delta t) - A(t)$ is the area under the curve over the interval $[t, t + \Delta t]$, denoted by ΔA.

Assume $f(t) \le f(t + \Delta t)$. Then, for all $\Delta t > 0$,

$$f(t)\Delta t \le A(t + \Delta t) - A(t) \le f(t + \Delta t)\Delta t.$$

Since $\Delta t > 0$,

$$f(t) \le \frac{A(t + \Delta t) - A(t)}{\Delta t} \le f(t + \Delta t).$$

Since f is assumed to be continuous, $f(t + \Delta t) \to f(t)$ as $\Delta t \to 0$. Thus,

$$\lim_{\Delta t \to 0} \frac{A(t + \Delta t) - A(t)}{\Delta t} = f(t).$$

This implies that $A(t)$ is differentiable at all $t \in (a, b)$ with the following property:

$$A'(t) = f(t) \text{ for all } t \in (a, b)$$

Therefore, the derivative of the area function $A(t)$ is the curve's "height function" $f(t)$, so the area function is one of the indefinite integrals of $f(t)$.

What is the area under $f(x) = x^2$ over $[0, 1]$?

$$A = \int_0^1 x^2 dx = \frac{1}{3}\left[x^3\right]_0^1 = \frac{1}{3}(1-0) = \frac{1}{3}.$$

Intuitively, we can think of the height of the function $f(x)$ multiplied with a marginal increase of x (dx). Then the integral sum in between the interval is the *area.*

Thus, I obtain the following geometric intuition regarding definite integration:

1. if $f(x) \geq 0$ over $[a, b]$, then $\int_a^b f(x)dx$ is the area of the graph of f over $[a, b]$.
2. If $f(x) \leq 0$ over $[a, b]$, then $-\int_a^b f(x)dx$ is the area of the graph of f over $[a, b]$.
3. If $f(x) \geq g(x)$ over $[a, b]$, then $\int_a^b \left[f\left(x\right) - g\left(x\right)\right] dx$ is the area in between the two graphs over $[a, b]$.

6.4 Fundamental Theorem of Calculus

Let $f : [a, b] \to \mathbb{R}$ be a *continuous* function. There exists $F : [a, b] \to \mathbb{R}$ such that

$$\begin{aligned} &(1) \quad \int_a^b f(x)dx = F(b) - F(a), \\ &(2) \quad F^{'}(x) = f(x) \text{ for all } x \in (a, b). \end{aligned}$$

I also obtain

$$\begin{aligned} \frac{d}{dx}\int_a^x f(t)dt &= f(x), \\ \frac{d}{dx}\int_x^b f(t)dt &= -f(x), \\ \frac{d}{dx}\int_a^{b(x)} f(t)dt &= f(b(x))b^{'}(x). \end{aligned}$$

6.4.1 Properties of Definite Integrals

$$\begin{aligned} (1) \quad \int_a^b f\left(x\right) dx &= -\int_b^a f\left(x\right) dx, \\ (2) \quad \int_a^a f\left(x\right) dx &= 0, \\ (3) \int_a^b \alpha f\left(x\right) dx &= \alpha \int_a^b f\left(x\right) dx, \\ (4) \quad \int_a^b f(x)dx &= \int_a^c f\left(x\right) dx + \int_c^b f(x)dx, \end{aligned}$$

where $a, b, c, \alpha \in \mathbb{R}$.

6.5 Integration by Parts

$$\int_a^b f(x)g^{'}(x)dx = [f(x)g(x)]_a^b - \int_a^b f^{'}(x)g(x)dx.$$

I discuss how we can derive the formula. Let me start from the product rule of differentiation:

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x).$$

Take the integral on the both hand sides:

$$\begin{aligned} \int (f(x)g(x))'\,dx &= \int f'(x)g(x)dx + \int f(x)g'(x)dx \\ f(x)g(x) &= \int f^{'}(x)g(x)dx + \int f(x)g^{'}(x)dx. \end{aligned}$$

I provide the following examples of the use of integration by parts.

Example 6.5.1

$$\begin{aligned} \int_0^1 xe^x dx &= [xe^x]_0^1 - \int_0^1 e^x dx \\ &= e - [e^x]_0^1 = e - (e-1) = 1, \end{aligned}$$

where $f(x) = x$ *and* $g(x) = e^x$.

Example 6.5.2

$$\begin{aligned} \int_1^e \ln(x)dx &= [\ln(x) \times x]_1^e - \int_1^e \frac{1}{x} \times x dx \\ &= e \times \ln(e) - (\ln 1) \times 1 - \int_1^e dx \\ &= e - [x]_1^e = e - (e-1) = 1, \end{aligned}$$

where $f(x) = \ln(x)$ *and* $g(x) = x$.

6.6 Integration by Substitution

$$\int_a^b f(x)dx = \int_{u_1}^{u_2} f(g(u))g^{'}(u)du,$$

where $x = g(u),\ g(u_1) = a$, and $g(u_2) = b$.

Example 6.6.1

$$\int 8x^2(3x^3-1)^{16}dx.$$

Define $u = 3x^3 - 1$. *Then,* $du/dx = 9x^2$ *so that*

$$8x^2 dx = \frac{8}{9} du.$$

Hence,

$$\int 8x^2(3x^3-1)^{16}dx = \frac{8}{9}\int u^{16}du = \frac{8}{9}\frac{u^{17}}{17} + C = \frac{8}{153}(3x^3-1)^{17} + C.$$

Example 6.6.2

$$\int_1^e \frac{1+\ln x}{x}dx.$$

Define $u = 1 + \ln x$. *Then,* $du/dx = 1/x$ *so that* $du = (1/x)dx$.
Also, $x = 1 \Leftrightarrow u = 1$ *and* $x = e \Leftrightarrow u = 2$. *So,*

$$\int_1^e \frac{1+\ln x}{x}dx = \int_1^2 u du = \frac{1}{2}\left[u^2\right]_1^2 = \frac{1}{2}(4-1) = \frac{3}{2}.$$

Example 6.6.3

$$\int \frac{x-\sqrt{x}}{x+\sqrt{x}}dx,$$

where $x > 0$. *Set* $u = \sqrt{x} \Rightarrow x = u^2$ *and* $dx = 2udu$.

$$\begin{aligned}
\int \frac{x-\sqrt{x}}{x+\sqrt{x}}dx &= \int \frac{u^2-u}{u^2+u}2udu = 2\int \frac{u^2-u}{u+1}du \\
&= 2\int \frac{u(u-1)}{u+1}du = 2\int \frac{u(u+1)-2u}{u+1}du \\
&= 2\int udu - 4\int \frac{u}{u+1}du = 2\int udu - 4\int \frac{(u+1)-1}{u+1}du \\
&= 2\int udu - 4\int du + 4\int \frac{1}{u+1}du \\
&= u^2 - 4u + 4\ln|u+1| + C
\end{aligned}$$

Therefore,

$$\int \frac{x-\sqrt{x}}{x+\sqrt{x}}dx = x - 4\sqrt{x} + 4\ln(\sqrt{x}+1) + C,$$

where $\sqrt{x} + 1 > 0$ *for all* $x > 0$.

6.6.1 More on Integration by Substitution

Example 6.6.4

$$\int_0^1 x^3\sqrt{1+x^2}dx.$$

Let $u = \sqrt{1+x^2}$. *This implies* $u^2 = 1+x^2$. *Differentiating this w.r.t.* x, *we obtain*

$$2u\frac{du}{dx} = 2x \Rightarrow udu = xdx.$$

We confirm $x = 0$ *corresponds to* $u = 1$ *and* $x = 1$ *corresponds to* $u = \sqrt{2}$. *Then,*

$$\int_0^1 x^3\sqrt{1+x^2}dx = \int_0^1 x^2\sqrt{1+x^2}xdx = \int_1^{\sqrt{2}} \underbrace{(u^2-1)}_{u^2=1+x^2} \overbrace{u}^{u=\sqrt{1+x^2}} \underbrace{udu}_{udu=xdx}$$

So,

$$\begin{aligned}
\int_0^1 x^3\sqrt{1+x^2}dx &= \int_1^{\sqrt{2}} \left(u^4-u^2\right)du = \left[\frac{u^5}{5}-\frac{u^3}{3}\right]_1^{\sqrt{2}} \\
&= \left(\frac{4\sqrt{2}}{5}-\frac{2\sqrt{2}}{3}\right)-\left(\frac{1}{5}-\frac{1}{3}\right) \\
&= \frac{12\sqrt{2}-10\sqrt{2}}{15}-\frac{3-5}{15} \\
&= \frac{2}{15}(\sqrt{2}+1).
\end{aligned}$$

Example 6.6.5 (Integrating Rational Functions and Partial Fractions)

$$\int \frac{x^4+3x^2-4}{x^2+2x}dx.$$

We first confirm the following:

$$\begin{aligned}
\frac{x^4+3x^2-4}{x^2+2x} &= \frac{x^2(x^2+2x)-2x^3+3x^2-4}{x^2+2x} \\
&= x^2+\frac{-2x(x^2+2x)+7x^2-4}{x^2+2x} \\
&= x^2-2x+\frac{7(x^2+2x)-14x-4}{x^2+2x} \\
&= x^2-2x+7-\frac{14x+4}{x^2+2x}.
\end{aligned}$$

We manipulate the term below as follows:

$$\frac{14x+4}{x^2+2x} = \frac{14x+4}{x(x+2)} = \frac{A}{x}+\frac{B}{x+2},$$

where A and B are constants to be determined. Then,

$$\frac{A}{x}+\frac{B}{x+2}=\frac{A(x+2)+Bx}{x(x+2)}=\frac{(A+B)x+2A}{x(x+2)}.$$

Thus, we set $A=2$ and $B=12$ so that

$$\frac{14x+4}{x^2+2x}=\frac{2}{x}+\frac{12}{x+2}.$$

Therefore,

$$\begin{aligned}\int\frac{x^4+3x^2-4}{x^2+2x}dx &= \int(x^2-2x+7)dx+2\int\frac{1}{x}dx+12\int\frac{1}{x+2}\\ &= \frac{x^3}{3}-x^2+7x+2\ln|x|+12\ln|x+2|+C,\end{aligned}$$

where C is a constant.

6.7 Infinite Intervals of Integration

I start from the following example.

Example 6.7.1

$$\int_0^a xe^{-cx^2}dx=\frac{-1}{2c}\left[e^{-cx^2}\right]_0^a=\frac{1}{2c}(1-e^{-ca^2}).$$

If c is a positive number,

$$e^{-ca^2}\to 0 \ \text{ as } a\to\infty$$

Then, it seems natural to write

$$\int_0^\infty xe^{-cx^2}dx=\frac{1}{2c}.$$

In statistics and economics, it is common to encounter such integrals over an infinite interval.

Definition 6.7.1 *Suppose f is a continuous function for all $x\geq a$. Then, $\int_a^b f(x)dx$ is defined for each $b\geq a$. Then, f is said to be* ***integrable over*** *$[a,\infty)$ if the limit of $\int_a^b f(x)dx$ exists as $b\to\infty$ and*

$$\lim_{b\to\infty}\int_a^b f(x)dx<\infty.$$

If f is integrable over $[a,\infty)$, we define

$$\int_a^\infty f(x)dx=\lim_{b\to\infty}\int_a^b f(x)dx \quad (*)$$

If $f(x)\geq 0$ in $[a,\infty)$, we interpret the integral $(*)$ as the ***area*** *below the graph of f over the infinite interval $[a,\infty)$.*

Example 6.7.2

$$f(x) = \lambda e^{-\lambda x},$$

where $x \geq 0$ and λ is a positive constant.

For $b > 0$, the area below the graph of f over $[0, b]$ is equal to

$$\int_0^b \lambda e^{-\lambda x} dx = \Big[(-e^{-\lambda x})\Big]_0^b = -e^{-\lambda b} + 1.$$

As $b \to \infty$, $-e^{-\lambda b} + 1 \to 1$. Therefore,

$$\int_0^\infty \lambda e^{-\lambda x} dx = \lim_{b\to\infty} \int_0^b \lambda e^{-\lambda x} dx = \lim_{b\to\infty} (-e^{-\lambda b} + 1) = 1.$$

Example 6.7.3

$$\int_1^\infty \frac{1}{x^a} dx.$$

For $a \neq 1$ and $b > 1$,

$$\int_1^b \frac{1}{x^a} dx = \int_1^b x^{-a} dx = \left[\frac{1}{1-a} x^{1-a}\right]_1^b = \frac{1}{1-a}(b^{1-a} - 1).$$

***Case of** $a > 1$: we have $b^{1-a} = 1/b^{a-1} \to 0$ as $b \to \infty$. So, x^{-a} is itegrable over $[1, \infty)$. Thus,*

$$\int_1^\infty \frac{1}{x^a} dx = \frac{1}{a-1}.$$

***Case of** $a = 1$: For $b > 1$,*

$$\int_1^b \frac{1}{x} dx = \ln b - \ln a \Rightarrow \infty \ \text{ as } b \to \infty.$$

So, $\int_1^\infty (1/x) dx$ diverges.

***Case of** $a < 1$: $b^{1-a} \to \infty$ as $b \to \infty$. So, in this case as well, $\int_1^\infty (1/x^a) dx$ diverges.*

If both limits of integration are infinite, the improper integral of a continuous function f on $(-\infty, \infty)$ is defined by

$$\int_{-\infty}^\infty f(x) dx = \int_{-\infty}^0 f(x) dx + \int_0^\infty f(x) dx. \quad (*)$$

*If **both** integrals on the right hand side of $(*)$ converge, the improver integral $\int_{-\infty}^\infty f(x) dx$ is said to **converge**; otherwise, it **diverges**.*

It is important to keep in mind the following note:

$$\lim_{b\to\infty} \int_{-b}^b f(x) dx < \infty$$

$$\not\Rightarrow \quad \lim_{b\to-\infty} \int_b^0 f(x) dx < \infty \ \text{ and } \ \lim_{b\to\infty} \int_0^b f(x) dx < \infty.$$

Example 6.7.4

$$\int_{-\infty}^{\infty} xe^{-cx^2}dx = \int_{-\infty}^{0} xe^{-cx^2}dx + \underbrace{\int_{0}^{\infty} xe^{-cx^2}dx}_{=1/2c}$$

$$\int_{-\infty}^{0} xe^{-cx^2}dx = \lim_{a\to-\infty}\int_{a}^{0} xe^{-cx^2}dx = \lim_{a\to-\infty}\left[-\frac{1}{2c}e^{-cx^2}\right]_{a}^{0} = -\frac{1}{2c}.$$

It follows that

$$\int_{-\infty}^{\infty} xe^{-cx^2}dx = -\frac{1}{2c} + \frac{1}{2c} = 0.$$

6.8 Differentiation under the Integral Sign

How does the value of the integral change if the parameter changes? Let $f : (x,t) \mapsto f(x,t) \in \mathbb{R}$. I define

$$F(x) = \int_{c}^{d} f(x,t)dt,$$

where c and d are constants. Then, what is

$$F^{'}(x)?$$

Theorem 6.8.1 (Leibniz's Formula: Simple Case)

$$F(x) = \int_{c}^{d} f(x,t)dt \Rightarrow F^{'}(x) = \int_{c}^{d} \frac{\partial f(x,t)}{\partial x}dt$$

Sketch of Proof:

$$\begin{aligned}
F^{'}(x) &= \lim_{h\to0}\frac{F(x+h)-F(x)}{h} \\
&= \lim_{h\to0}\int_{c}^{d}\frac{f(x+h,t)-f(x,t)}{h}dt \\
&= \int_{c}^{d}\lim_{h\to0}\frac{f(x+h,t)-f(x,t)}{h}dt \\
&\quad (\because \text{we can change the order of lim and } \textstyle\int.) \\
&= \int_{c}^{d}\frac{\partial f(x,t)}{\partial x}dt. \blacksquare
\end{aligned}$$

Example 6.8.1 (Discounted Present Value of an Asset) *Let an asset generate a value $f(t) \in \mathbb{R}$ in period $t \in [0,T]$ and T be the terminal period.*

Assume that r is the interest rate for the safe asset. Then, the discounted present value of this asset in period 0 is computed as:

$$K = \int_{0}^{T} f(t)e^{-rt}dt.$$

By Leibniz's rule,

$$\frac{dK}{dr} = \int_0^T f(t)(-t)e^{-rt}dt = -\int_0^T tf(t)e^{-rt}dt.$$

Theorem 6.8.2 (Leibniz's Formula: General Case) *Suppose $f(x,t)$ and $\partial f(x,t)/\partial x$ are continuous over the rectangle by $a \leq x \leq b$ and $c \leq t \leq d$. Assume further that $u(\cdot)$ and $v(\cdot)$ are C^1 functions over $[a,b]$ and*

$$\begin{aligned} u([a,b]) &= \{y \in \mathbb{R} |\ \exists x \in [a,b]\ \ s.t.\ u(x) = y\} \subseteq [c,d], \\ v([a,b]) &= \{y \in \mathbb{R} |\ \exists x \in [a,b]\ \ s.t.\ v(x) = y\} \subseteq [c,d]. \end{aligned}$$

Then,

$$\begin{aligned} F(x) &= \int_{u(x)}^{v(x)} f(x,t)dt \\ \Rightarrow \qquad & F^{'}(x) = f(x,v(x))v^{'}(x) - f(x,u(x))u^{'}(x) + \int_{u(x)}^{v(x)} \frac{\partial f(x,t)}{\partial x}dt. \end{aligned}$$

Sketch of Proof: Let $H : [a,b] \times [c,d] \times [c,d] \to \mathbb{R}$ be the following function.

$$H(x,u,v) = \int_u^v f(x,t)dt.$$

Since we can set $F(x) = H(x,u(x),v(x))$, by the chain rule,

$$\begin{aligned} F^{'}(x) &= H_x^{'} + H_u^{'}u^{'}(x) + H_v^{'}v^{'}(x) \\ &= \int_u^v \frac{\partial f(x,t)}{\partial x}dt + f(x,v)v^{'}(x) - f(x,u)u^{'}(x). \end{aligned}$$

This completes the proof. ■

Example 6.8.2

$$F(x) = \int_x^{x^2} \frac{1}{2}t^2 x dt.$$

$$\begin{aligned} F^{'}(x) &= \frac{1}{2}(x^2)^2 x \cdot 2x - \frac{1}{2}x^2 x \cdot 1 + \int_x^{x^2} \frac{1}{2}t^2 dt \\ &= x^6 - \frac{1}{2}x^3 + \left[\frac{1}{6}t^3\right]_x^{x^2} \\ &= x^6 - \frac{1}{2}x^3 + \frac{1}{6}\left((x^2)^3 - x^3\right) \\ &= \frac{7}{6}x^6 - \frac{2}{3}x^3. \end{aligned}$$

In this case, the integral $F(x)$ is easy to calculate explicitly:

$$F(x) = \frac{1}{2}x\int_x^{x^2} t^2 dt = \frac{1}{2}x\left[\frac{1}{3}t^3\right]_x^{x^2} = \frac{1}{6}(x^7 - x^4)$$

Example 6.8.3 (A Growth Model of Kaldor and Mirrlees (REStud, 1962))

$$N(t) = \int_{t-T(t)}^{t} n(\tau)e^{-\delta(t-T(t))}d\tau,$$

where

- *$N(t)$: the working population;*
- *$n(\tau)$: the number of workers available at period τ to operate new equipment per unit period;*
- *δ: the rate of depreciation of the equipment per unit period;*
- *$T(t)$: the life time of the equipment which is retired at period t.*

If $n(\cdot)$ is continuous and T is C^1, Leibniz's formula gives

$$\begin{aligned}\dot{N}(t) &= n(t)e^{-\delta(t-T(t))} - n(t-T(t))e^{-\delta(t-T(t))}(1-\dot{T}(t)) \\ &\quad + \int_{t-T(t)}^{t} n(\tau)(-\delta)(1-\dot{T}(t))e^{-\delta(t-T(t))}d\tau \\ &= \left[n(t) - (1-\dot{T}(t))n(t-T(t))\right]e^{-\delta(t-T(t))} - \delta(1-\dot{T}(t))N(t).\end{aligned}$$

Example 6.8.4 (Firm's Profits) *The profit of a firm is $\pi(t)$ at each time $t \in [0,T]$. At time t, the discounted value of future profit is*

$$V(t) = \int_t^T \pi(\tau)e^{-r(\tau-t)}d\tau$$

where r is the discount rate.

Define $u(t) = t, v(t) = T$, and $f(t,\tau) = \pi(\tau)e^{-r(\tau-t)}$. Then $u'(t) = 1, v'(t) = 0$, and $f_t'(t,\tau) = \pi(\tau)re^{-r(\tau-t)}$. By Liebniz's rule, we obtain

$$V'(t) = -\pi(t)e^{-r(t-t)} + \int_t^T \pi(\tau)re^{-r(\tau-t)}d\tau = -\pi(t) + rV(t).$$

The first term reflects the fact that the future is shortened when t increases and the second term reflects the fact that as time advances future profit at any given time is obtained sooner, and is thus worth more.

Theorem 6.8.3 (Leibniz's Formula: the case of Unbounded Intervals) *Suppose that $f(x,t)$ and $f_x^{'}(x,t)$ are continuous for all $t \geq c$ and all $x \in [a,b]$, and the integral*

$$\int_c^{\infty} f(x,t)dt$$

converges for each $x \in [a,b]$. Assume further that there exists a function $p(t)$, independent of x, for which

$$\int_c^{\infty} p(t)dt$$

converges and $|f_x^{'}(x,t)| \leq p(t)$ for all $t \geq c$ and $x \in [a,b]$. Then,

$$\frac{d}{dx}\int_c^{\infty} f(x,t)dt = \int_c^{\infty} f_x^{'}(x,t)dt.$$

Example 6.8.5 (Acquisition Value of the Firm) *:*

- *$K(t)$: the capital stock of some firm at time t;*
- *$p(t)$: the purchase price per unit of capital at time t; and*
- *$R(t)$: the rental price per unit of capital at time t.*

The acquisition value $V(t)$ of the firm's capital is computed: for all t,

$$V(t) = p(t)K(t) = \int_t^{\infty} R(\tau)K(\tau)e^{-r(\tau-t)}d\tau. \quad (*)$$

$V(t)$ should equal the discounted present value of the returns from using the firm's capital.

Example 6.8.6 (Find an Expression for $R(t)$) *Set $v(t) = \infty$ and $u(t) = t$. Then, we have $v^{'}(t) = 0$ and $u^{'}(t) = 1$. Also, set $f(t,\tau) = R(\tau)K(\tau)e^{-r(\tau-t)}$ so that $f_t^{'}(t,\tau) = rR(\tau)K(\tau)e^{-r(\tau-t)}$.*

Using Leibniz's rule, we differentiate $(*)$ with respect to t so that

$$\begin{aligned} V^{'}(t) &= p^{'}(t)K(t) + p(t)K^{'}(t) \\ &= f(t,v(t))v^{'}(t) - f(t,u(t))u^{'}(t) + \int_t^{\infty} f_t^{'}(t,\tau)d\tau \\ &= -R(t)K(t) + \int_t^{\infty} R(\tau)K(\tau)re^{-r(\tau-t)}d\tau \\ &= -R(t)K(t) + rV(t) = -R(t)K(t) + rp(t)K(t). \end{aligned}$$

Solving the above equation for $R(t)$ yields

$$\begin{aligned} R(t) &= \left(r - \frac{K^{'}(t)}{K(t)}\right)p(t) - p^{'}(t) \\ &= \underbrace{rp(t)}_{\text{interest cost}} - \overbrace{p(t)\frac{K^{'}(t)}{K(t)}}^{\text{depreciation}} - \underbrace{p^{'}(t)}_{\text{capital gain}}. \end{aligned}$$

Chapter 7

Differential Equations

What is a Differential Equation? Economists often study the changes over time in economic variables. A model of economic growth, for example, typically contains a relation between the change in the capital stock and the value of output. The laws of motion of these variables are usually expressed in terms of one or more equations. If time is regarded as continuous and the equations involve unknown functions and their derivative, the relationship may be modeled as *differential equations*.

7.1 First-Order Ordinary Differential Equations

An *ordinally* differential equation is one for which the unknown is a function of only one variable. *Partial differential equations* are equations where the unknown is a function of two or more variables, and one or more of the partial derivatives of the function are included. For the moment, we restrict attention to *first-order* ordinary differential equations where only the first-order derivatives of the unknown functions of one variable are included.

Let me start from discussing the following examples of first-order differential equations:

$$\begin{aligned} &(a) \quad \dot{x} = x + t; \\ &(b) \quad \dot{K} = \alpha\sigma K + H_0 e^{\mu t}; \text{ and} \\ &(c) \quad \dot{k} = sf(k) - \lambda k, \end{aligned}$$

where we often use $\dot{x} = dx/dt$.

Solving equation (a) means finding all functions $x(t)$ such that, for every value of t, the derivative $\dot{x}(t)$ is equal to $x(t) + t$. In equation (b), $K(t)$ is the unknown function, whereas α, σ, H_0, and μ are constants. In equation (c), $f(k)$ is a given function, whereas s and λ are constants. The unknown function is $k = k(t)$.

7.1.1 Solutions

A first-order differential equation is written as:

$$\dot{x} = F(t, x) \quad (*)$$

where $F(\cdot)$ is a given function of two variables and $x = x(t)$ is the unknown function.

A *solution* of $(*)$ in an interval $I \subseteq \mathbb{R}$ is any differentiable function $\varphi : I \to \mathbb{R}$ such that $x = \varphi(t)$ satisfies $(*)$, that is, $\dot{\varphi}(t) = F(t, \varphi(t))$ for all $t \in I$. The graph of a solution is called a *solution curve* or *integral curve*. For example, the equations $(a), (b)$, and (c) are all of the form $(*)$.

We illustrate by example why a differential equation has infinitely man solutions:

Example 7.1.1 *Let $x = Ce^t - t - 1$ where C is a constant. Then,*

$$\dot{x} = Ce^t - 1 = Ce^t - t - 1 + t = x + t.$$

Therefore, $x = Ce^t - t - 1$ is a solution to $\dot{x} = x + t$.

In particular, $x = -t - 1$ and $x = e^t - t - 1$ are also solutions to $\dot{x} = x + t$.

The set of all solutions of a differential equation is called its *general solution*, while any specific function that satisfies the equation is called a *particular solution*.

7.1.2 Initial Value Problem

Many models specify both that a function satisfies a differential equation and that the value of the function, or the values of the derivatives of the function, take certain values for some values of the variable. Consider the following example:

$$\dot{x}(t) = x(t) + t \text{ and } x(0) = 1 \qquad (*)$$

If $t = 0$ denotes the initial time, then $x(0) = 1$ is called an *initial condition* and we call $(*)$ an *initial value problem*. We thus have $x(t) = 2e^t - t - 1$ as a unique solution to $(*)$.

7.1.3 Separable Equations

The differential equation $\dot{x} = F(t, x)$ is *separable* if $F(t, x) = f(t)g(x)$.

Step 1: Write

$$\frac{dx}{dt} = f(t)g(x).$$

Step 2: Separate the variables:

$$\frac{dx}{g(x)} = f(t)dt.$$

Step 3: Integrate each side:

$$\int \frac{dx}{g(x)} = \int f(t)dt.$$

Example 7.1.2

$$\frac{dx}{dt} = -2tx^2.$$

$$\begin{aligned} \textit{Separate:} \quad & -\frac{dx}{x^2} = 2tdt \\ \textit{Integrate:} \quad & -\int \frac{dx}{x^2} = \int 2tdt \\ \textit{Evaluete:} \quad & \frac{1}{x} = t^2 + C \end{aligned}$$

So, the general solution is $x(t) = 1/(t^2 + C)$. *If* $x(1) = -1$, *we have* $C = -2$. *So, we obtain*

$$x(t) = \frac{1}{t^2 - 2}.$$

7.1.4 A Model of Economic Growth

According to Wikipedia, economic growth is the increase in the inflation-adjusted market value of the goods and services produced by an economy over time. It is conventionally measured as the real GDP. It is one of the most fundamental questions in Economics and I think everyone wants to know what drives economic growth.

Example 7.1.3 (The Solow Model) *There is a representative consumer who produces final output* $Y(t)$ *using only two inputs, physical capital* $K(t)$ *and labor* $L(t)$ *at each period* t. *The production function* $F(\cdot)$ *takes the form*

$$Y(t) = F(K(t), L(t)).$$

Output $Y(t)$ *is a homogeneous good that can be consumed (i.e.,* $C(t)$*) or invested (i.e.,* $I(t)$*), to create new units of physical capital.*

For simplicity, we assume that the representative consumer saves a constant fraction of the income, denoted by $s \in [0, 1]$ *and that the population grows at a constant, exogenous rate* $n = \dot{L}/L \geq 0$*:*

$$L(t) = L_0 e^{nt}.$$

where $L_0 > 0$ *is constant.*

For all $K > 0$ *and* $L > 0$, $F(\cdot)$ *is "neoclassical" in the following sense:*

$$\frac{\partial F}{\partial K} > 0; \ \frac{\partial F}{\partial L} > 0; \ \frac{\partial^2 F}{\partial K^2} < 0; \ \textit{and} \ \frac{\partial^2 F}{\partial L^2} < 0.$$

In addition, $F(\cdot)$ *exhibits constant returns to scale:* $\forall \lambda > 0$,

$$F(\lambda K, \lambda L) = \lambda F(K, L).$$

Assume that $F(K, L) = AK^{1-\alpha}L^{\alpha}$ *with* $\alpha \in (0, 1)$. *Then, it can be easily checked that the Cobb-Douglass function satisfies the neoclassical properties and it exhibits constant returns to scale.*

In a closed economy, outputs equals income, and the amount invested equals the amount saved. So,

$$\dot{K} = \frac{dK}{dt} = sY = sAK^{1-\alpha}L^{\alpha} = sAL_0^{\alpha}e^{\alpha nt}K^{1-\alpha}.$$

Clearly, this is a separable differential equation.

$$\begin{aligned} &\textit{Separate:} && K^{\alpha-1}dK = sAL_0^{\alpha}e^{\alpha nt}dt \\ &\textit{Integrate:} && \int K^{\alpha-1}dK = sAL_0^{\alpha}\int e^{\alpha nt}dt \\ &\textit{Evaluate:} && \frac{1}{\alpha}K^{\alpha} = \frac{sAL_0^{\alpha}}{\alpha n}e^{\alpha nt} + C. \end{aligned}$$

Setting $C_1 = \alpha C$, we obtain $K^{\alpha} = (sA/n)L_0^{\alpha}e^{\alpha nt} + C_1$. Furthermore, if $K(0) = K_0 > 0$ (initial condition), we obtain $C_1 = K_0^{\alpha} - (sA/n)L_0^{\alpha}$. Therefore, the solution is

$$K = \left[K_0^{\alpha} + (sA/n)L_0^{\alpha}(e^{\alpha nt} - 1)\right]^{1/\alpha}.$$

7.1.5 First-Order Linear Equations

A *first-order linear differential equation* is one that can be written in the form

$$\dot{x} + a(t)x = b(t) \qquad (*)$$

where $a(t)$ and $b(t)$ denote continuous functions of t in a certain interval, and $x = x(t)$ is the unknown function.

Equation $(*)$ is called *linear* because the left-hand-side of $(*)$ is a linear function of x and $\dot{x}$. I illustrate first-order linear equations by the following examples:

1. $\dot{x} + x = t$;
2. $\dot{x} + 2tx = 4t$; and
3. $(t^2 + 1)\dot{x} + e^t x = t \ln t$.

The last equation can be rearranged into:

$$\dot{x} + \frac{e^t}{t^2 + 1}x = \frac{t \ln t}{t^2 + 1}.$$

The Simplest Case of First-Order Linear Equations

Consider the following equation with a and b as constants, where $a \neq 0$:

$$\dot{x} + ax = b.$$

Multiplying this equation by e^{at}, we have

$$\dot{x}e^{at} + axe^{at} = be^{at}.$$

This is equivalent to

$$\frac{d}{dt}(xe^{at}) = be^{at}.$$

Taking the integration on both hand sides, we obtain

$$xe^{at} = \int be^{at}dt = \frac{b}{a}e^{at} + C$$

Therefore, we obtain

$$\dot{x} + ax = b \Leftrightarrow x = Ce^{-at} + \frac{b}{a},$$

where C is a constant. If $C = 0$, we obtain the constant function $x(t) = b/a$. We say that $x = b/a$ is an *equilibrium state* or *steady state*. If $a > 0$, the solution $x = Ce^{-at} + b/a$ converges to b/a as $t \to \infty$. In this case, the equation is said to be *stable*, because every solution of the equation converges to an equilibrium as $t \to \infty$.

Example 7.1.4 (Price Adjustment Mechanism) *Let $D(P) = a - bP$ denote the demand and $S(P) = \alpha + \beta P$ the supply of a certain commodity when its price is P. Here, a, b, α, and β are positive constants.*

Assume $\dot{P} = \lambda[D(P) - S(P)]$ where $\lambda > 0$ is a constant. Substituting $D(P) = a - bP$ and $S(P) = \alpha + \beta P$ for the differential equation, we obtain

$$\dot{P} + \lambda(b + \beta)P = \lambda(a - \alpha).$$

Using the formula, we have

$$P = Ce^{-\lambda(b+\beta)t} + \frac{a - \alpha}{b + \beta}.$$

Because $\lambda(b + \beta)$ is positive, as $t \to \infty$, P converges to the equilibrium price $P^e = (a - \alpha)/(b + \beta)$, for which $D(P^e) = S(P^e)$. Thus, the equilibrium is stable.

7.1.6 Variable RHS Case of First-Order Linear Equations

Consider the following equation:

$$\dot{x} + ax = b(t)$$

Multiplying the equation by e^{at}, we obtain

$$\dot{x}e^{at} + axe^{at} = b(t)e^{at} \Leftrightarrow \frac{d}{dt}(xe^{at}) = b(t)e^{at}.$$

Hence, we obtain

$$xe^{at} = \int b(t)e^{at}dt + C.$$

Multiplying the last equation by e^{-at}, we obtain

$$\dot{x} + ax = b(t) \Leftrightarrow x = Ce^{-at} + e^{-at}\int e^{at}b(t)dt.$$

Example 7.1.5 (Economic Growth with Foreign Investment)

$$\begin{array}{ll}(a) & Y(t) = \sigma K(t);\\ (b) & \dot{K}(t) = \alpha Y(t) + H(t); \ \ \textit{and}\\ (c) & N(t) = N_0 e^{\rho t},\end{array}$$

where $Y(t)$ is total domestic product per year; $K(t)$ is capital stock; $H(t)$ is the net inflow foreign investment per year; and $N(t)$ is the size of the population.

From (a) and (b), we have

$$\dot{K}(t) - \alpha\sigma K(t) = H(t).$$

If $H(t) = H_0 e^{\mu t}$, we use the formula we derived to obtain

$$\begin{aligned} K(t) &= Ce^{\alpha\sigma t} + e^{\alpha\sigma t}\int e^{-\alpha\sigma t} H_0 e^{\mu t} dt \\ &= Ce^{\alpha\sigma t} + e^{\alpha\sigma t} H_0 \int e^{(\mu-\alpha\sigma)t} dt \\ &= Ce^{\alpha\sigma t} + e^{\alpha\sigma t}\frac{H_0}{\mu-\alpha\sigma} e^{(\mu-\alpha\sigma)t} \\ &= Ce^{\alpha\sigma t} + \frac{H_0}{\mu-\alpha\sigma} e^{\mu t}. \end{aligned}$$

If $K(0) = K_0$, we obtain $C = K_0 - H_0/(\mu - \alpha\sigma)$. Thus,

$$K(t) = \left(K_0 - \frac{H_0}{\mu-\alpha\sigma}\right) e^{\alpha\sigma t} + \frac{H_0}{\mu-\alpha\sigma} e^{\mu t}$$

7.1.7 General Case of First-Order Linear Equations

Consider

$$\dot{x} + a(t)x = b(t).$$

Let $A(t) = \int a(t)dt$. This means that $\dot{A}(t) = a(t)$.

Multiplying the equation by $e^{A(t)}$, we obtain

$$\dot{x}e^{A(t)} + a(t)xe^{A(t)} = b(t)e^{A(t)}.$$

This is equivalent to

$$\frac{d}{dt}(xe^{A(t)}) = b(t)e^{A(t)}.$$

Integrating the equation, we obtain

$$xe^{A(t)} = \int b(t)e^{A(t)}dt + C.$$

Multiplying the equation by $e^{-A(t)}$, we have

$$x = Ce^{-A(t)} + e^{-A(t)} \int b(t)e^{A(t)}dt.$$

I summarize this result:

$$\dot{x} + a(t)x = b(t) \Leftrightarrow x = e^{-\int a(t)dt} \left(C + \int e^{\int a(t)dt} b(t)dt \right).$$

7.1.8 Qualitative Theory and Stability

We are often interested not in the exact form of the solution of a differential equation, but only in the *qualitative* properties of this solution. Many differential equations in economics can be expressed as:

$$\dot{x} = F(x). \qquad (*)$$

This equation is called *autonomous*.

One of the most important properties of a differential equation is whether it has any *equilibrium* or *steady states*. These corresponds to solutions of the equation that do not change over time.

Stability

In many economic applications, it is also very important to know whether an equilibrium state is *stable*.

Definition 7.1.1 *:*

- *A point a represents an* ***equilibrium (steady) state*** *for equation $(*)$ if $F(a) = 0$.*
- *An equilibrium state a is called* ***globally asymptotically stable*** *if for any initial point $x(0) = x_0$, $x(t) \to a$ as $t \to \infty$.*
- *An equilibrium state a is called* ***locally asymptotically stable*** *if for any initial point $x(0) = x_0 \in B_\varepsilon(a)$, $x(t) \to a$ as $t \to \infty$.*

Theorem 7.1.1 *: Let a be an equilibrium state to $\dot{x} = F(x)$.*

1. *$F(a) = 0$ and $F'(a) < 0 \Rightarrow a$ is a locally asymptotically stable equilibrium.*
2. *$F(a) = 0$ and $F'(a) > 0 \Rightarrow a$ is an unstable equilibrium.*

Proof: We skip the proof. ■

Example 7.1.6 (The Solow Model Revisited) *Now assume that capitals depreciate at the constant rate $\delta > 0$.*

The net increase in the stock of physical capital at a point in time equals gross investment less depreciation:

$$\dot{K} = I - \delta K = s \cdot F(K, L) - \delta K.$$

Exploiting constant returns to scale technology, we obtain

$$Y = F(K, L) = LF(K/L, 1) = Lf(k),$$

where $k \equiv K/L$ is the capital-labor ratio, $y \equiv Y/L$ is per capita output, and the function $f(k) = F(k, 1)$. Since $Y = Lf(k)$, we obtain

$$\partial Y/\partial K = f^{'}(k) > 0 \text{ and } \partial^2 Y/\partial K^2 = f^{''}(k)/L < 0.$$

So, we have $f^{''}(k) < 0$. The production function can now be expressed as $y = f(k)$.

$$\dot{K} = sF(K, L) - \delta K \Rightarrow \frac{\dot{K}}{L} = sf(k) - \delta k.$$

Using this, we compute

$$\dot{k} = \frac{d(K/L)}{dt} = \frac{\dot{K}L - K\dot{L}}{L^2} = \frac{\dot{K}}{L} - k\frac{\dot{L}}{L} = sf(k) - (n + \delta)k.$$

$\dot{k} = sf(k) - (n + \delta)k$ is the fundamental differential equation of the Solow model. Since $f(0) = 0$, $f^{'}(k) > 0$, and $f^{''}(k) < 0$ for all $k > 0$, there is a unique steady state $k^* > 0$ in the Solow Model:

$$sf(k^*) = (n + \delta)k^*.$$

If $\lim_{k\to 0} f^{'}(k) = \infty$ and $\lim_{k\to\infty} f^{'}(k) = 0$ (known as the Inada Conditions), the unique steady state k^* is locally asymptotically stable.

7.2 Second-Order Differential Equations

In an important area of dynamic optimization called the *calculus of variations*, the first-order condition for optimality involves a second-order differential equation.

The typical second-order differential equation takes the form

$$\ddot{x} = F(t, x, \dot{x}) \quad (*)$$

where F is a given fixed function, $x = x(t)$ is the unknown function, and $\dot{x} = dx/dt$.

The new feature here is the presence of the second derivative $\ddot{x} = d^2x/dt^2$. A *solution* of $(*)$ on an interval I is a twice differentiable function that satisfies the equation.

Example 7.2.1

$$\ddot{x} = k \text{ (} k \text{ is a constant).}$$

Taking integration on the equation, we obtain

$$\dot{x} = \int k dt = kt + A,$$

where A is some constant. Taking further integration on the equation above, we obtain

$$\int (kt + A)dt = \frac{k}{2}t^2 + At + B,$$

where B is some constant.

7.2.1 Differential Equations where x or t is Missing

There are two cases to consider:

Case 1: $\ddot{x} = F(t, \dot{x})$

In this case, x is missing. I introduce the new variable $u = \dot{x}$. Then, Case 1 becomes $\dot{u} = F(t, u)$, which is a first-order differential equation.

Example 7.2.2

$$\ddot{x} = \dot{x} + t.$$

Define $u = \dot{x}$. Then, the equation is transformed to $\dot{u} = u + t$. This first-order differential equation has the general solution

$$u = Ae^t - t - 1,$$

where A is a constant. This is equivalent to

$$\dot{x} = Ae^t - t - 1.$$

Integrating this equation, we obtain

$$x = \int (Ae^t - t - 1)dt = Ae^t - \frac{1}{2}t^2 - t + B,$$

where B is a constant. Assume that $x(0) = 1$ and $\dot{x}(0) = 2$. First,

$$\dot{x}(0) = A - 1 = 2 \Rightarrow A = 3.$$

Second,

$$x(0) = A + B = 1 \underbrace{\Rightarrow}_{A=3} B = -2.$$

Then,

$$x = 3e^t - \frac{1}{2}t^2 - t - 2.$$

Case 2: $\ddot{x} = F(x, \dot{x})$

In this case, t is not explicitly present in the equation and the equation is called *autonomous*. Let $t' = dt/dx$ and $t'' = d^2t/dx^2$. Assume that $t' \neq 0$. By the inverse function theorem,

$$\dot{x} = \frac{dx}{dt} = \frac{1}{dt/dx} = \frac{1}{t'}$$

Once again, by the inverse function theorem,

$$\ddot{x} = \frac{d^2x}{dt^2} = \frac{d}{dt}\left(\frac{dx}{dt}\right) = \frac{d}{dt}\left(\frac{1}{t'}\right) = \frac{d}{dx}\left(\frac{1}{t'}\right)\frac{dx}{dt} = -(t')^{-2} \cdot (t'') \cdot (t')^{-1} = -\frac{t''}{(t')^3}.$$

So, the original differential equation is transformed to

$$t'' = -(t')^3 F(x, 1/t').$$

This is now the same as Case 1 where we interpret x as t and t' as $\dot{x}$.

7.2.2 Second-Order Linear Differential Equations

The general second-order linear differential equation is

$$\ddot{x} + a(t)\dot{x} + b(t)x = f(t) \quad (*)$$

where $a(t), b(t)$, and $f(t)$ are all continuous functions of t on some interval I.

Let me begin with the *homogeneous* equation

$$\ddot{x} + a(t)\dot{x} + b(t)x = 0 \quad (**)$$

Assume that $u_1 = u_1(t)$ and $u_2 = u_2(t)$ both satisfy $(**)$. Define $x = Au_1 + Bu_2$ where A and B are constants. Then,

$$\begin{aligned} \dot{x} &= A\dot{u}_1 + B\dot{u}_2 \\ \ddot{x} &= A\ddot{u}_1 + B\ddot{u}_2 \end{aligned}$$

Substituting these into $(**)$, we obtain

$$\begin{aligned} \ddot{x} + a(t)\dot{x} + b(t)x &= A\ddot{u}_1 + B\ddot{u}_2 + a(t)(A\dot{u}_1 + B\dot{u}_2) + b(t)(Au_1 + Bu_2) \\ &= A[\ddot{u}_1 + a(t)\dot{u}_1 + b(t)u_1] + B[\ddot{u}_2 + a(t)\dot{u}_2 + b(t)u_2] \\ &= 0. \end{aligned}$$

This is true for all choices of A and B.

Equation $(*)$ is called a *nonhomogeneous equation*, and $(**)$ is the homogeneous equation associated with it.

Suppose we are able to find *some particular solution* $u^* = u^*(t)$ of $(*)$. Assume further that $x(t)$ is an arbitrary solution to $(*)$. Then, define $v = v(t) = x(t) - u^*(t)$. Then,

$$\begin{aligned} \dot{v} &= \dot{x} - \dot{u}^* \\ \ddot{v} &= \ddot{x} - \ddot{u}^*. \end{aligned}$$

So, we compute

$$\begin{aligned}\ddot{v} + a(t)\dot{v} + b(t)v &= \ddot{x} - \ddot{u}^* + a(t)(\dot{x} - \dot{u}^*) + b(t)(x - u^*) \\ &= [\ddot{x} + a(t)\dot{x} + b(t)x] - [\ddot{u}^* + a(t)\dot{u}^* + b(t)u^*] \\ &= f(t) - f(t) = 0.\end{aligned}$$

Thus, $x(t) - u^*(t)$ is a solution to the homogeneous equation $(**)$.

Since I have argued that the solution to $(**)$ is of the form $Au_1(t) + Bu_2(t)$,

$$x(t) - u^*(t) = Au_1(t) + Bu_2(t),$$

where $u_1(t)$ and $u_2(t)$ are two nonproportional solutions to $(**)$, and A and B are arbitrary constants. I therefore summarize the preceding discussion in the following theorem:

Theorem 7.2.1 *:*

1. *The general solution of the homogeneous differential equation* $(**)$ *is*

$$x = Au_1(t) + Bu_2(t),$$

where $u_1(t)$ *and* $u_2(t)$ *are any two solutions that are not proportional, and* A *and* B *are arbitrary constants.*

2. *The general solution of the nonhomogeneous differential equation* $(*)$ *is*

$$x = Au_1(t) + Bu_2(t) + u^*(t),$$

where $Au_1(t) + Bu_2(t)$ *is the general solution of the associated homogeneous equation, and* $u^*(t)$ *is any particular solution of* $(*)$.

7.2.3 Constant Coefficients

Consider

$$\ddot{x} + a\dot{x} + bx = 0, \quad (**)$$

where a and b are arbitrary constants, and $x = x(t)$ is the unknown function.

It *seems* a good idea to try possible solutions x with the property that $x, \dot{x}$, and $\ddot{x}$ are all constant multiples of each other. The exponential function $x = e^{rt}$ has this property because $\dot{x} = re^{rt} = rx$ and $\ddot{x} = r^2e^{rt} = r^2x$.

So, I adjust the constant r in order that $x = e^{rt}$ satisfies $(**)$. This requires us to arrange that $r^2e^{rt} + are^{rt} + be^{rt} = 0$. Therefore, e^{rt} satisfies $(**)$ if and only if r satisfies

$$r^2 + ar + b = 0 \ \ (***).$$

This is the *characteristic equation* of the differential equation $(**)$.

If $a^2 - 4b \geq 0$, the characteristic equation has two real roots:

$$\begin{aligned} r_1 &= -\frac{1}{2}a + \sqrt{\frac{1}{4}a^2 - b}, \\ r_2 &= -\frac{1}{2}a - \sqrt{\frac{1}{4}a^2 - b}. \end{aligned}$$

Theorem 7.2.2 *The general solution of $\ddot{x} + a\dot{x} + bx = 0$ depends on the roots of the characteristic equation $r^2 + ar + b = 0$ as follows:*

1. *If $a^2 - 4b > 0$, when there are two distinct real roots, then*

$$x = Ae^{r_1 t} + Be^{r_2 t}, \quad \textit{where} \quad r_{1,2} = -\frac{1}{2}a \pm \sqrt{\frac{1}{4}a^2 - b}.$$

2. *If $a^2 - 4b = 0$, when there is a double real root, then*

$$x = (A + Bt)e^{rt}, \quad \textit{where} \quad r = -\frac{1}{2}a.$$

3. *If $a^2 - 4b < 0$, when there are two complex roots, then*

$$x = e^{\alpha t}(A\cos\beta t + B\sin\beta t), \quad \textit{where} \quad \alpha = -\frac{1}{2}a,\ \beta = \sqrt{b - \frac{1}{4}a^2}.$$

Proof: **Part 1**: When $a^2/4 - b > 0$, there are two distinct characteristic roots r_1 and r_2. Then, the two functions $e^{r_1 t}$ and $e^{r_2 t}$ satisfy $(**)$ and are not proportional to each other. So, the general solution in this case is $Ae^{r_1 t} + Be^{r_2 t}$.
Part 2: If $a^2/4 - b = 0$, $r = -a/2$ is a double root of $(***)$. So, $u_1 = e^{rt}$ satisfies $(**)$. We claim that $u_2 = te^{rt}$ also satisfies $(**)$. We compute the following:

$$\begin{aligned} \dot{u}_2 &= e^{rt} + tre^{rt}; \\ \ddot{u}_2 &= re^{rt} + re^{rt} + tr^2e^{rt} = 2re^{rt} + tr^2e^{rt}. \end{aligned}$$

Plugging $\dot{u}_2, \ddot{u}_2$ above into the left-hand side of $(**)$, we obtain

$$\begin{aligned} \ddot{u}_2 + a\dot{u}_2 + bu_2 &= 2re^{rt} + tr^2e^{rt} + a(e^{rt} + tre^{rt}) + bte^{rt} \\ &= e^{rt}(a + 2r) + te^{rt}(r^2 + ar + b) \\ &= 0 \ (\because\ r = -a/2,\ r^2 + ar + b = 0\ (***)). \end{aligned}$$

Thus, $u_2 = te^{rt}$ also satisfies $(**)$. We can also claim that $u_1(t) = e^{rt}$ and $u_2(t) = te^{rt}$ are the functions that are not proportional to each other. Suppose not, that is, there exists $\alpha \neq 0$ such that $u_1(t) = \alpha u_2(t)$ for every $t \in I \subseteq \mathbb{R}$, where I denotes the interval over which these two functions are defined. Then, we must have $\alpha = 1/t$, which is not a constant. This is the desired contradiction.

Part 3: If $a^2/4 - b < 0$, there are two distinct complex roots for $(***)$:

$$r_1, r_2 = -\frac{1}{2}a \pm i\sqrt{b - \frac{1}{4}a^2}.$$

Let $\alpha \equiv -a/2$ and $\beta \equiv \sqrt{b - a^2/4}$. Then, the two solutions to $(**)$ are given as follows:

$$u_1(t) = e^{(\alpha+i\beta)t} \text{ and } u_2(t) = e^{(\alpha-i\beta)t}.$$

Using rules for exponents, we write

$$u_1(t) = e^{(\alpha+i\beta)t} = e^{\alpha t}e^{i\beta t} \text{ and } u_2(t) = e^{(\alpha-i\beta)t} = e^{\alpha t}e^{-i\beta t}.$$

Using the following well-known *Euler's formula*: $e^{ix} = \cos x + i \sin x$, we can re-express $u_1(t)$ and $u_2(t)$ as follows:

$$\begin{aligned} u_1(t) &= e^{\alpha t}(\cos \beta t + i \sin \beta t); \\ u_2(t) &= e^{\alpha t}(\cos(-\beta t) + i \sin(-\beta t)) = e^{\alpha t}(\cos \beta t - i \sin \beta t). \end{aligned}$$

Then, the general solution to $(**)$ becomes

$$x(t) = e^{\alpha t}\left[k_1(\cos \beta t + i \sin \beta t) + k_2(\cos \beta t - i \sin \beta t)\right],$$

for any k_1, k_2, real or complex. If we choose k_1 and k_2 to be complex numbers that are complex conjugates to each other, we have

$$k_1 = c_1 + ic_2 \text{ and } k_2 = c_1 - ic_2,$$

where c_1, c_2 are real numbers. Since $(\cos \beta t + i \sin \beta t)$ and $(\cos \beta t - i \sin \beta t)$ are complex conjugates of each other,

$$\begin{aligned} (c_1 + ic_2)(\cos \beta t + i \sin \beta t) &= (c_1 \cos \beta t - c_2 \sin \beta t) + i(c_2 \cos \beta t + c_1 \sin \beta t) \\ (c_1 - ic_2)(\cos \beta t - i \sin \beta t) &= (c_1 \cos \beta t - c_2 \sin \beta t) - i(c_2 \cos \beta t + c_1 \sin \beta t) \end{aligned}$$

are complex conjugates of each other. Thus,

$$k_1(\cos \beta t + i \sin \beta t) + k_2(\cos \beta t - i \sin \beta t) = 2(c_1 \cos \beta t - c_2 \sin \beta t).$$

Setting $A = 2c_1$ and $B = -2c_2$, we obtain the general solution to $(**)$ as follows:

$$x(t) = e^{\alpha t}(A \cos \beta t + B \sin \beta t).$$

This completes the proof. ■

Example 7.2.3

$$\ddot{x} - 3x = 0.$$

The characteristic equation $r^2 - 3 = 0$ *has two real roots:* $r_1 = -\sqrt{3}$ *and* $r_2 = \sqrt{3}$. *Then, the general solution is*

$$x = Ae^{-\sqrt{3}t} + Be^{\sqrt{3}t}.$$

Example 7.2.4

$$\ddot{x} - 4\dot{x} + 4x = 0.$$

The characteristic equation $r^2 - 4r + 4 = 0$ *has a double real root:* $r = 2$*. Hence, the general solution is*

$$x = (A + Bt)e^{2t}.$$

Example 7.2.5

$$\ddot{x} - 6\dot{x} + 13x = 0.$$

The characteristic equation $r^2 - 6r + 13 = 0$ *has two complex roots because* $(r-3)^2 + 4 = 0$*. Then, we compute*

$$\begin{aligned} \alpha &= -\frac{1}{2}a = 3 \\ \beta &= \sqrt{13 - \frac{1}{4}(-6)^2} = \sqrt{13 - 9} = 2. \end{aligned}$$

So, the general solution is

$$x = e^{3t}(A\cos 2t + B\sin 2t).$$

7.2.4 The Nonhomogeneous Equation

Consider the nonhomogeneous eqaution

$$\ddot{x} + a\dot{x} + bx = f(t), \quad (*)$$

where $f(t)$ is an arbitrary continuous function. If $b = 0$ in $(*)$, then the term in x is missing and the substitution $u = \dot{x}$ transforms the equation into a linear equation of first order. So, we may assume $b \neq 0$. In what follows, I consider the following four cases and provide in the way we obtain a particular solution is provided explicitly.

Case (A): $f(t) = A$ (constant)

We check to see if $(*)$ has a solution that is constant, $u^* = c$. Then, $\dot{u}^* = \ddot{u}^* = 0$. So, the equation reduces to $bc = A$. Hence, $c = A/b$. For $b \neq 0$: $\ddot{x} + a\dot{x} + bx = A$ has a particular solution $u^* = A/b$.

Case (B): $f(t)$ is polynominal

Suppose $f(t)$ is a polynomial of degree n. Then, a **reasonable** guess is that $(*)$ has a particular solution that is also a polynomial of degree n, of the form $u^* = A_n t^n + A_{n-1}t^{n-1} + \cdots + A_1 t + A_0$. We determine the undetermined coefficients $A_n, A_{n-1}, \ldots, A_0$ by requiring u^* to satisfy $(*)$.

Example 7.2.6

$$\ddot{x} - 4\dot{x} + 4x = t^2 + 2.$$

Let $u^* = At^2 + Bt + C$. *Then,*

$$\begin{aligned} \dot{u}^* &= 2At + B \\ \ddot{u}^* &= 2A. \end{aligned}$$

Plugging these into the LHS of the equation, we obtain

$$2A - 4(2At + B) + 4(At^2 + Bt + C) = 4At^2 + 4(B - 2A)t + (2A - 4B + 4C).$$

Then, we must have $A = 1/4$; $B = 2A = 1/2$; *and* $1/2 - 2 + 4C = 2$, *which implies* $4C = 7/2$, *which further implies* $C = 7/8$. *Hence,*

$$u^* = \frac{1}{4}t^2 + \frac{1}{2}t + \frac{7}{8}.$$

Case (C): $f(t) = pe^{qt}$

It seems natural to try a particular solution of the form $u^* = Ae^{qt}$. Then,

$$\dot{u}^* = Aqe^{qt} \text{ and } \ddot{u}^* = Aq^2e^{qt}.$$

$$\ddot{x} + a\dot{x} + bx = f(t) \Rightarrow Ae^{qt}(q^2 + aq + b) = pe^{qt}.$$

Hence, if $q^2 + aq + b \neq 0$,

$$u^* = \frac{p}{q^2 + aq + b}e^{qt}$$

is a particular solution to $\ddot{x} + a\dot{x} + bx = f(t)$. The condition $q^2 + aq + b \neq 0$ means that q is not a solution of the characteristic equation.

Case (D): $f(t) = p\sin rt + q\cos rt$

Let $u^* = A\sin rt + B\cos rt$ and adjust the constants A and B so that the coefficients of $\sin rt$ and $\cos rt$ match.

Example 7.2.7

$$\ddot{x} - 4\dot{x} + 4x = 2\cos 2t.$$

Let $u^* = A\sin 2t + B\cos 2t$. *Then, we have*

$$\dot{u}^* = 2A\cos 2t - 2B\sin 2t \text{ and } \ddot{u}^* = -4A\sin 2t - 4B\cos 2t.$$

Therefore,

$$\begin{aligned} & \ddot{x} + a\dot{x} + bx = f(t) \\ \Leftrightarrow\ & -4A\sin 2t - 4B\cos 2t - 4(2A\cos 2t - 2B\sin 2t) \\ & +4(A\sin 2t + B\cos 2t) = 2\cos 2t \\ \Leftrightarrow\ & 8B\sin 2t - 8A\cos 2t = 2\cos 2t \end{aligned}$$

This implies that $A = -1/4$ *and* $B = 0$. *Thus,*

$$u^* = -\frac{1}{4}\sin 2t.$$

7.2.5 Stability for Linear Equations

Will small changes in the initial conditions have any effect on the long-run behavior of the solution to a given system of differential equations or will the effect "die out" as $t \to \infty$? In the latter case, the system is called *asymptocially stable*. On the other hand, if small changes in the initial conditions might lead to significant differences in the behavior of the solution in the long run, then the system is *unstable*.

Consider the second-order nonhomogeneous differential equation

$$\ddot{x} + a(t)\dot{x} + b(t)x = f(t). \quad (*)$$

Recall that the general solution of $(*)$ is $x = Au_1(t) + Bu_2(t) + u^*(t)$, where $Au_1(t)+Bu_2(t)$ is the general solution of the associated homogeneous equation (with $f(t)$ replaced by zero), and $u^*(t)$ is a particular solution of the nonhomogeneous equation $(*)$.

Definition 7.2.1 $(*)$ *is called* ***globally asymptotically stable*** *if every solution* $Au_1(t) + Bu_2(t)$ *of the associated homogeneous equation tends to 0 as* $t \to \infty$ *for all values of* A *and* B. *Then, the effect of the initial conditions "dies out" as* $t \to \infty$.

Example 7.2.8

$$(1)\ \ddot{x} + 2\dot{x} + 5x = e^t.$$

The corresponding characteristic equation is $r^2 + 2r + 5 = 0$, *with complex roots* $r_1 = -1 + 2i, r_2 = -1 - 2i$, *so* $u_1 = e^{-t}\cos 2t$ *and* $u_2 = e^{-t}\sin 2t$ *are linearly independent solutions of the homogeneous equation. Since* $\cos 2t$ *and* $\sin 2t$ *are both less than or equal to 1 in absolute value and* $e^{-t} \to 0$ *as* $t \to \infty$, u_1 *and* u_2 *tend to 0 as* $t \to \infty$. *So, the equation is globally asymptotically stable.*

$$(2)\ \ddot{x} + \dot{x} - 2x = 3t^2 + 2.$$

The corresponding characteristic equation is $r^2 + r - 2 = 0$, *with two real roots* $r_1 = 1, r_2 = -2$, *so* $u_1 = e^t$ *and* $u_2 = e^{-2t}$ *are linearly independent solutions of the homogeneous equation. Since* $u_1 = e^t$ *does not tend to 0 as* $t \to \infty$, *the equation is not globally asymptotically stable.*

Theorem 7.2.3 *The equation* $\ddot{x} + a\dot{x} + bx = f(t)$ *is globally asymptotically stable if and only if both roots of the characteristic equation* $r^2 + ar + b = 0$ *have negative real parts.*

Proof: We prove this by considering the following three cases:

Case I: $\frac{1}{4}a^2 - b > 0$

In this case, we have $x = Ae^{r_1 t} + Be^{r_2 t}$, where $r_1, r_2 = -\frac{1}{2}a \pm \sqrt{\frac{1}{4}a^2 - b}$. Then, $Ae^{r_1 t} + Be^{r_2 t} \to 0$ as $t \to \infty$ for all values of A and B if and only if $e^{r_1 t} \to 0$ and $e^{r_2 t} \to 0$, which is equivalent to $r_1 < 0$ and $r_2 < 0$.

Case II: $\frac{1}{4}a^2 - b = 0$

In this case, we have $x = (A + Bt)e^{rt}$, where $r = -\frac{1}{2}a$. Then, $(A + Bt)e^{rt} \to 0$ as $t \to \infty$ for all values of A and B if and only if $te^{rt} \to 0$ as $t \to \infty$, which is equivalent to $r < 0$.

Case III: $\frac{1}{4}a^2 - b < 0$

In this case, we have $r_1, r_2 = \alpha \pm i\beta$ so that $x = e^{\alpha t}(A \cos \beta t + B \sin \beta t)$, where $\alpha = -\frac{1}{2}a, \beta = \sqrt{b - \frac{1}{4}a^2}$. Since $\cos \beta t$ and $\sin \beta t$ are both less than or equal to 1 in absolute value, $x \to 0$ as $t \to \infty$ for all values of A and B if and only if $e^{\alpha t} \to 0$ as $t \to \infty$, which is equivalent to $\alpha < 0$. ■

Corollary 7.2.1 *$\ddot{x} + a\dot{x} + bx = f(t)$ is globally asymptotically stable if and only if $a > 0$ and $b > 0$.*

Proof: The two roots (real or complex) r_1 and r_2 of the quadratic characteristic equation $r^2 + ar + b = 0$ have the property that $r^2 + ar + b = (r - r_1)(r - r_2) = r^2 - (r_1 + r_2)r + r_1 r_2$. Hence, $a = -r_1 - r_2$ and $b = r_1 r_2$. In Cases (I) and (II) in the previous theorem, the system is globally asymptotically stable if and only if $r_1 < 0$ and $r_2 < 0$, which is equivalent to $a > 0$ and $b > 0$. In Case (III) in the previous theorem, we have $r_1, r_2 = \alpha \pm i\beta$. Then, the system is globally asymptotically stable if and only if $\alpha < 0$. Then, $a = -(r_1 + r_2) = -2\alpha > 0$ and $b = r_1 r_2 = \alpha^2 + \beta^2 > 0$. ■

Example 7.2.9

$$\ddot{\nu} + \left(\mu - \frac{\lambda}{a}\right)\dot{\nu} + \lambda\gamma\nu = -\frac{\lambda}{a}\dot{b}(t),$$

where μ, λ, γ, and a are constants, and $\dot{b}(t)$ is a fixed function. By the previous corollary, the equation is globally asymptotically stable if and only if $\mu > \frac{\lambda}{a}$ and $\lambda\gamma > 0$.

Chapter 8

Calculus of Variations

I begin by introducing a problem from optimal growth theory that is closely related to Ramsey's pioneering discussion of optimal saving.

Example 8.0.1 (How Much Should a Nation Save?) *Consider an economy evolving over time where $K = K(t)$ denotes the capital stock, $C = C(t)$ consumption, and $Y = Y(t)$ net national product at time t. Suppose that*

$$Y = f(K), \quad \text{where} \quad f'(K) > 0 \quad \text{and} \quad f''(K) \leq 0.$$

For each t, assume that

$$f(K(t)) = C(t) + \dot{K}(t),$$

which means that output, $Y(t) = f(K(t))$, is divided between consumption, $C(t)$, and investment, $\dot{K}(t)$.

Let $K(0) = K_0$ be a historically given capital stock existing "today" at $t = 0$ and suppose that there is a fixed planning period $[0, T]$.

For each choice of investment function $\dot{K}(t)$ on the interval $[0, T]$, capital is fully determined by

$$K(t) = K_0 + \int_0^t \dot{K}(\tau)d\tau,$$

and in turn, determines $C(t)$.

Assume that the society has a utility function U, where $U(C)$ is the utility (flow) the country enjoys when the total consumption is C. Suppose also that

$$U'(C) > 0 \quad \text{and} \quad U''(C) < 0.$$

For each $t \geq 0$, we multiply $U(C(t))$ by the discount factor e^{-rt}. Frank Ramsey (1928) argues that r must be zero.

The goal of investment policy is to find the path of capital $K = K(t)$, with $K(0) = K_0$, that maximizes

$$\int_0^T U(C(t))e^{-rt}dt = \int_0^T U(f(K(t)) - \dot{K}(t))e^{-rt}dt.$$

Usually, some terminal condition on $K(t)$ is imposed. For example, $K(T) = K_T$ where K_T is given. One possibility is $K_T = 0$, with no capital left for times after T.

8.1 The Euler Equation

More generally, I consider the following problem:

$$\max \int_{t_0}^{t_1} F(t, x, \dot{x})dt \quad \text{subject to } x(t_0) = x_0 \text{ and } x(t_1) = x_1 \quad (*)$$

Here F is a given C^2 function of three variables, whereas t_0, t_1, x_0, and x_1 are given numbers.

Leonhard Euler (1744) proved that a function $x(t)$ can only solve problem $(*)$ if $x(t)$ satisfies the differential equation:

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) = 0 \quad (**)$$

where $\partial F/\partial x = F_2'(t, x, \dot{x})$ and $\partial F/\partial \dot{x} = F_3'(t, x, \dot{x})$. Equation $(**)$ is called the *Euler equation*.

Assuming that $x = x(t)$ is C^2, we find that

$$\frac{d}{dt}\left(\frac{\partial F(t, x, \dot{x})}{\partial \dot{x}}\right) = \frac{\partial^2 F}{\partial t \partial \dot{x}} \cdot 1 + \frac{\partial^2 F}{\partial x \partial \dot{x}} \cdot \dot{x} + \frac{\partial^2 F}{\partial \dot{x} \partial \dot{x}} \cdot \ddot{x}$$

Inserting this into $(**)$, we obtain

$$\frac{\partial^2 F}{\partial \dot{x} \partial \dot{x}} \cdot \ddot{x} + \frac{\partial^2 F}{\partial x \partial \dot{x}} \cdot \dot{x} + \frac{\partial^2 F}{\partial t \partial \dot{x}} - \frac{\partial F}{\partial x} = 0.$$

This equation can be written as

$$F_{33}''\ddot{x} + F_{32}''\dot{x} + F_{31}'' - F_2' = 0.$$

So, the Euler equation is a differential equation of the second-order (if $F_{33}'' \neq 0$).

Example 8.1.1 *Consider*

$$\max \int_0^2 (4 - 3x^2 - 16\dot{x} - 4(\dot{x})^2)e^{-t}dt, \; x(0) = -8/3, \; x(2) = 1/3.$$

Set $F(t, x, \dot{x}) = (4 - 3x^2 - 16\dot{x} - 4(\dot{x})^2)e^{-t}$. *So,*

$$\begin{aligned} \frac{\partial F}{\partial x} &= -6xe^{-t} \\ \frac{\partial F}{\partial \dot{x}} &= (-16 - 8\dot{x})e^{-t}. \end{aligned}$$

Next, compute

$$\frac{d}{dt}\left[(-16-8\dot{x})e^{-t}\right] = 16e^{-t} + 8\dot{x}e^{-t} - 8\ddot{x}e^{-t}.$$

By the Euler equation, we have

$$-6xe^{-t} - 16e^{-t} + 8\ddot{x}e^{-t} - 8\dot{x}e^{-t} = 0 \Rightarrow \ddot{x} - \dot{x} - \frac{3}{4}x = 2.$$

This is a second-order linear differential equation with constant coefficients. The characteristic equation is

$$r^2 - r - \frac{3}{4} = 0 \Leftrightarrow \left(r + \frac{1}{2}\right)\left(r - \frac{3}{2}\right) = 0.$$

The nonhomogeneous equation has a particular solution, A/b (in the formula) $= 2/(-3/4) = -8/3$. Thus, the general solution is

$$x = Ae^{-\frac{1}{2}t} + Be^{\frac{3}{2}t} - \frac{8}{3},$$

where A and B are arbitrary constants. The boundary conditions $x(0) = -8/3$ and $x(2) = 1/3$ imply

$$\begin{aligned} 0 &= A + B \\ Ae^{-1} + Be^{3} &= 3 \end{aligned}$$

Then, we obtain $A = -3/(e^3 - e^{-1})$ and $B = -A$ so that

$$x = x(t) = -\frac{3}{e^3 - e^{-1}}e^{-\frac{1}{2}t} + \frac{3}{e^3 - e^{-1}}e^{\frac{3}{2}t} - \frac{8}{3}.$$

This is the only solution of the Euler equation that satisfies the given boundary conditions.

8.2 Why the Euler Equation is Necessary

The Euler equation plays a similar role in the calculus of variations as the familiar first-order conditions in static optimization.

Theorem 8.2.1 (Necessity and Sufficiency of the Euler Equation) *Suppose that F is a C^2 function of three variables. Suppose that $x^*(t)$ maximizes or minimizes*

$$J(x) = \int_{t_0}^{t_1} F(t, x, \dot{x})dt,$$

among all ***admissible*** *functions $x(t)$, i.e., all C^1 functions $x(t)$ defined on $[t_0, t_1]$ that satisfy the boundary conditions:*

$$x(t_0) = x_0, \quad x(t_1) = x_1, \quad (x_0 \text{ and } x_1 \text{ given numbers})$$

Then, $x^*(t)$ is a solution of the Euler equation

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) = 0.$$

If $F(t, x, \dot{x})$ is concave (convex) in $(x, \dot{x})$, an admissible $x^*(t)$ that satisfies the Euler equation solves the maximization (minimization) problem.

Proof: (**Necessity**) We rather provide the heuristic version of the proof. Suppose $x^* = x^*(t)$ is an optimal solution to the maximization problem and let $\mu(t)$ be any C^2 function that satisfies $\mu(t_0) = \mu(t_1) = 0$.

For each real number $\alpha \in \mathbb{R}$, define a *perturbed* function $x(t)$ by

$$x(t) = x^*(t) + \alpha\mu(t).$$

Note that if α is small, the function $x(t)$ is near the function $x^*(t)$. Clearly, $x(t)$ is admissible because $x^*(t)$ and $\mu(t)$ are C^2 and

$$\begin{aligned} x(t_0) &= x^*(t_0) + \alpha\mu(t_0) = x_0 + \alpha \cdot 0 = x_0 \\ x(t_1) &= x^*(t_1) + \alpha\mu(t_1) = x_1 + \alpha \cdot 0 = x_1 \end{aligned}$$

If $\mu(t)$ is a fixed function, then $J(x^* + \alpha\mu)$ is a function $I(\alpha)$ of only the single scalar α, given by

$$I(\alpha) = \int_{t_0}^{t_1} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))dt \qquad (1)$$

Obviously, $I(0) = J(x^*)$. Also, because of the hypothesis that $x^*(t)$ is optimal,

$$J(x^*) \geq J(x^* + \alpha\mu),\ \forall\alpha \Leftrightarrow I(0) \geq I(\alpha),\ \forall\alpha$$

Because I is a differentiable function and $\alpha = 0$ is an interior point in the domain of I, one must have

$$I^{'}(0) = 0.$$

By Leibniz's formula, we differentiate $I(\alpha)$ with respect to α,

$$I^{'}(\alpha) = \int_{t_0}^{t_1} \frac{\partial}{\partial\alpha} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))dt.$$

By the chain rule,

$$\frac{\partial}{\partial\alpha} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t)) = F_2^{'} \cdot \mu(t) + F_3^{'} \cdot \dot{\mu}(t),$$

where $F_2^{'}$ and $F_3^{'}$ are evaluated at $(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))$. When $\alpha = 0$, we have

$$I^{'}(0) = \int_{t_0}^{t_1} \left[F_2^{'}(t, x^*(t), \dot{x}^*(t)) \cdot \mu(t) + F_3^{'}(t, x^*(t), \dot{x}^*(t)) \cdot \dot{\mu}(t)\right] dt,$$

or, in more compact notation,

$$I^{'}(0) = \int_{t_0}^{t_1} \left[\frac{\partial F^*}{\partial x}\mu(t) + \frac{\partial F^*}{\partial \dot{x}}\dot{\mu}(t) \right] dt$$

where $*$ indicates that the derivatives are evaluated at $(t, x^*, \dot{x}^*)$.

By integration by parts,

$$\begin{aligned}
\int_{t_0}^{t_1} \frac{\partial F^*}{\partial \dot{x}}\dot{\mu}(t)dt &= \left[\left(\frac{\partial F^*}{\partial \dot{x}} \right) \mu(t) \right]_{t_0}^{t_1} - \int_{t_0}^{t_1} \frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}} \right) \mu(t)dt \\
&= \left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_1} \mu(t_1) - \left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_0} \mu(t_0) \\
&\quad - \int_{t_0}^{t_1} \frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}} \right) \mu(t)dt \\
&= - \int_{t_0}^{t_1} \frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}} \right) \mu(t)dt \\
&\quad (\because \ \mu(t_0) = \mu(t_1) = 0)
\end{aligned}$$

Therefore, $I^{'}(0) = 0$ reduces to

$$\int_{t=t_0}^{t=t_1} \left[\frac{\partial F^*}{\partial x} - \frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}} \right) \right] \mu(t)dt = 0. \qquad (2)$$

So far, $\mu(t)$ was a fixed function. But the above equation (2) must hold for **all** functions $\mu(t)$ that are C^2 on $[t_0, t_1]$ and that are zero at t_0 and t_1.

Then, it seems to be reasonable to conclude that the bracket expression in (2) must be zero for all $t \in [t_0, t_1]$.

(**Sufficiency**) Suppose that $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$. Assume further that $x^* = x^*(t)$ satisfies the Euler equation as well as the boundary conditions $x^*(t_0) = x_0$ and $x^*(t_1) = x_1$.

Let $x = x(t)$ be an arbitrary admissible function in the problem. Since $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$, we use the following first-order characterization of concave functions: $f : S \to \mathbb{R}$ is concave if and only if, for any $x, x^{'} \in S$,

$$f(x^{'}) - f(x) \leq \nabla f(x) \cdot (x^{'} - x).$$

Setting $x = (x^*, \dot{x}^*), x^{'} = (x, \dot{x})$, and $f(x) = F(t, x, \dot{x})$, we obtain

$$F(t, x, \dot{x}) - F(t, x^*, \dot{x}^*) \leq \frac{\partial F(t, x^*, \dot{x}^*)}{\partial x}(x - x^*) + \frac{\partial F(t, x^*, \dot{x}^*)}{\partial \dot{x}}(\dot{x} - \dot{x}^*).$$

Using the Euler equation, we further obtain

$$\begin{aligned}
F^* - F &\geq \frac{\partial F^*}{\partial x}(x^* - x) + \frac{\partial F^*}{\partial \dot{x}}(\dot{x}^* - \dot{x}) \\
&= \left[\frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}} \right) \right] (x^* - x) + \frac{\partial F^*}{\partial \dot{x}}(\dot{x}^* - \dot{x}) \\
&= \frac{d}{dt}\left[\frac{\partial F^*}{\partial \dot{x}}(x^* - x) \right].
\end{aligned}$$

Since the above inequality holds for all $t \in [t_0, t_1]$, integrating the above yeilds

$$\int_{t_0}^{t_1} (F^* - F)dt \geq \int_{t_0}^{t_1} \frac{d}{dt}\left[\frac{\partial F^*}{\partial \dot{x}}(x^* - x)\right] dt = \left[\frac{\partial F^*}{\partial \dot{x}}(x^* - x)\right]_{t_0}^{t_1} = 0,$$

where the last equality follows because $x^*(t_0) = x(t_0) = x_0$ and $x^*(t_1) = x(t_1) = x_1$.

It follows that

$$\int_{t_0}^{t_1} [F(t, x^*, \dot{x}^*) - F(t, x, \dot{x})]\, dt \geq 0,$$

for every admissible function $x = x(t)$. This confirms that $x^*(t)$ solves the maximization problem. This completes the proof. ■

We revisit the following example.

Example 8.2.1

$$\max \int_0^2 (4 - 3x^2 - 16\dot{x} - 4(\dot{x})^2)e^{-t}dt,\ x(0) = -8/3,\ x(2) = 1/3.$$

Set $F(t, x, \dot{x}) = (4 - 3x^2 - 16\dot{x} - 4(\dot{x})^2)e^{-t}$. We compute the Hessian matrix of F:

$$H(x, \dot{x}) = \begin{pmatrix} \partial^2 F/\partial x^2 & \partial^2 F/\partial \dot{x}\partial x \\ \partial^2 F/\partial x \partial \dot{x} & \partial^2 F/\partial \dot{x}^2 \end{pmatrix} = \begin{pmatrix} -6e^{-t} & 0 \\ 0 & -8e^{-t} \end{pmatrix}.$$

This implies that $H(x, \dot{x})$ is negative definite, which further implies that F is strictly concave in x and $\dot{x}$. Therefore, the solution of the Euler equation that satisfies the given boundary conditions is indeed the solution to the maximization problem.

8.3 Optimal Savings

Consider the Ramsey (optimal growth) problem:

$$\max \int_0^T U(f(K(t)) - \dot{K}(t))e^{-rt}dt, \quad K(0) = K_0, \quad K(T) = K_T.$$

Assume $f'(K) > 0, f''(K) \leq 0, U'(C) > 0$, and $U''(C) < 0$. Let $F(t, K, \dot{K}) = U(C)e^{-rt}$ with $C = f(K) - \dot{K}$. Then, we obtain

$$\begin{aligned} \frac{\partial F}{\partial K} &= U'(C)f'(K)e^{-rt} \\ \frac{\partial F}{\partial \dot{K}} &= -U'(C)e^{-rt}. \end{aligned}$$

This implies that the Euler equation reduces to

$$U'(C)f'(K)e^{-rt} - \frac{d}{dt}\left(-U'(C)e^{-rt}\right) = 0.$$

We compute

$$\frac{d}{dt}\left(U^{'}(C)e^{-rt}\right) = U^{''}(C)\dot{C}e^{-rt} - rU^{'}(C)e^{-rt}$$

Inserting this into the Euler equation, we obtain

$$\begin{aligned}
& U^{'}(C)f^{'}(K)e^{-rt} + U^{''}(C)\dot{C}e^{-rt} - rU^{'}(C)e^{-rt} = 0 \\
\Leftrightarrow \quad & \left[U^{'}(C)f^{'}(K) + U^{''}(C)\dot{C} - rU^{'}(C)\right]e^{-rt} = 0 \\
\Leftrightarrow \quad & U^{'}(C)f^{'}(K) + U^{''}(C)\dot{C} - rU^{'}(C) = 0 \\
\Leftrightarrow \quad & U^{'}(C)(f^{'}(K) - r) + U^{''}(C)\dot{C} = 0.
\end{aligned}$$

This implies

$$\frac{\dot{C}}{C} = \frac{f^{'}(K) - r}{\frac{-CU^{''}(C)}{U^{'}(C)}}$$

Let

$$\sigma(C) = -\frac{CU^{''}(C)}{U^{'}(C)}.$$

Define $1/\sigma(C)$ as the *intertemporal elasticity of substitution* at C. With this additional concept, the Euler equation simplifies to

$$\frac{\dot{C}}{C} = \frac{f^{'}(K) - r}{\sigma(C)}.$$

This means

$$\frac{\dot{C}}{C} > 0 \Leftrightarrow f^{'}(K(t)) > r.$$

Hence, consumption increases if and only if the marginal productivity of capital exceeds the discount rate.

If we use the fact that $\dot{C} = f^{'}(K)\dot{K} - \ddot{K}$ in the Euler equation, we get

$$\ddot{K} - f^{'}(K)\dot{K} + \frac{U^{'}(C)}{U^{''}(C)}(r - f^{'}(K)) = 0. \quad (*)$$

Because f is concave ($f^{''}(K) \leq 0$), it follows that $f(K) - \dot{K}$ is also concave in $(K, \dot{K})$, as it is a sum of two concave functions. The function U is increasing and concave, so $U(f(K) - \dot{K})e^{-rt}$ is also concave in $(K, \dot{K})$.

This can also be directly checked via the negative seminidefiniteness of the associated Hessian matrix. We first compute all the second-order derivatives of F:

$$\begin{aligned}
\frac{\partial^2 F}{\partial K^2} &= U^{''}(f(K) - \dot{K})(f^{'}(K))^2 e^{-rt} + U^{'}(f(K) - \dot{K})f^{''}(K)e^{-rt} < 0 \\
\frac{\partial^2 F}{\partial \dot{K}^2} &= -U^{''}(f(K) - \dot{K})(-1)e^{-rt} = U^{''}(f(K) - \dot{K})e^{-rt} < 0 \\
\frac{\partial^2 F}{\partial K \partial \dot{K}} = \frac{\partial^2 F}{\partial \dot{K} \partial K} &= U^{''}(f(K) - \dot{K})f^{'}(K)(-1)e^{-rt} = -U^{''}(f(K) - \dot{K})f^{'}(K)e^{-rt} > 0.
\end{aligned}$$

We form the Hessian matrix:

$$\begin{aligned} H(K,\dot{K}) &= \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} \\ &= e^{-2rt}\begin{pmatrix} U''(f(K)-\dot{K})(f'(K))^2 + U'(f(K)-\dot{K})f''(K) & -U''(f(K)-\dot{K})f'(K) \\ -U''(f(K)-\dot{K})f'(K) & U''(f(K)-\dot{K}) \end{pmatrix}. \end{aligned}$$

We also compute the determinant of $H(K,\dot{K})$:

$$|H(K,\dot{K})| = U'(f(K)-\dot{K})U''(f(K)-\dot{K})f''(K)e^{-2rt} \geq 0.$$

Since $h_{11} < 0$, $h_{22} < 0$; and $|H(K,\dot{K})| \geq 0$, the Hessian matrix $H(K,\dot{K})$ is negative semidefinite. Thus, $F(t,K,\dot{K})$ is concave in $(K,\dot{K})$.

Therefore, any solution to $(*)$ that satisfies the boundary conditions must be a solution to the problem.

Suppose that $f(K) = bK$ and $U(C) = C^{1-v}/(1-v)$. Assume further that $b > 0, v > 0, v \neq 1$, and $b \neq (b-r)/v$. In this case, the Euler equation becomes

$$\ddot{K} - \left(b - \frac{r-b}{v}\right)\dot{K} + \frac{b-r}{v}bK = 0 \;\Rightarrow (\lambda - b)\left(\lambda - \frac{b-r}{v}\right) = 0,$$

which is the characteristic function of λ. Because $b \neq (b-r)/v$, this second-order differential equation has the general solution

$$K(t) = Ae^{bt} + Be^{(b-r)t/v}.$$

The constants A and B are determined by

$$\begin{aligned} K_0 &= A + B \\ K_T &= Ae^{bT} + Be^{(b-r)T/v}. \end{aligned}$$

8.4 More General Terminal Conditions

In economic applications, the initial point is usually fixed, while in many models, the terminal value of the unknown function can be free, or subject to more general restrictions.

The problems I study are formulated as

$$\max \int_{t_0}^{t_1} F(t,x,\dot{x})dt, \quad x(t_0) = x_0, \quad \text{(a) } x(t_1) \text{ free or (b) } x(t_1) \geq x_1 \quad (*)$$

Theorem 8.4.1 (Transversality Conditions) *If $x^*(t)$ solves problem $(*)$ with either (a) or (b) as the terminal condition, then $x^*(t)$ must satisfy the Euler equation. With the terminal condition (a), the **transversality condition** is*

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} = 0,$$

where $F^* \equiv F(t, x^*, \dot{x}^*)$. *With the terminal condition (b), the* ***transversality condition*** *is*

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} \begin{cases} = 0 & \text{if } x^*(t_1) > x_1 \\ \leq 0 & \text{otherwise} \end{cases}$$

Remark: We can interpret $(\partial F^*/\partial \dot{x})_{t=t_1}$ as the marginal value of investment at $t = t_1$.

Proof: We recall the following equation used in the proof for the necessity of Euler equation:

$$\int_{t=t_0}^{t=t_1} \left[\frac{\partial F^*}{\partial x} - \frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}}\right)\right] \mu(t)dt = 0. \qquad (2)$$

We re-define $\mu(t)$ to be any C^2 function that satisfies $\mu(t_0) = 0$ and $x(t) = x^*(t) + \alpha\mu(t)$ for any $\alpha \in \mathbb{R}$. We now consider the terminal condition into (a) $x(t_1)$ free. Since the value of $x(t_1)$ is unconstrained, so the perturbed function $x(t)$ is admissible whatever the value of $\mu(t_1)$.

From the equation (2), we have

$$I'(0) = \int_{t_0}^{t_1} \left[\frac{\partial F^*}{\partial x} - \frac{d}{dt}\left(\frac{\partial F^*}{\partial \dot{x}}\right)\right] \mu(t)dt + \left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} \mu(t_1) - \left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_0} \mu(t_0).$$

Since the Euler equation is a necessary condition for optimality and $\mu(t_0) = 0$, we simplify the expression of $I'(0)$ into

$$I'(0) = \left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} \mu(t_1).$$

Since we must have $I'(0) = 0$ for any value of $\mu(t_1)$, we obtain

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} = 0.$$

We finally consider the terminal condition (b) $x(t_1) \geq x_1$. For $x(t) = x^*(t) + \alpha\mu(t)$ to be admissible in Case (b), we have $\mu(t_0) = 0$ and $x^*(t_1) + \alpha\mu(t_1) \geq x_1$. Assume further that $x^*(t_1) > x_1$. Then, we can choose $|\mu(t_1)|$ and $|\alpha|$ small enough so that

$$|\mu(t_1)| \cdot |\alpha| < x^*(t_1) - x_1.$$

By the optimality of x^*, $I(\alpha)$ must have a local maximum at $\alpha = 0$ so that $I'(0) = 0$. Show the following transversality condition:

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} = 0.$$

Since the Euler equation holds and $\mu(t_0) = 0$, we obtain

$$I^{'}(0) = \left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} \mu(t_1) = 0.$$

Since $\mu(t_1)$ can be either positive or negative, we must satisfy

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} = 0. \blacksquare$$

Example 8.4.1 *Consider the following problem:*

$$\max \int_0^1 (1 - x^2 - \dot{x}^2)dt, \ x(0) = 1, \ \textit{with (a) } x(1) \textit{ free or (b) } x(1) \geq 2$$

Let $F(t, x, \dot{x}) = 1 - x^2 - \dot{x}^2$. *The Euler equation is*

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) = 0 \Leftrightarrow -2x + 2\ddot{x} = 0 \Leftrightarrow \ddot{x} - x = 0.$$

The characteristic equation of this differential equation is $r^2 - 1 = 0$. *So, we have* $r = 1, -1$. *The general solution is*

$$x(t) = Ae^t + Be^{-t}.$$

$$x(0) = 1 \Rightarrow A + B = 1 \Rightarrow B = 1 - A$$

Thus, an optimal solution of either problem must be of the form

$$x^*(t) = Ae^t + (1 - A)e^{-t}.$$

With (a) as the terminal condition, the transversality condition requires

$$\frac{\partial F^*}{\partial \dot{x}}\Big|_{t=t_1} = 0 \Rightarrow -2\dot{x}^*(1) = 0 \Rightarrow \dot{x}^*(1) = 0.$$

Since $\dot{x}^*(t) = Ae^t - (1 - A)e^{-t}$,

$$\dot{x}^*(1) = Ae - (1 - A)e^{-1} = 0 \Rightarrow A = \frac{1}{e^2 + 1}.$$

Hence,

$$x^*(t) = \frac{1}{e^2 + 1}\left(e^t + e^2 e^{-t}\right).$$

Because $F(t, x, \dot{x}) = 1 - x^2 - \dot{x}^2$ *is concave in* $(x, \dot{x})$, *the solution has been found. With (b) as the terminal condition, we require*

$$x^*(1) = Ae + (1 - A)e^{-1} \geq 2 \Rightarrow A \geq \frac{2e - 1}{e^2 - 1}$$

Suppose $x^*(1) > 2$. Then, as in (a), the transversality condition gives $A = 1/(e^2+1)$. But this violates the inequality $A \geq (2e-1)/(e^2-1)$ because $2e-1 > 1$. So,

$$x^*(1) = 2 \Rightarrow A = \frac{2e-1}{e^2-1}$$

Then,

$$\dot{x}^*(1) = Ae - (1-A)e^{-1} = \frac{2(e^2-e+1)}{e^2-1} > 0.$$

This implies

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=1} = -2\dot{x}^*(1) \leq 0.$$

Hence, the transversality condition holds. Then, the only solution candidate is

$$x^*(t) = \frac{1}{e^2-1}\left\{(2e-1)e^t + (e^2-2e)e^{-t}\right\}.$$

Because $F(t,x,\dot{x}) = 1 - x^2 - \dot{x}^2$ is concave in $(x,\dot{x})$, the solution has been found.

Example 8.4.2 (Optimal Dynamic Consumption) *Consider the following optimization problem of the decision maker (henceforth, DM). Let $A(t)$ denote the DM's wealth at time t, and let w be the (constant) income per unit of time. Suppose that the DM can borrow and save at the same constant rate of interest r. Consumption per unit of time at time t is then given by $C(t) = rA(t) + w - \dot{A}(t)$. Suppose the DM plans consumption from $t = 0$, until the expected time of death T so that the DM faces the following maximization problem:*

$$\begin{aligned} &\max_{\{C(t)\}_{t\in[0,T]}} \int_0^T U(C(t))e^{-\rho t}dt \\ &\text{subject to} \quad C(t) = rA(t) + w - \dot{A}(t),\ \forall t \in [0,T],\ A(0) = A_0,\ A(T) \geq A_T, \end{aligned}$$

where $U(\cdot)$ is the DM's utility function such that $U'(\cdot) > 0$ and $U''(\cdot) < 0$, $\rho > 0$ is a discount factor, and A_0 and A_T are given numbers.

The objective function is then given by $F(t,A,\dot{A}) = U(rA + w - \dot{A})e^{-\rho t}$. We compute the following:

$$\begin{aligned} \frac{\partial F}{\partial A} &= rU'(rA + w - \dot{A})e^{-\rho t}, \\ \frac{\partial F}{\partial \dot{A}} &= -U'(rA + w - \dot{A})e^{-\rho t}. \end{aligned}$$

We further compute the following:

$$\frac{d}{dt}\left(\frac{\partial F}{\partial \dot{A}}\right) = -U''(rA + w - \dot{A})(r\dot{A} - \ddot{A})e^{-\rho t} + \rho U'(rA + w - \dot{A})e^{-\rho t}$$

Using the short-hand notation that $U^{'} \equiv U^{'}(rA+w-\dot{A})$ *and* $U^{''} \equiv U^{''}(rA+w-\dot{A})$, *we obtain the Euler equation:*

$$\begin{aligned} & \frac{\partial F}{\partial A} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{A}}\right) = 0 \\ \Leftrightarrow \quad & rU^{'}e^{-\rho t} + U^{''}(r\dot{A} - \ddot{A})e^{-\rho t} - \rho U^{'}e^{-\rho t} = 0 \\ \Leftrightarrow \quad & (r-\rho)U^{'} + U^{''}(r\dot{A} - \ddot{A}) = 0 \\ \Leftrightarrow \quad & \ddot{A} - r\dot{A} + (\rho - r)\frac{U^{'}}{U^{''}} = 0. \end{aligned}$$

Let $A^*(t)$ *be an admissible function that satisfies the Euler equation and the corresponding transversality condition.*

Since we have $A(T) \geq A_T$ *as the terminal condition, the following transversality condition holds:*

$$\left(\frac{\partial F^*(t)}{\partial \dot{A}}\right)_{t=T} \leq 0 \ \text{ and } \ A^*(T) > A_T \Rightarrow \left(\frac{\partial F^*(t)}{\partial \dot{A}}\right)_{t=T} = 0.$$

Taking the contrapositive form of $A^*(T) > A_T \Rightarrow \left(\frac{\partial F^*(t)}{\partial \dot{A}}\right)_{t=T} = 0$, *we have*

$$\left(\frac{\partial F^*(t)}{\partial \dot{A}}\right)_{t=T} < 0 \Rightarrow A^*(T) = A_T$$

Because $U^{'} > 0$, *we have*

$$\left(\frac{\partial F^*}{\partial \dot{A}}\right)_{t=T} = -U^{'}(rA(T) + w - \dot{A}(T))e^{-\rho T} < 0.$$

This implies that $A^*(T) = A_T$.

We compute the following second derivatives of F:

$$\begin{aligned} \frac{\partial^2 F}{\partial A^2} &= r^2 U^{''}e^{-\rho t} < 0 \\ \frac{\partial^2 F}{\partial \dot{A}\partial A} = \frac{\partial^2 F}{\partial A \partial \dot{A}} &= -rU^{''}e^{-\rho t} > 0 \\ \frac{\partial^2 F}{\partial \dot{A}^2} &= U^{''}e^{-\rho t} < 0 \end{aligned}$$

We form the Hessian matrix $\mathcal{H}(A, \dot{A})$ *associated with* F:

$$\mathcal{H}(A, \dot{A}) = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} = \begin{pmatrix} r^2 U^{''}e^{-\rho t} & -rU^{''}e^{-\rho t} \\ -rU^{''}e^{-\rho t} & U^{''}e^{-\rho t} \end{pmatrix}.$$

Since $U^{''} < 0$, *we confirm that* $h_{11} < 0$, $h_{22} < 0$, *and* $h_{11}h_{22} - h_{12}h_{21} = 0$. *Thus,* $\mathcal{H}(A, \dot{A})$ *is negative semidefinite so that* $F(t, A, \dot{A})$ *is concave in* $(A, \dot{A})$. *By the sufficiency of the Euler equation, we conclude that* $A^*(t)$ *solves the problem.*

Exercise 8.4.1 *Solve the above example in the specific form of* $U(C) = \alpha - e^{-\beta C}$ *with* $\alpha, \beta > 0$.

Chapter 9

Control Theory

9.1 Control Theory: Basic Technique

We consider a control problem with no restrictions on the control variable and no restrictions on the terminal state.

Given the fixed time t_0 and t_1, our problem is

$$\max \int_{t_0}^{t_1} f(t, x(t), u(t))dt, \;\; u(t) \in (-\infty, \infty) \quad (*)$$

subject to

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0, \;\; x_0 \text{ fixed}, \; x(t_1) \text{ free.} \;\; (**)$$

A pair $(x(t), u(t))$ that satisfies $(**)$ is called an *admissible pair*. Among all admissible pairs we search for an *optimal pair*, a pair of functions that maximizes the integral in $(*)$.

Analogously to the Lagrange multiplier to each constraint, we associate a number $p(t)$, called the *co-state* variable, with the constraint $(**)$ for each $t \in [t_0, t_1]$. The resulting function $p(t)$ is called the *adjoint function* associated with the differential equation. We interpret $p(t)$ as the rental price for the use of one unit of capital.

Very much like the Lagrangian method, for each time $t \in [t_0, t_1]$ and each possible triple (x, u, p), of the state, control, and adjoint variables, we define what is called the *Hamiltonian* by

$$H(t, x, u, p) = f(t, x, u) + pg(t, x, u).$$

Theorem 9.1.1 (Necessity for the Maximum Principle) *Suppose that $(x^*(t), u^*(t))$ is an optimal pair for the problem $(*)$ subject to the constraints $(**)$. Then, there exists a continuous function $p(t)$ such that, for each $t \in [t_0, t_1]$,*

$$\begin{aligned}
&(i) \quad u^*(t) \in \arg \max_{u \in (-\infty, \infty)} H(t, x^*(t), u, p(t)), \\
&(ii) \quad \dot{p}(t) = -H_x^{'}(t, x^*(t), u^*(t), p(t)), \\
&(iii) \quad p(t_1) = 0 \;\; (\textit{the transversality condition}).
\end{aligned}$$

Proof: We rather provide a heuristic version of the proof. We setup the Lagrangian as follows:

$$\mathcal{L} = \int_{t_0}^{t_1} \{f(t, x(t), u(t)) - p(t)[\dot{x}(t) - g(t, x(t), u(t))]\}\, dt,$$

where $p(t)$ is considered the Lagrange multiplier, which changes over time.

By integration by parts, we obtain

$$\begin{aligned} \int_{t_0}^{t_1} p(t)\dot{x}(t)dt &= [p(t)x(t)]_{t_0}^{t_1} - \int_{t_0}^{t_1} \dot{p}(t)x(t)dt \\ &= p(t_1)x(t_1) - p(t_0)x(t_0) - \int_{t_0}^{t_1} \dot{p}(t)x(t)dt. \end{aligned}$$

Plugging this back into the Lagrangian, we obtain

$$\begin{aligned} \mathcal{L} = & \int_{t_0}^{t_1} [f(t, x(t), u(t)) + p(t)g(t, x(t), u(t)) + \dot{p}(t)x(t)]\, dt \\ & -p(t_1)x(t_1) + p(t_0)x(t_0). \end{aligned}$$

Assume that the optimal pair $(x^*(t), u^*(t))$ has been found and the Lagrangian is concave in (x, u) so that it is maximized at $(x^*(t), u^*(t))$.

Then, we compute the differential of the Lagrangian $d\mathcal{L}$. Given our hypothesis, we shall show $d\mathcal{L} \leq 0$. The proof that follows is a very simple version of the proof for $d\mathcal{L} \leq 0$.

Claim 9.1.1 $d\mathcal{L} \leq 0$.

A Heuristic Proof: Let $\mathcal{L} : [x_0, x_1] \to \mathbb{R}$ be a concave function and $x^* \in \arg\max_{x \in [x_0, x_1]} \mathcal{L}(x)$. We consider the following cases.

Case 1: $x^* \in (x_0, x_1)$

In this case, $\mathcal{L}^{'}(x^*) = 0$. So, $d\mathcal{L} = \mathcal{L}^{'}(x^*)dx = 0$ regardless of whether $dx > 0$ or $dx < 0$.

Case 2: $x^* = x_0$

In this case, $\mathcal{L}^{'}(x^*) \leq 0$ and $dx > 0$. So, $d\mathcal{L} = \mathcal{L}^{'}(x^*)dx \leq 0$.

Case 3: $x^* = x_1$

In this case, $\mathcal{L}^{'}(x^*) \geq 0$ and $dx < 0$. So, $d\mathcal{L} = \mathcal{L}^{'}(x^*)dx \leq 0$.

Summarizing all three cases, we conclude that $d\mathcal{L} \leq 0$. ■

So, by Leibniz's formula, we derive the expression for $d\mathcal{L}$:

$$\begin{aligned} d\mathcal{L} &= \int_{t_0}^{t_1} \left[\left(f_u^{'}(t, x^*(t), u^*(t)) + p(t) g_u^{'}(t, x^*(t), u^*(t)) \right) du(t) \right] dt \\ &+ \int_{t_0}^{t_1} \left[\left(f_x^{'}(t, x^*(t), u^*(t)) + p(t) g_x^{'}(t, x^*(t), u^*(t)) + \dot{p}(t) \right) dx(t) \right] dt \\ &- p(t_1) dx(t_1) + p(t_0) dx(t_0). \end{aligned}$$

Observe that $x(t_0) = x_0$ so that $dx(t_0) = 0$. So, we must have $d\mathcal{L} \leq 0$ for *any* $dx(t), du(t)$, and $dx(t_1)$. This is equivalent to the following conditions:

$$\begin{aligned} f_u^{'}(t, x^*(t), u^*(t)) + p(t) g_u^{'}(t, x^*(t), u^*(t)) &= 0, \\ f_x^{'}(t, x^*(t), u^*(t)) + p(t) g_x^{'}(t, x^*(t), u^*(t)) + \dot{p}(t) &= 0, \\ p(t_1) &= 0. \end{aligned}$$

Then, set $H(t, x(t), u(t), p(t)) = f(t, x(t), u(t)) + p(t) g(t, x(t), u(t))$. The above conditions are translated into:

$$\begin{aligned} H_u^{'}(t, x^*(t), u^*(t), p(t)) &= 0 \\ H_x^{'}(t, x^*(t), u^*(t), p(t)) &= -\dot{p}(t) \\ p(t_1) &= 0. \end{aligned}$$

This completes the Heuristic argument. ■

Theorem 9.1.2 (Sufficiency for the Maximum Principle) *Suppose that there exists an admissible pair of $(x^*(t), u^*(t))$ satisfying $(**)$ for which there exists a continuous function $p(t)$ such that, for each $t \in [t_0, t_1]$, conditions (i), (ii), and (iii) hold. In addition if $H(t, x, u, p(t))$ is concave in (x, u) for each $t \in [t_0, t_1]$, $(x^*(t), u^*(t))$ is optimal.*

Proof: We provide the proof later. ■

Example 9.1.1 *Consider the problem*

$$\max \int_0^T \left[1 - tx(t) - u(t)^2 \right] dt, \;\; \dot{x}(t) = u(t), \;\; x(0) = x_0, \;\; x(T) \text{ free}, \; u \in \mathbb{R}$$

where x_0 and T are given positive constants.

The Hamiltonian is

$$H(t, x, u, p) = 1 - tx - u^2 + pu.$$

Since $\partial^2 H / \partial u^2 = -2$, H is strictly concave in u. So,

$$u^*(t) \in \arg\max_{u \in \mathbb{R}} H(t, x^*(t), u, p(t)) \Leftrightarrow H_u^{'} = -2u + p(t) = 0.$$

Thus, $u^*(t) = p(t)/2$.

Because $H_x^{'} = -t$, *Condition (ii) reduces to* $\dot{p}(t) = t$ *and Condition (iii) reduces to* $p(T) = 0$.

$$\dot{p}(t) = t \Rightarrow p(t) = \frac{1}{2}t^2 + C,$$

where C *is a constant. Taking into account* $p(T) = T^2/2 + C = 0$, *we have*

$$p(t) = -\frac{1}{2}(T^2 - t^2) \Rightarrow u^*(t) = -\frac{1}{4}(T^2 - t^2).$$

$$\dot{x}^*(t) = u^*(t) = -\frac{1}{4}(T^2 - t^2) \Rightarrow x^*(t) = -\frac{1}{4}T^2 t + \frac{1}{12}t^3 + K,$$

where K *is a constant. Taking into account* $x(0) = x_0$, *we obtain*

$$x^*(t) = x_0 - \frac{1}{4}T^2 t + \frac{1}{12}t^3.$$

We have therefore found the only possible pair that can solve the problem. Because $H(t, x, u, p) = 1 - tx - u^2 + pu$ *is concave in* (x, u) *for each fixed* t, $(x^*(t), u^*(t))$ *is indeed optimal.*

9.2 The Standard Problem

We consider the *standard end-constrained problem.*

$$\max \int_{t_0}^{t_1} f(t, x(t), u(t))dt, \ u(t) \in U \subseteq \mathbb{R}, \ \forall t, \qquad (1)$$

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0 \qquad (2)$$

with one of the following conditions imposed

$$(a) \ x(t_1) = x_1, \ \ (b) \ x(t_1) \geq x_1, \ \text{ or (c) } x(t_1) \text{ free} \qquad (3)$$

where t_0, t_1, x_0, and x_1 are fixed numbers and U is the fixed control region.

Remark: U may be a closed set so that $u(t)$ takes the values of the boundary of U.

A pair $(x(t), u(t))$ that satisfies (2) and (3) with $u(t) \in U$ is called an *admissible (or feasible) pair*. Among all admissible pairs we seek an *optimal pair*, a pair of functions that maximizes the integral in (1).

We define the Hamiltonian as

$$H(t, x, u, p) = p_0 f(t, x, u) + pg(t, x, u) \qquad (4)$$

The new feature is the constant number p_0 in front of $f(t, x, u)$. If $p_0 \neq 0$, we can divide by p_0 to get a new Hamiltonian in which $p_0 = 1$, in effect. But if $p_0 = 0$, this normalization is impossible.

Theorem 9.2.1 (The Maximum Principle with Standard End Constraints) *Suppose that $(x^*(t), u^*(t))$ is an optimal pair for the standard end-constrained problem (1) - (3). Then, there exists a continuous function $p(t)$ and a number p_0, which is either 0 or 1, such that for all $t \in [t_0, t_1]$, we have $(p_0, p(t)) \neq (0, 0)$ and, moreover,*
(A) The control $u^*(t)$ maximizes $H(t, x^*(t), u, p(t))$ w.r.t. $u \in U$, i.e.,

$$H(t, x^*(t), u, p(t)) \leq H(t, x^*(t), u^*(t), p(t)) \text{ for all } u \in U$$

(B) $\dot{p}(t) = -H_x^{'}(t, x^*(t), u^*(t), p(t))$
*(C) Corresponding to each of the terminal conditions in (3) there is a **transversality condition** on $p(t_1)$:*
$(a^{'})$ $p(t_1)$ no condition
$(b^{'})$ $p(t_1) \geq 0$, with $p(t_1) = 0$ if $x^*(t_1) > x_1$
$(c^{'})$ $p(t_1) = 0$.

Proof: We omit the proof. ■

Theorem 9.2.2 (Mangasarian's Sufficiency for the Maximum Principle) *Suppose that $(x^*(t), u^*(t))$ is an admissible pair with a corresponding adjoint function $p(t)$ such that the conditions (A) - (C) are satisfied with $p_0 = 1$. Suppose further that the control region U is convex and that $H(t, x, u, p(t))$ is concave in (x, u) for every $t \in [t_0, t_1]$. Then, $(x^*(t), u^*(t))$ is an optimal pair.*

Remark: Almost all papers in economics literature that use control theory assume that the problem is "normal" in the sense that $p_0 = 1$.

Proof: We rather provide a heuristic version of the proof. Suppose that $(x, u) = (x(t), u(t))$ is an arbitrary alternative admissible pair. What we want is

$$D_u = \int_{t_0}^{t_1} f(t, x^*(t), u^*(t))dt - \int_{t_0}^{t_1} f(t, x(t), u(t))dt \geq 0.$$

First, simplify notation by writing H^* instead of $H(t, x^*(t), u^*(t), p(t))$ and H instead of $H(t, x(t), u(t), p(t))$.

Then, using the definition of the Hamiltonian and the fact that $\dot{x}^*(t) = g(t, x^*(t), u^*(t))$ and $\dot{x}(t) = g(t, x(t), u(t))$, we have $f^* = H^* - p\dot{x}^*$ and $f = H - p\dot{x}$. Therefore,

$$D_u = \int_{t_0}^{t_1} (H^* - H)dt + \int_{t_0}^{t_1} p(\dot{x} - \dot{x}^*)dt \qquad (*)$$

Because H is concave in (x, u), the first-order characterization of a concave function implies that

$$H - H^* \leq \frac{\partial H^*}{\partial x}(x - x^*) + \frac{\partial H^*}{\partial u}(u - u^*) \qquad (**)$$

Plugging $(**)$ into $(*)$, we obtain

$$\begin{aligned} D_u &\geq -\int_{t_0}^{t_1} \left[\frac{\partial H^*}{\partial x}(x - x^*) + \frac{\partial H^*}{\partial u}(u - u^*) \right] dt + \int_{t_0}^{t_1} p(\dot{x} - \dot{x}^*) dt \\ &= \int_{t_0}^{t_1} \left[\dot{p}(x - x^*) + p(\dot{x} - \dot{x}^*) \right] dt + \int_{t_0}^{t_1} \frac{\partial H^*}{\partial u}(u^* - u) dt, \end{aligned}$$

where the equality follows because $\dot{p} = -\partial H^*/\partial x$.

Assume $U = [u_0, u_1]$. Since the Hamiltonian is concave in u, Condition (A) is equivalent to

$$\frac{\partial H^*}{\partial u} \begin{cases} \leq 0 & \text{if } u^*(t) = u_0 \\ = 0 & \text{if } u^*(t) \in (u_0, u_1) \\ \geq 0 & \text{if } u^*(t) = u_1. \end{cases}$$

This can be further replaced by the equivalent inequality

$$\frac{\partial H^*}{\partial u}(u^*(t) - u) \geq 0 \quad \text{for all } u \in [u_0, u_1]$$

So,

$$\begin{aligned} D_u &\geq \int_{t_0}^{t_1} \left[\dot{p}(x - x^*) + p(\dot{x} - \dot{x}^*) \right] dt \\ &= \int_{t_0}^{t_1} \frac{d}{dt} \left[p(x - x^*) \right] dt \\ &= \left[p(t)(x(t) - x^*(t)) \right]_{t_0}^{t_1} \\ &= p(t_1)(x(t_1) - x^*(t_1)) - p(t_0)(x(t_0) - x^*(t_0)) \\ &= p(t_1)(x(t_1) - x^*(t_1)), \end{aligned}$$

where the last equality follows because $x(t_0) = x^*(t_0) = x_0$. It only remains to prove the claim below.

Claim 9.2.1 $p(t_1)(x(t_1) - x^*(t_1)) \geq 0 \qquad (***)$

Proof: In Case (a), we have $x(t_1) = x^*(t_1) = x_1$ so that $(***)$ holds. In Case (b), we have $x(t_1) \geq x_1$ and $x^*(t_1) \geq x_1$. If $x^*(t_1) > x_1$, by transversality condition (b'), we have $p(t_1) = 0$. So, $(***)$ holds. If $x^*(t_1) = x_1$, by (b'), we have $p(t_1) \geq 0$. Since $x(t_1) \geq x_1$ by (b'), $(***)$ holds. In Case (c), $x(t_1)$ is free. By transversality condition (c'), we have $p(t_1) = 0$ so that $(***)$ holds. This completes the proof of the claim. ■

This completes the proof of the theorem. ■

Example 9.2.1

$$\max \int_0^1 x(t) dt$$
$$\textit{subject to } \dot{x}(t) = x(t) + u(t),\ x(0) = 0,\ x(1) \textit{ free},\ u \in [-1, 1]$$

We want to have $x(t)$ as large as possible all the time, and from the differential equation, this is achieved by having u as large as possible. So, $u(t) = 1$ for all t and this must be the optimal control. Let us confirm this by using the maximum principle.

The Hamiltonian function with $p_0 = 1$ is

$$H(t, x, u, p) = x + px + pu,$$

which is linear and hence concave in (x, u). So, the sufficiency for the maximum principle applies. Then,

$$\left[H_x^{'} = -\dot{p} \Leftrightarrow \dot{p} = -1 - p\right], \quad p(1) = 0.$$

We solve $\dot{p} = -1 - p$:

$$\frac{dp}{dt} = -(1+p) \Leftrightarrow \frac{dp}{1+p} = -dt \Leftrightarrow \int \frac{dp}{1+p} = -\int dt \Leftrightarrow \ln(1+p) = -t + A$$

where A is a constant.

Since $p(1) = 0$, we have $A = 1$. Therefore, $p(t) = e^{1-t} - 1$. Due to the very expression for H, we must have $u^*(t) = 1$ for all $t \in [0, 1]$. By the differential equation $\dot{x}(t) = x(t) + u(t)$, $\dot{x}^*(t) = x^*(t) + 1$.

$$\frac{dx^*(t)}{dt} = x^*(t) + 1 \Leftrightarrow \int \frac{dx^*(t)}{x^*(t)+1} = \int dt \Leftrightarrow \ln(x^*(t)+1) = t + B,$$

where B is a constant. Since $x^*(0) = 0$, we have $B = 0$. So, $x^*(t) = e^t - 1$. We now see that $u^*(t), x^*(t)$, and $p(t)$ satisfy all the requirements in the sufficiency for the maximum principle.

Example 9.2.2 *Consider the following problem.*

$$\max \int_0^1 (2x - x^2)dt \ \text{ subject to } \ \dot{x} = u, \ x(0) = 0, \ x(1) = 0, \ u \in [-1, 1].$$

We set up the Hamiltonian:

$$H = 2x - x^2 + pu.$$

Only the term $p(t)u$ involves u in H. So, if we want to maximize the value of H with respect to $u \in [-1, 1]$, we obtain

$$u^*(t) = \begin{cases} 1 & \text{if } p(t) > 0 \\ -1 & \text{if } p(t) < 0, \ (*) \end{cases}$$

where $p(t)$ denotes the adjoint function associated with the Hamiltonian.

By the maximum principle, we have

$$\dot{p}(t) = -H_x^{'} \Rightarrow \dot{p}(t) = -2 + 2x.$$

So, we have

$$\dot{p}(t) = 2(x^*(t) - 1). \ (**)$$

Since $u^*(t) \leq 1$ *by* $(*)$ *and* $x^*(0) = 0$*, by the differential equation* $\dot{x}^*(t) = u^*(t)$*, we have*

$$x^*(t) = \int_0^t u^*(\tau)d\tau \leq \int_0^t d\tau = t.$$

Thus, $x^*(t) \leq t$ *for every* $t \in [0,1]$*. This implies that* $x^*(t) < 1$ *for all* $t \in [0,1)$*. By the differential equation* $\dot{p}(t) = 2(x^*(t) - 1)$*, we conclude that* $\dot{p}(t) < 0$ *for all* $t \in [0,1)$*. Thus,* $p(t)$ *is strictly decreasing over* $[0,1]$*.*

Claim 9.2.2 *There is no solution to the problem with* $p(1) \geq 0$*.*

Proof*: Suppose, on the contrary, that there is a solution with* $p(1) \geq 0$*. Since* $p(t)$ *is strictly decreasing, we have* $p(t) > 0$ *for all* $t \in [0,1)$*. By* $(*)$*, we also have* $u^*(t) = 1$ *for all* $t \in [0,1]$*. By the difference equation* $\dot{x}^*(t) = u^*(t)$*, we have* $\dot{x}^*(t) = 1$*. Once we solve this differential equation with* $x^*(0) = 0$*, we have* $x^*(t) = t$ *for all* $t \in [0,1]$*. However, we obtain* $x^*(1) = 1 \neq 0$*. This contradicts the terminal condition* $x^*(1) = 0$*, which completes the argument.* ■

Claim 9.2.3 *There is no solution to the problem such that* $p(t) < 0$ *for all* $t \in (0,1]$*.*

Proof*: Suppose, on the contrary, that there exists a solution such that* $p(t) < 0$ *for all* $t \in (0,1]$*. By* $(*)$*, we have* $u^*(t) = -1$ *for all* $t \in [0,1]$*. By the differential equation* $\dot{x}^*(t) = u^*(t) = -1$*, we have*

$$\frac{dx^*(t)}{dt} = -1 \Leftrightarrow x^*(t) = -\int dt = -t + B,$$

where B *is a constant. Since* $x^*(0) = 0$*, we have* $B = 0$*. Thus,* $x^*(t) = -t$*. However,* $x^*(1) = -1 \neq 0$*, which contradicts the terminal condition at* $t = 1$*. This completes the argument.* ■

Assume that there exists $t^* \in (0,1)$ *such that*

$$u^*(t) = \begin{cases} 1 & \text{if } t \in [0,t^*] \\ -1 & \text{if } t \in (t^*,1]. \end{cases}$$

Assume also that $x^*(\cdot)$ *and* $p(\cdot)$ *are continuous at* $t = t^*$*. We shall find a solution candidate. We have* $u^*(t) = 1$ *for any* $t \in [0,t^*]$*. By the differential equation* $\dot{x}^*(t) = u^*(t)$ *with* $x^*(0) = 0$*, we have* $x^*(t) = t$ *for all* $t \in [0,t^*]$*. We have* $u^*(t) = -1$ *for any* $t \in (t^*,1]$*. By the differential equation* $\dot{x}^*(t) = u^*(t)$*, we obtain*

$$x^*(t) = -t + C, \forall t \in (t^*,1],$$

where C *is a constant. Since* $x^*(\cdot)$ *is assumed to be continuous at* $t = t^*$*, we have* $C = 2t^*$*. Hence,* $x^*(t) = -t + 2t^*$ *for any* $t \in (t^*,1]$*. We also need to satisfy*

$x^*(1) = 0$. This implies that $t^* = 1/2$. We thus have $x^*(t) = -t + 1$ for any $t \in (1/2, 1]$.

Since $x^*(t) = t$ for any $t \in [0, 1/2]$, by $(*)$ and $(**)$, we have $\dot{p}(t) = 2(t-1)$ for any $t \in [0, 1/2]$. Then,

$$\frac{dp}{dt} = 2(t-1) \Leftrightarrow p(t) = \int 2(t-1)dt \Leftrightarrow p(t) = t^2 - 2t + D,$$

where D is a constant. By $(*)$, we must have $p(1/2) = 0$, which implies that $D = 3/4$. Hence, $p(t) = t^2 - 2t + 3/4$ for any $t \in [0, 1/2]$.

Since $x^*(t) = -t + 1$ for any $t \in (1/2, 1]$, by $(*)$ and $(**)$, we have $\dot{p}(t) = -2t$. Then,

$$\frac{dp}{dt} = -2t \Leftrightarrow p(t) = -\int 2tdt \Leftrightarrow p(t) = -t^2 + E,$$

where E is a constant. Since $p(\cdot)$ is assumed to be continuous at $t = 1/2$, $E = 1/4$. Hence, $p(t) = -t^2 + 1/4$.

The solution candidate we obtain is summarized as follows:

$$\begin{aligned} u^*(t) &= \begin{cases} 1 & \text{if } t \in [0, 1/2] \\ -1 & \text{if } t \in (1/2, 1]. \end{cases} \\ x^*(t) &= \begin{cases} t & \text{if } t \in [0, 1/2] \\ -t + 1 & \text{if } t \in (1/2, 1] \end{cases} \\ p(t) &= \begin{cases} t^2 - 2t + 3/4 & \text{if } t \in [0, 1/2] \\ -t^2 + 1/4 & \text{if } t \in (1/2, 1] \end{cases} \end{aligned}$$

Recall that we have $H = 2x - x^2 + pu$ as the Hamiltonian. Since $2x - x^2$ and pu are both concave functions, the sum of two concave functions is also a concave function. This implies that the Hamiltonian is a concave function. By the sufficiency for the maximum principle, we conclude that the solution candidate we found is indeed a solution to the original problem.

9.3 The Maximum Principle and the Calculus of Variations

Consider the standard variational problem:

$$\max \int_{t_0}^{t_1} F(t, x(t), \dot{x}(t))dt, \quad x(t_0) = x_0, \quad \begin{cases} \text{(a)} & x(t_1) = x_1 \\ \text{(b)} & x(t_1) \geq x_1 \\ \text{(c)} & x(t_1) \text{ free,} \end{cases} \quad (*)$$

where one of the alternative conditions (a), (b), and (c) is imposed.

To transform this to a control problem, simply use $u(t) = \dot{x}(t)$ as a control variable. Because there are no restrictions on $\dot{x}(t)$ in the calculus of variations problem, $U = \mathbb{R}$. The control problem has the particularly simple differential equation $\dot{x}(t) = u(t)$. The Hamiltonian is $H(t, x, u, p) = p_0 F(t, x, u) + pu$.

The maximum principle states that if $u^*(t)$ solves the problem, then H as a function of u must be maximized at $u = u^*(t)$.

Because $U = \mathbb{R}$, a necessary condition for this maximum is

$$H_u^{'}(t, x^*(t), u^*(t), p(t)) = p_0 F_u^{'}(t, x^*(t), u^*(t)) + p(t) = 0 \quad (*)$$

Since $(p_0, p(t)) \neq (0, 0)$, $(*)$ implies that $p_0 \neq 0$ so that $p_0 = 1$.

The differential equation for $p(t)$ is

$$\dot{p}(t) = -H_x^{'}(t, x^*(t), u^*(t), p(t)) = -F_x^{'}(t, x^*(t), u^*(t)) \quad (**)$$

Differentiating $(*)$ w.r.t. t yields

$$\frac{d}{dt}\left(F_u^{'}(t, x^*(t), u^*(t))\right) + \dot{p}(t) = 0 \qquad (***)$$

Since $u^* = \dot{x}^*$, it follows from $(**)$ and $(***)$ that

$$F_x^{'}(t, x^*, \dot{x}^*) - \frac{d}{dt}\left(F_{\dot{x}}^{'}(t, x^*, \dot{x}^*)\right) = 0,$$

which is the Euler equation.

Moreover, $(*)$ implies that

$$p(t) = -F_{\dot{x}}^{'}(t, x^*, \dot{x}^*).$$

Thus, we summarize the relationship between two approaches in the table below:

Terminal Conditions	**Transversality Conditions (Control Theory)**	**Transversality Conditions (Calculus of Variations)**
$x(t_1) = x_1$	$p(t_1)$ no condition	$F_{\dot{x}}^{'}(t_1, x^*, \dot{x}^*)$ no condition
$x(t_1) \geq x_1$	$p(t_1) \geq 0$ with $p(t_1) = 0$ if $x^*(t_1) > x_1$	$F_{\dot{x}}^{'}(t_1, x^*, \dot{x}^*) \leq 0$ with $F_{\dot{x}}^{'}(t_1, x^*, \dot{x}^*) = 0$ if $x^*(t_1) > x_1$
$x(t_1)$ free	$p(t_1) = 0$	$F_{\dot{x}}^{'}(t_1, x^*, \dot{x}^*) = 0$

Note that concavity of the Hamiltonian with respect to (x, u) is equivalent to concavity of $F(t, x, \dot{x})$ with respect to $(x, \dot{x})$.

Consider the standard end-constrained problem:

$$\max \int_{t_0}^{t_1} f(t, x(t), u(t))dt, \ u(t) \in U \subseteq \mathbb{R}, \ \forall t, \quad (1)$$

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0 \qquad (2)$$

with one of the following conditions imposed

$$(a)\ x(t_1) = x_1, \ \ (b)\ x(t_1) \geq x_1, \ \text{ or (c) } x(t_1) \text{ free} \quad (3)$$

where t_0, t_1, x_0, and x_1 are fixed numbers and U is the fixed control region.

9.4 Adjoint Variables as Shadow Prices

I consider the standard end-constrained problem. Suppose it has a unique solution $(x^*(t), u^*(t))$ with a unique corresponding adjoint function $p(t)$. The corresponding value of the objective function will depend on x_0, x_1, t_0, and t_1, so it is denoted by

$$V(x_0, x_1, t_0, t_1) = \int_{t_0}^{t_1} f(t, x^*(t), u^*(t))dt.$$

I call V the *value function*. (When $x(t_1)$ is free, x_1 is not an argument of V.) At any point where V is differentiable,

$$\frac{\partial V(x_0, x_1, t_0, t_1)}{\partial x_0} = p(t_0).$$

The number $p(t_0)$ therefore measures the marginal change in the value function as x_0 increases.

To illustrate the implications of this section, we revisit the following problem.

Example 9.4.1

$$\max \int_0^T \left[1 - tx(t) - u(t)^2\right] dt, \;\; \dot{x}(t) = u(t), \;\; x(0) = x_0, \;\; x(T) \text{ free}, \; u \in \mathbb{R}$$

where x_0 and T are given positive constants. The solution was

$$\begin{aligned} u^*(t) &= -\frac{1}{4}(T^2 - t^2), \\ x^*(t) &= x_0 - \frac{1}{4}T^2 t + \frac{1}{12}t^3, \\ p(t) &= -\frac{1}{2}(T^2 - t^2). \end{aligned}$$

So, the value function is

$$\begin{aligned} V(x_0, T) &= \int_0^T [1 - tx^*(t) - (u^*(t))^2]dt \\ &= \int_0^T \left[1 - x_0 t + \frac{1}{4}T^2 t^2 - \frac{1}{12}t^4 - \frac{1}{16}(T^2 - t^2)^2\right] dt \end{aligned}$$

By Leibniz' formula, we obtain

$$\frac{\partial V(x_0, T)}{\partial x_0} = \int_0^T (-t)dt = -\frac{1}{2}T^2 = p(0).$$

Define

$$H^*(t) = H(t, x^*(t), u^*(t), p(t)).$$

Provided V is differentiable, we have

$$\frac{\partial V}{\partial x_0} = p(t_0),\ \frac{\partial V}{\partial x_1} = -p(t_1),\ \frac{\partial V}{\partial t_0} = -H^*(t_0),\ \frac{\partial V}{\partial t_1} = H^*(t_1).$$

By Leibniz' rule, we have

$$\begin{aligned}\frac{\partial V}{\partial T} &= \left(1 - x_0 T + \frac{1}{4}T^4 - \frac{1}{12}T^4\right)\cdot 1 + \int_0^T \left[\frac{1}{2}t^2 T - \frac{1}{8}(T^2 - t^2)2T\right] dt \\ &= 1 - x_0 T + \frac{1}{6}T^4.\end{aligned}$$

Since $u^*(T) = 0$ and $x^*(T) = x_0 - T^3/6$, we have

$$H^*(T) = 1 - Tx^*(T) - (u^*(T))^2 + p(T)u^*(T) = 1 - x_0 T + \frac{1}{6}T^4 = \frac{\partial V}{\partial T}.$$

Example 9.4.2 (Economic Growth Model by Shell (1967)) *Consider the following economic growth model:*

$$\begin{aligned}&\max \int_0^T (1 - s(t))e^{\rho t} f(k(t))e^{-\delta t} dt \\ &\text{subject to}\quad \dot{k}(t) = s(t)e^{\rho t} f(k(t)) - \lambda k(t),\ k(0) = k_0,\ k(T) \geq k_T > k_0,\ s(t) \in [0,1],\end{aligned}$$

where $k(t)$ is the capital stock (a state variable), $s(t)$ is the savings rate (a control variable), and $f(k)$ is a production function. Assume that (i) $f(k) > 0$ whenever $k \geq k_0 e^{-\lambda T}$; (ii) $f'(k) > 0$; and (iii) $\rho, \delta, \lambda, T, k_0$, and k_T are all positive constants.

Let $(k^*(t), s^*(t))$ be a solution to the optimization problem. Using the maximum principle, characteirze $(k^*(t), s^*(t), p(t))$ where $p(t)$ is the adjoint function, to the extent possible. We set up the Hamiltonian as follows:

$$H(t, k, s, p) = (1 - s)e^{\rho t} f(k)e^{-\delta t} + p(se^{\rho t} f(k) - \lambda k).$$

By the maximum principle, we have

$$s^*(t) \in \arg\max_{s\in[0,1]} H(t, k^*(t), s, p) \Rightarrow s^*(t) \in \arg\max_{s\in[0,1]} se^{\rho t} f(k)\left[p - e^{-\delta t}\right].$$

So,

$$s^*(t) = \begin{cases} 1 & \text{if } p(t) > e^{-\delta t}, \\ [0,1] & \text{if } p(t) = e^{-\delta t}, \\ 0 & \text{if } p(t) < e^{-\delta t}. \end{cases}$$

By the maximum principle, we also have

$$\dot{p} = -\frac{\partial H(t, k^*(t), s^*(t), p(t)}{\partial k}.$$

Thus, we obtain

$$\dot{p} = -(1 - s^*(t))e^{\rho t} f'(k^*(t))e^{-\delta t} - p(t)s^*(t)e^{\rho t} f'(k^*(t)) + \lambda p(t).$$

By the transversality condition, we satisfy the following condition:

$$p(T) \geq 0 \text{ such that } p(T) = 0 \text{ if } k^*(T) > k_T.$$

We set $\rho = 0; f(k) = ak;\ a > 0;\ \delta = 0;$ *and* $\lambda = 0$. *Assume further that* $T > 1/a$ *and* $k_0 e^{aT} > k_T$.

Claim 9.4.1 $p(\cdot)$ *is strictly decreasing.*

Proof: *Given the specification of the parameters, we rewrite the optimization problem:*

$$\begin{aligned} &\max \int_0^T (1 - s(t))ak(t)dt \\ \text{subject to} \quad &\dot{k}(t) = as(t)k(t),\ k(0) = k_0,\ k(T) \geq k_T > k_0,\ s(t) \in [0,1], \end{aligned}$$

where $a > 0,\ T > 1/a,$ *and* $k_0 e^{aT} > k_T$. *The Hamiltonian can also be rewritten as follows:*

$$H(t, k, s, p) = (1 - s)ak + pask.$$

We can simplify the characterization of the solution $(k^*(t), s^*(t))$. *First, we obtain*

$$s^*(t) = \begin{cases} 1 & \text{if } p(t) > 1, \\ [0,1] & \text{if } p(t) = 1, \\ 0 & \text{if } p(t) < 1. \end{cases}$$

Second, using the above property of $s^*(t)$, *we obtain*

$$\begin{aligned} \dot{p} &= -(1 - s^*(t))a - p(t)s^*(t)a \\ &= -a + s^*(t)a(1 - p(t)) \\ &= \begin{cases} -ap(t) < 0 & \text{if } p(t) > 1 \\ -a < 0 & \text{if } p(t) \leq 1. \end{cases} \end{aligned}$$

As $a > 0$, *we have* $\dot{p}(t) < 0$ *for each* $t \in [0, T]$, *which implies that* $p(\cdot)$ *is strictly decreasing.* ■

Claim 9.4.2 $p(0) > 1$.

Proof: *Suppose, on the contrary, that* $p(0) \leq 1$. *Since* $p(t)$ *is strictly decreasing, we have* $p(t) < 1$ *for any* $t \in (0, T]$. *Since*

$$s^*(t) = \begin{cases} 1 & \text{if } p(t) > 1 \\ [0,1] & \text{if } p(t) = 1 \\ 0 & \text{if } p(t) < 1, \end{cases}$$

we have $s^*(t) = 0$ *for any* $t \in [0, T]$. *Since* $\dot{k}^*(t) = as^*(t)k^*(t)$ *and* $k^*(0) = k_0$, *we have* $k^*(t) = k_0$ *for any* $t \in [0, T]$. *In particular, we obtain* $k^*(T) = k_0$, *which*

contradicts the restriction that $k^*(T) = k_T > k_0$. *Thus, we conclude that* $p(0) > 1$. ■

In what follows, we divide our argument into two cases: $p(T) = 0$ *(Case I) and* $p(T) > 0$ *(Case II).*

Case I: $p(T) = 0$

Since $p(t)$ *is continuous and strictly decreasing such that* $p(0) > 1$ *and* $p(T) = 0$, *there is a unique* $t^* \in (0, T)$ *such that*

$$p(t) \begin{cases} > 1 & \text{if } t \in [0, t^*) \\ = 1 & \text{if } t = t^* \\ < 1 & \text{if } t \in (t^*, T]. \end{cases}$$

Then,

$$s^*(t) = \begin{cases} 1 & t \in [0, t^*) \\ [0, 1] & t = t^* \\ 0 & t \in (t^*, T]. \end{cases}$$

From the preceding claim, we have

$$\dot{p} = \begin{cases} -ap(t) < 0 & \text{if } p(t) > 1, \\ -a < 0 & \text{if } p(t) \leq 1. \end{cases}$$

So,

$$\dot{p}(t) = \begin{cases} -ap(t) & \text{if } t \in [0, t^*), \\ -a & \text{if } t \in [t^*, T]. \end{cases}$$

On the interval $[t^*, T]$, *by solving the differential equation* $\dot{p} = -a$, *we have* $p(t) = -at + A$, *where* A *is a constant. Taking into account that* $p(T) = 0$, *we have* $p(t) = a(T - t)$. *Since* $p(t^*) = 1$, *we have* $a(T - t^*) = 1$, *which implies that* $t^* = T - 1/a$, *which is positive by our assumption. We can also easily see that* $t^* < T$, *as* $t^* = T - 1/a$ *and* $a > 0$. *Hence,* $t^* = T - 1/a$ *is well-defined.*

On the interval $[0, t^*]$, *by solving the differential equation* $\dot{p} = -ap$, *we obtain* $p(t) = Be^{-at}$, *where* B *is a constant. Since* $p(t^*) = 1$, *we have* $p(t) = e^{-a(t-t^*)}$. *Since* $t^* = T - 1/a > 0$, *we obtain* $p(t) = e^{a(T-t)-1}$. *Therefore, the solution candidate is given as follows:*

$$\begin{aligned} s^*(t) &= \begin{cases} 1 & \text{if } t \in [0, T - 1/a] \\ 0 & \text{if } t \in (T - 1/a, T], \end{cases} \\ k^*(t) &= \begin{cases} k_0 e^{at} & \text{if } t \in [0, T - 1/a] \\ k_0 e^{aT-1} & \text{if } t \in (T - 1/a, T], \end{cases} \\ p(t) &= \begin{cases} e^{a(T-t)-1} & \text{if } t \in [0, T - 1/a] \\ -a(t - T) & \text{if } t \in (T - 1/a, T]. \end{cases} \end{aligned}$$

It remains to verify the terminal condition $k(T) \geq k_T$. This terminal condition reduces to $k_0 e^{aT} \geq k_T$, which is guaranteed by our assumption that $k_0 e^{aT} > k_T$.

Case II: $p(T) > 0$

In this case, we first establish the following claim.

Claim 9.4.3 *$p(T) > 0$ implies $p(T) < 1$.*

Proof*: Suppose, on the contrary, that $p(T) \geq 1$. Since $p(t)$ is strictly decreasing, $p(t) > 1$ for any $t \in [0, T)$. This implies that $s^*(t) = 1$ for any $t \in [0, T]$. So, we have $\dot{k}^*(t) = ak^*(t)$. Solving this differential equation together with $k^*(0) = k_0$, we obtain $k^*(t) = k_0 e^{at}$. Thus, we have $k^*(T) = k_0 e^{aT}$ which is smaller than k_T by our assumption. $p(T) > 0$ together with the transversality condition guarantees that $k^*(T) = k_T$. This is a contradiction.* ■

From the previous claim, we know that $p(T) < 1$. As we did for Case I: $p(T) = 0$, there is a unique $t^* \in (0, T)$ such that

$$p(t) \begin{cases} > 1 & \text{if } t \in [0, t^*) \\ = 1 & \text{if } t = t^* \\ < 1 & \text{if } t \in (t^*, T]. \end{cases}$$

Then,

$$s^*(t) = \begin{cases} 1 & t \in [0, t^*) \\ [0, 1] & t = t^* \\ 0 & t \in (t^*, T]. \end{cases}$$

From the preceding analysis, we have

$$\dot{p}(t) = \begin{cases} -ap(t) & \text{if } t \in [0, t^*), \\ -a & \text{if } t \in [t^*, T]. \end{cases}$$

Therefore, the solution candidate is given as follows:

$$\begin{aligned} s^*(t) &= \begin{cases} 1 & \text{if } t \in [0, t^*] \\ 0 & \text{if } t \in (t^*, T] \end{cases} \\ k^*(t) &= \begin{cases} k_0 e^{at} & \text{if } t \in [0, t^*] \\ k_0 e^{at^*} & \text{if } t \in (t^*, T] \end{cases} \\ p(t) &= \begin{cases} e^{a(t^* - t)} & \text{if } t \in [0, t^*] \\ 1 - a(t - t^*) & \text{if } t \in (t^*, T] \end{cases} \end{aligned}$$

It remains to identify t^*. From $k^*(T) = k_T$, it follows that $e^{at^*} = k_T/k_0$. So,

$$t^* = \frac{1}{a} \ln \left(\frac{k_T}{k_0} \right).$$

For t^* to be well-defined, we must have

$$\frac{1}{a}\ln\left(\frac{k_T}{k_0}\right) \leq T \Rightarrow k_T \leq k_0 e^{aT}.$$

This inequality is guaranteed by our assumption that $k_0 e^{aT} > k_T$. Since we assume $p(T) > 0$, which implies $1 - a(T - t^*) > 0$, which further implies

$$T - \frac{1}{a} < t^* = \frac{1}{a}\ln\left(\frac{k_T}{k_0}\right).$$

Considering Cases I and II together, there is only one solution candidate such that

$$s^*(t) = \begin{cases} 1 & \text{if } t \in [0, \bar{t}], \\ 0 & \text{if } t \in (\bar{t}, T], \end{cases}$$

where $\bar{t} = \max\{T - 1/a, (1/a)\ln(k_T/k_0)\}$.

For Case I, we obtain the value function:

$$V(k_0, k_T, T) = \int_{T-1/a}^{T} ak_0 e^{aT-1} dt = ak_0 e^{aT-1}[T - (T - 1/a)] = k_0 e^{aT-1}.$$

So, we obtain the following relationships:

$$\frac{\partial V}{\partial k_0} = e^{aT-1} = p(0); \ \frac{\partial V}{\partial k_T} = 0 = -p(T); \ \text{ and } \ \frac{\partial V}{\partial T} = ak_0 e^{aT-1} = H^*(T),$$

where $H^*(T) = H(T, k^*(T), s^*(T), p(T)) = (1 - s^*(T))ak^*(T) + p(T)as^*(T)k^*(T) = ak^*(T) = ak_0 e^{aT-1} = \partial V/\partial T$.

For Case II, we obtain the value function:

$$V(k_0, k_T, T) = \int_{t^*}^{T} ak_0 e^{at^*} dt = ak_0 e^{at^*}(T - t^*) = ak_T\left(T - \frac{1}{a}\ln k_T + \frac{1}{a}\ln k_0\right),$$

where $t^* = (1/a)\ln(k_T/k_0)$ and $k_T = k_0 e^{at^*}$. Since $p(0) = e^{at^*} = k_T/k_0$, we obtain

$$\frac{\partial V}{\partial k_0} = \frac{k_T}{k_0} = p(0).$$

Since $p(T) = 1 - a(T - t^*)$, we have

$$\frac{\partial V}{\partial k_T} = a\left(T - \frac{1}{a}\ln k_T + \frac{1}{a}\ln k_0\right) - 1 = a(T - t^*) - 1 = -P(T).$$

Finally, we obtain

$$\frac{\partial V}{\partial T} = ak_T = H^*(T),$$

where $H^*(T) = ak^*(T) = ak_0 e^{at^*} = ak_0(k_T/k_0) = ak_T$.

9.5 Sufficient Conditions

In many economic models, the Hamiltonian is not concave in (x, u). Arrow has suggested a weakening of this concavity condition. Define

$$\hat{H}(t, x, p) = \max_{u \in U} H(t, x, u, p),$$

assuming that the maximum value is attained. The function $\hat{H}(t, x, p)$ is called the *maximized Hamiltonian.*

Theorem 9.5.1 (Arrow's Sufficient Conditions) *Suppose that $(x^*(t), u^*(t))$ is an admissible pair in the standard end-constrained problem that satisfies all the requirements in the maximum principle, with $p(t)$ as the adjoint function, and with $p_0 = 1$. Suppose further that $\hat{H}(t, x, p(t))$ is concave in x for every $t \in [t_0, t_1]$. Then, $(x^*(t), u^*(t))$ solves the problem.*

Proof: Let $(x, u) = (x(t), u(t))_{t \in [t_0, t_1]}$ be an admissible pair. Define

$$D_u \equiv \int_{t_0}^{t_1} f(t, x^*(t), u^*(t))dt - \int_{t_0}^{t_1} f(t, x(t), u(t))dt.$$

By definition, we have

$$\begin{aligned} H^* &= f^* + p(t)\dot{x}^*(t), \\ H &= f + p(t)\dot{x}(t). \end{aligned}$$

Using the notation introduced above, we rewrite D_u as follows:

$$D_u = \int_{t_0}^{t_1} (H^* - H)dt + \int_{t_0}^{t_1} p(t)(\dot{x}(t) - \dot{x}^*(t))dt.$$

Using integration by parts, we compute the following:

$$\begin{aligned} & \int_{t_0}^{t_1} p(t)(\dot{x}(t) - \dot{x}^*(t))dt \\ = \; & [p(t)(x(t) - x^*(t))]_{t_0}^{t_1} - \int_{t_0}^{t_1} \dot{p}(t)(x(t) - x^*(t))dt \\ = \; & p(t_1)(x(t_1) - x^*(t_1)) - p(t_0)(x(t_0) - x^*(t_0)) - \int_{t_0}^{t_1} \dot{p}(t)(x(t) - x^*(t))dt \\ = \; & p(t_1)(x(t_1) - x^*(t_1)) - \int_{t_0}^{t_1} \dot{p}(t)(x(t) - x^*(t))dt \;\; (\because \;\; x(t_0) = x^*(t_0) = x_0) \end{aligned}$$

We next claim that $p(t_1)(x(t_1) - x^*(t_1)) \geq 0$. We show this by considering the following cases of the terminal condition: (a) $x(t_1) = x_1$; (b) $x(t_1) \geq x_1$; and (c) $x(t_1)$ free. In Case (a), by the transversality condition, we have $x(t_1) = x^*(t_1) = x_1$

so that $p(t_1)(x(t_1) - x^*(t_1)) = 0$. In Case (b), we have $x(t_1) \geq x_1$ and $x^*(t_1) \geq x_1$. If $x^*(t_1) > x_1$, by the transversality condition, we have $p(t_1) = 0$. Then, $p(t_1)(x(t_1) - x^*(t_1)) = 0$. If $x^*(t_1) = x_1$ instead, by the transversality condition, we have $p(t_1) \geq 0$. Since $x(t_1) \geq x_1$ and $x^*(t_1) = x_1$, we have $x(t_1) - x^*(t_1) \geq 0$ so that $p(t_1)(x(t_1) - x^*(t_1)) \geq 0$. In Case (c), since $x(t_1)$ is free, by the transversality condition, we have $p(t_1) = 0$ so that $p(t_1)(x(t_1) - x^*(t_1)) = 0$. This concludes that $p(t_1)(x(t_1) - x^*(t_1)) \geq 0$.

Therefore, we obtain

$$D_u \geq \int_{t_0}^{t_1} (H^* - H)dt - \int_{t_0}^{t_1} \dot{p}(t)(x(t) - x^*(t))dt.$$

By the definition of $\hat{H}$, we have

$$H^* = \hat{H}^* \text{ and } H \leq \hat{H}.$$

$H^* = \hat{H}^*$ and $H \leq \hat{H}$ implies that $\hat{H}^* - H \geq \hat{H}^* - \hat{H}$. Therefore, we obtain

$$D_u \geq \int_{t_0}^{t_1} \left[\hat{H}^* - \hat{H} - \dot{p}(t)(x(t) - x^*(t))\right] dt. \quad (*)$$

Since $\hat{H}$ is concave in x, we have

$$\hat{H} - \hat{H}^* \leq \frac{\partial \hat{H}^*}{\partial x}(x(t) - x^*(t)) \Rightarrow \hat{H}^* - \hat{H} \geq -\frac{\partial \hat{H}^*}{\partial x}(x(t) - x^*(t)).$$

By the maximum principle, we have

$$\dot{p}(t) = -\frac{\partial H^*}{\partial x} = -\frac{\partial \hat{H}^*}{\partial x},$$

where the second equality follows from $\hat{H}^* = H^*$. Using this, we obtain the following inequality:

$$\hat{H}^* - \hat{H} \geq \dot{p}(t)(x(t) - x^*(t)).$$

This implies that the integral on the right hand side of $(*)$ is nonnegative for all $t \in [t_0, t_1]$. So, $D_u \geq 0$. Thus, $(x^*(t), u^*(t))$ solves the problem. ■

Example 9.5.1 *Consider the problem*

$$\max \int_0^2 (u^2 - x)dt, \ \dot{x} = u, \ x(0) = 0, \ x(2) \textit{ free}, \ 0 \leq u \leq 1.$$

Let $H(t, x, u, p) = u^2 - x + pu$ be the Hamiltonian with $p_0 = 1$. We use the maximum principle to propose the solution candidate.

$$\dot{p} = -\partial H/\partial x = 1 \Rightarrow p(t) = t + A \underbrace{\Rightarrow}_{x(2) \textit{ free} \Rightarrow p(2)=0} p(t) = t - 2.$$

Now we have

$$H(t,x,u,p) = u^2 - x + u(t-2).$$

Since $\partial^2 H/\partial u^2 = 2$, $H(t,x,u,p)$ *is a strictly convex function of* u. *Noting* $u \in [0,1]$, *we have*

$$u^*(t) \in \arg\max H(t,x^*,u,p) \Rightarrow u^*(t) = 0, \quad or\ 1.$$

Since $H(t,x^*,0,p) = -x^*$ *and* $H(t,x^*,1,p) = -x^* + t - 1$, *we have*

$$u^*(t) = \begin{cases} 0 & if\ t \in [0,1] \\ 1 & if\ t \in (1,2]. \end{cases} \underbrace{\Rightarrow}_{\dot{x}=u} \dot{x}^*(t) = \begin{cases} 0 & if\ t \in [0,1] \\ 1 & if\ t \in (1,2] \end{cases}$$

So,

$$x^*(t) = \begin{cases} A & if\ t \in [0,1] \\ t + B & if\ t \in (1,2] \end{cases}$$

where A *and* B *are constants. Since* $x^*(0) = 0$, *we have* $A = 0$. *By the continuity of* x^*, *we must have*

$$x^*(1) = 0 = 1 + B \Rightarrow B = -1.$$

Thus,

$$x^*(t) = \begin{cases} 0 & if\ t \in [0,1], \\ t - 1 & if\ t \in (1,2]. \end{cases}$$

The maximized Hamiltonian is

$$\hat{H}(t,x,p(t)) = \max_{u \in [0,1]} u^2 - x + (t-2)u = \begin{cases} -x & if\ t \in [0,1], \\ -x + t - 1 & if\ t \in (1,2]. \end{cases}$$

For each $t \in [0,2]$, *the maximized Haimltonian is linear in* x, *hence concave in* x. *Therefore, by Arrow's sufficient condition, the solution candidate we found using the maximum principle is a solution to the problem.*

9.6 Variable Final Time

In the optimal control problems so far, the time interval has been fixed. Yet, for some control problems in economics, the final time is also a variable to be chosen optimally, along with the function $u(t),\ t \in [t_0, t_1]$. One such instance is the optimal extraction problem in Example 9.6.1 below, where it is natural to choose for how long to extract the resource, as well as how fast. Another example is the minimal time problem in which the objective is to steer a system from its initial state to a desired state as quickly as possible.

The variable final time problem can be formulated as follows:

$$\max_{u,t_1} \int_{t_0}^{t_1} f(t,x,u)dt \ \text{ subject to } \ \dot{x}(t) = g(t,x,u),\ x(t_0) = x_0, \begin{cases} (a) & x(t_1) = x_1 \\ (b) & x(t_1) \geq x_1 \\ (c) & x(t_1)\ \text{free} \end{cases} \qquad (*)$$

The only difference from the standard end-constrained problem is that t_1 can now be chosen. In contrast to the previous problems, the admissible control functions may be defined on different time intervals.

Theorem 9.6.1 (The Maximum Principle with Variable Final Time) *Let $(x^*(t), u^*(t))$ be an admissible pair defined on $[t_0, t_1^*]$ which solves problem $(*)$ with t_1 free ($t_1 \in (t_0, \infty)$). Then, all the conditions in the maximum principle (Theorem **??**) are satisfied on $[t_0, t_1^*]$, and, in addition*

$$H(t_1^*, x^*(t_1^*), u^*(t_1^*), p(t_1^*)) = 0 \quad (**).$$

Example 9.6.1 (Oil Extraction) *Let $x(t)$ denote the amount of oil in a reservoir at time t. Assume that at $t = 0$ the field contains K barrels of oil, so that $x(0) = K$. If $u(t)$ is the rate of extraction, we have*

$$x(t) = x(0) - \int_0^t u(\tau)d\tau = K - \int_0^t u(\tau)d\tau, \ \forall t \geq 0.$$

That is, the amount of oil left at time t is equal to the initial amount K, minus the total amount that has been extracted during the time span $[0, t]$, namely $\int_0^t u(\tau)d\tau$. Differentiating this equation gives

$$\dot{x}(t) = -u(t), \ \ x(0) = K \quad (*)$$

Suppose that the market price of oil at time t is known to be $q(t)$, so that the sales revenue per unit of time at t is $q(t)u(t)$. Assume further that the cost C per unit of time depends on t, x, and u, so that $C = C(t, x, u)$. The instantaneous profit per unit of time at time t is then

$$\pi(t, x(t), u(t)) = q(t)u(t) - C(t, x(t), u(t)).$$

If the discount rate is r, the total discounted profit over the interval $[0, T]$ is

$$\int_0^T \left[q(t)u(t) - C(t, x(t), u(t))\right] e^{-rt}dt.$$

It is natural to assume that $u(t) \geq 0$, and that $x(T) \geq 0$. We further assume that $C = C(t, u)$ is independent of x and convex in u, i.e., $C''_{uu} > 0$. Thus, the problem we consider is

$$\max_{u,T} \int_0^T \left[q(t)u(t) - C(t, u(t))\right] e^{-rt}dt \ \ s.t. \ \dot{x}(t) = -u(t), \ x(0) = K, \ x(T) \geq 0, \ u(t) \geq 0.$$

Suppose $(x^*(t), u^*(t))$, defined on $[0, T^*]$, solves this problem. The Hamiltonian with $p_0 = 1$ is

$$H(t, x, u, p) = \left[q(t)u(t) - C(t, u(t))\right] e^{-rt} + p(-u).$$

By the maximum principle, there exists a continuous function $p(t)$ such that

$$(i) \quad u^*(t) \in \arg\max_{u\geq 0} \left\{ [q(t)u(t) - C(t,u(t))]\, e^{-rt} - p(t)u \right\}$$

$$(ii) \quad \dot{p}(t) = -\frac{\partial H}{\partial x} = 0,\ p(T^*) \geq 0,\ \textit{with}\ p(T^*) = 0\ \textit{if}\ x^*(T^*) > 0$$

$$(iii) \quad [q(T^*)u^*(T^*) - C(T^*, u^*(T^*))]\, e^{-rT^*} = p(T^*)u^*(T^*)$$

Because $p(t)$ is continuous, it follows from (ii) that there exists a constant $\bar{p}$ such that $p(t) = \bar{p} \geq 0$.

Let

$$g(u) = [q(t)u(t) - C(t,u(t))]\, e^{-rt} - \bar{p}u.$$

Since $C(t,u)$ is convex in u and the other terms in $g(u)$ are linear in u, the function $g(u)$ is concave. According to (i), $u^*(t)$ maximizes $g(u)$ subject to $u \geq 0$. So, if $u^*(t) = 0$ (i.e., a boundary solution), then $g'(u^*(t)) = g'(0) \leq 0$. If $u^*(t) > 0$ (i.e., an interior solution), then $g'(u^*(t)) = 0$. Therefore, (i) implies

$$(iv) \quad \left[q(t) - C'_u(t, u^*(t))\right] e^{-rt} - \bar{p} \leq 0\ \ (= 0\ \textit{if}\ u^*(t) > 0)$$

Because $g(\cdot)$ is concave, Condition (iv) is also sufficient for (i) to hold.

At any time t where $u^*(t) > 0$, equation (iv) implies that

$$(v) \quad q(t) - C'_u(t, u^*(t)) = \bar{p}e^{rt}.$$

The left-hand side of equation (v) is the marginal profit from extraction, $\partial\pi/\partial u$.

Putting $t = T^*$ in (v), and using (iii), we deduce that, if $u^*(T^*) > 0$, then

$$(vi) \quad C'_u(T^*, u^*(T^*)) = \frac{C(T^*(, u^*(T^*))}{u^*(T^*)}.$$

This means that we terminate extraction at a time when the marginal cost of extraction is equal to average cost. If the problem has a solution with $u^*(t) > 0$, then (v) and (vi) both hold. If $C(T^*, 0) > 0$, then $u^*(T^*) > 0$, because $u^*(T^*) = 0$ contradicts (iii).

9.7 Current Value Formulations

Consider the standard end-constrained problem:

$$\max \int_{t_0}^{t_1} f(t,x,u)e^{-rt}dt,\ u \in U \subseteq \mathbb{R},\ \forall t,$$
$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0$$

with one of the following conditions imposed

$$(a)\ x(t_1) = x_1,\ \ (b)\ x(t_1) \geq x_1,\ \ \text{or (c)}\ x(t_1)\ \text{free} \qquad (1)$$

where t_0, t_1, x_0, and x_1 are fixed numbers and U is the fixed control region.

The ordinary Hamiltonian is $H = p_0 f(t,x,u)e^{-rt} + pg(t,x,u)$. We multiply it by e^{rt} to obtain the *current value Hamiltonian* H^c:

$$H^c = He^{rt} = p_0 f(t,x,u) + e^{rt} pg(t,x,u).$$

Introducing $\lambda = e^{rt}p$ as the *current value shadow price* for the problem, one can write H^c in the form (where we put $p_0 = \lambda_0$)

$$H^c(t,x,u,\lambda) = \lambda_0 f(t,x,u) + \lambda g(t,x,u).$$

$$\lambda = e^{rt}p \Rightarrow \dot{\lambda} = re^{rt}p + e^{rt}\dot{p} = r\lambda + e^{rt}\dot{p} \Rightarrow \dot{p} = e^{-rt}(\dot{\lambda} - r\lambda).$$

$$H^c = He^{rt} \Rightarrow \frac{\partial H^c}{\partial x} = e^{rt}\left(\frac{\partial H}{\partial x}\right)$$

So,

$$\dot{p} = -\frac{\partial H}{\partial x} \Rightarrow \dot{\lambda} - r\lambda = -\frac{\partial H^c}{\partial x}.$$

Theorem 9.7.1 (The Maximum Principle: Current Value Formulation) *Suppose that the admissible pair $(x^*(t), u^*(t))$ solves Problem (1) and let H^c be the current value Hamiltonian. Then, there exists a continuous function $\lambda(t)$ and a number λ_0, either 0 or 1, such that for all $t \in [t_0, t_1]$, we have $(\lambda_0, \lambda(t)) \neq (0,0)$, and:*
(A) $u = u^*(t)$ maximizes $H^c(t, x^*(t), u, \lambda(t))$ for $u \in U$
(B) $\dot{\lambda}(t) - r\lambda(t) = -\frac{\partial H^c(t,x^*(t),u^*(t),\lambda(t))}{\partial x}$
(C) The transversality conditions are: (a') $\lambda(t_1)$ no condition; (b') $\lambda(t_1) \geq 0$ and $x^*(t_1) > x_1 \Rightarrow \lambda(t_1) = 0$; and ($c'$) $\lambda(t_1) = 0$.

9.7.1 Sufficiency for Current Value Hamiltonian

The conditions in the maximum principle via current value formulation are sufficient for optimality if $\lambda_0 = 1$ and

Mangasarian: $H^c(t,x,u,\lambda(t))$ is concave in (x,u)
or (more generally)
Arrow: $\hat{H}^c(t,x,\lambda(t)) = \max_{u\in U} H^c(t,x,u,\lambda(t))$ is concave in x.

Example 9.7.1 *Consider the problem*

$$\max_{u\geq 0} \int_0^{20} (4K - u^2)e^{-0.25t}dt,\ \dot{K} = -0.25K + u,\ K(0) = K_0,\ K(20) \text{ free}$$

$K(t)$ is the value of a firm's capital stock, which depreciates at the constant proportional rate 0.25 per unit of the time, whereas $u(t)$ is gross investment, which costs $u(t)^2$ because the marginal cost of investment increases.

Profits $(4K - u^2)$ *are discounted at the constant proportional rate 0.25 per unit of the time. Let* $H^c = 4K - u^2 + \lambda(-0.25K + u)$ *with* $\lambda_0 = 1$. *So,*

$$\begin{aligned} \partial H^c/\partial u &= -2u + \lambda \\ \partial H^c/\partial K &= 4 - 0.25K. \end{aligned}$$

Assuming that $u^*(t) > 0$,

$$\partial (H^c)^*/\partial u = 0 \Rightarrow u^*(t) = 0.5\lambda(t).$$

The adjoint function λ *satisfies*

$$\dot{\lambda} - 0.25\lambda = -\partial (H^c)^*/\partial K = -4 + 0.25\lambda \text{ and } \underbrace{\lambda(20) = 0}_{\because K(20) \text{ free}}.$$

Using the formula for the first-order linear differential equations,

$$\dot{\lambda} - 0.5\lambda = -4 \Rightarrow \lambda = Ce^{0.5t} + \frac{-4}{-0.5} \underbrace{\Rightarrow}_{\lambda(20)=0} \lambda = 8(1 - e^{0.5t-10}).$$

and

$$u^*(t) = 0.5\lambda = 4(1 - e^{0.5t-10}).$$

Plugging $u = 4(1 - e^{0.5t-10})$ *into* $\dot{K} = -0.25K + u$, *we obtain*

$$\dot{K} + 0.25K = 4(1 - e^{0.5t-10}).$$

Using the formula for the first-order linear differential equations,

$$\begin{aligned} K^*(t) &= Ce^{-0.25t} + e^{-0.25t} \int e^{0.25t} 4(1 - e^{0.5t-10}) dt \\ &= Ce^{-0.25t} + 16 - \frac{16}{3} e^{0.5t-10} \end{aligned}$$

Noting $K^*(0) = K_0$, *we obtain*

$$K^*(t) = \left(K_0 - 16 + \frac{16}{3} e^{-10} \right) e^{-0.25t} + 16 - \frac{16}{3} e^{0.5t-10}.$$

To justify the solution we have obtained, it suffices to show that H^c *is concave in* (K, u). *We compute the Hessian matrix of* H^c:

$$H(K,u) = \begin{pmatrix} \partial^2 H^c/\partial K^2 & \partial^2 H^c/\partial u \partial K \\ \partial^2 H^c/\partial K \partial u & \partial^2 H^c/\partial u^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}.$$

Since $H(K, u)$ *is negative semidefinite for all* (K, u), H^c *is concave in* (K, u).

9.8 Scrap Values

In some economic optimization problems it is natural to include within the optimality criterion an additional function representing the value or utility associated with the terminal state. Consider the following problem

$$\max_{u(t)\in U}\left\{\int_{t_0}^{t_1} f(t,x(t),u(t))dt + S(x(t_1))\right\},\ \ \dot{x}(t)=g(t,x(t),u(t)),\ \ x(t_0)=x_0\ \ (*)$$

The function $S(x)$ is called a *scrap value function*, and I shall assume that it is C^1.

Suppose that $(x^*(t),u^*(t))$ solves this problem (with no additional condition on $x(t_1)$). Then, $(x^*(t),u^*(t))$ is indeed a solution to the corresponding problem with fixed terminal point $(t_1,x^*(t_1))$. For all admissible pairs in this new problem, the scrap value function $S(x^*(t_1))$ is constant. But then $(x^*(t),u^*(t))$ must satisfy all the conditions in the maximum principle, except the transversality conditions. We establish the following transversality condition for problem $(*)$.

Lemma 9.8.1 *The transversality condition for problem* $(*)$ *is*

$$p(t_1)=S^{'}(x^*(t_1))\ \ (**).$$

Remark: If $S(x)\equiv 0$, then $(**)$ reduces to $p(t_1)=0$, which is precisely as expected in a problem with no restrictions on $x(t_1)$. If $x(t)$ denotes the capital stock of a firm, then according to $(**)$, the shadow price of capital at the end of the planning period is equal to the marginal scrap value of the terminal stock.

Proof: Suppose that $(x(t),u(t))$ is an admissible pair for the problem $(*)$. Then,

$$\frac{d}{dt}S(x(t))=S^{'}(x(t))\dot{x}(t)=S^{'}(x(t))g(t,x(t),u(t)).$$

So, by integration,

$$S(x(t_1))-S(x(t_0))=\int_{t_0}^{t_1} S^{'}(x(t))g(t,x(t),u(t))dt.$$

Here $S(x(t_0))=S(x_0)$ is a constant. So, if the objective function in the problem $(*)$ is replaced by

$$\int_{t_0}^{t_1}\left[f(t,x(t),u(t))+S^{'}(x(t))g(t,x(t),u(t))\right]dt,$$

then the new problem is of a type studied previously with no scrap value, still with $x(t_1)$ free.

Let the Hamiltonian for this new problem be $H_1=f+S^{'}(x)g+qg=f+(q+S^{'}(x)g$, with adjoint variable q. An optimal pair $(x^*(t),u^*(t))$ for this problem must have the following properties:

$$(a) \quad u = u^*(t) \text{ maximizes } H_1(t, x^*(t), u, q(t)) \text{ for } u \in U$$
$$(b) \quad \dot{q}(t) = -\partial H_1^*/\partial x, \;\; q(t_1) = 0.$$

Define $p(t) = q(t) + S^{'}(x^*(t))$. It remains to show that if $H = f + pg$ is the ordinary Hamiltonian associated with the problem $(*)$, then $u^*(t))$ maximizes $H(t, x^*(t), u, p(t))$ for $u \in U$ and $\dot{p}(t) = -H_x^*$, with $p(t_1) = 0$. We omit this part of the proof. ■

Appropriate concavity conditions ensure optimality, as shown in the next theorem.

Theorem 9.8.1 (Sufficient Conditions with Scrap Value) *Suppose $(x^*(t), u^*(t))$ is an admissible pair for the scrap value problem $(*)$ and suppose there exists a continuous $p(t)$ such that, for all $t \in [t_0, t_1]$,*

$$(A) \quad u = u^*(t) \textit{ maximizes } H(t, x^*(t), u, p(t)) \textit{ w.r.t. } u \in U$$
$$(B) \quad \dot{p}(t) = -H_x^{'}(t, x^*(t), u^*(t), p(t)), \;\; p(t_1) = S^{'}(x^*(t_1))$$
$$(C) \quad H(t, x, u, p(t)) \textit{ is concave in } (x, u) \textit{ and } S(x) \textit{ is concave.}$$

Then, $(x^*(t), u^*(t))$ solves the problem.

Proof: Suppose that $(x, u) = (x(t), u(t))$ is an arbitrary admissible pair. We must show that

$$D_u = \int_{t_0}^{t_1} f(t, x^*(t), u^*(t))dt + S(x^*(t_1)) - \int_{t_0}^{t_1} f(t, x(t), u(t))dt - S(x(t_1)) \geq 0.$$

Because $S(x)$ is C^1 and concave, by the first-order chracterization of a concave function (Theorem 4.6.3), we have

$$S(x(t_1)) - S(x^*(t_1)) \leq S^{'}(x^*(t_1))((x(t_1) - x^*(t_1)).$$

Combining this with the inequality

$$\int_{t_0}^{t_1} (f^* - f)dt \geq p(t_1)(x(t_1) - x^*(t_1)),$$

which was derived in the proof of Theorem **??**, we get

$$D_u \geq \left[p(t_1) - S^{'}(x^*(t_1))\right](x(t_1) - x^*(t_1)) = o,$$

where the last equality follows from (B). So, $D_u \geq 0$. This completes the proof. ■

Example 9.8.1 *Consider the following problem:*

$$\max_{u\in\mathbb{R}}\left\{\int_0^1 -\frac{1}{2}u^2dt+\sqrt{x(1)}\right\} \text{ subject to } \dot{x}=x+u,\ x(0)=0,\ x(1) \text{ free.}$$

Let

$$\begin{aligned} H(t,x,u,p) &= -u^2/2+p(x+u) \\ S(x) &= \sqrt{x}=x^{1/2}. \end{aligned}$$

Since $u\in\mathbb{R}$, maximizing H with respect to u, which is due to the maximum principle, requires that $H_u^{'}=0$. This implies that $H_u^{'}=-u+p=0$ so that $u=p$. By the maximum principle, we also have $\dot{p}=-H_x^{'}$ so that $\dot{p}=-p$. This implies that $p(t)=Ae^{-t}$, where A is a constant. Since we have the differential equation $\dot{x}=x+u$, $u=p$, and $p(t)=Ae^{-t}$, we have

$$\dot{x}-x=Ae^{-t}.$$

Using the formula of first-order linear equations, we have

$$x(t)=Be^t+e^t\int e^{-t}Ae^{-t}dt=Be^t+Ae^t\int e^{-2t}dt=Be^t-\frac{A}{2}e^{-t}.$$

Since $x(0)=0$, we have

$$0=B-\frac{A}{2}\Rightarrow B=\frac{A}{2}.$$

Therefore,

$$x(t)=\frac{A}{2}(e^t-e^{-t}).$$

We compute

$$S^{'}(x)=\frac{1}{2}x^{-1/2}.$$

By the transversality condition $p(1)=S^{'}(x(1))$ and $p(t)=Ae^{-t}$,

$$Ae^{-1}=\frac{1}{2}(x(1))^{-1/2}\Rightarrow Ae^{-1}=\frac{1}{2}\left\{\frac{A}{2}(e-e^{-1})\right\}^{-1/2}.$$

We solve this for A as follows:

$$\begin{aligned} & 4A^2e^{-2}=\left\{\frac{A}{2}(e-e^{-1})\right\}^{-1} \\ \Rightarrow\ & 4A^2e^{-2}\left\{\frac{A}{2}(e-e^{-1})\right\}=1 \\ \Rightarrow\ & 2A^3(e^{-1}-e^{-3})=1 \\ \Rightarrow\ & A=[2(e^{-1}-e^{-3})]^{-1/3} \\ \Rightarrow\ & A=[2e^{-3}(e^2-1)]^{-1/3} \\ \Rightarrow\ & A=e[2(e^2-1)]^{-1/3}. \end{aligned}$$

Thus, we obtain the following solution candidate:

$$\begin{aligned} u^*(t) &= Ae^{-t} \\ p(t) &= Ae^{-t} \\ x^*(t) &= \frac{A}{2}(e^t - e^{-t}), \end{aligned}$$

where $A = e[2(e^2-1)]^{-1/3}$. *Since the Hamiltonian consists of* $-u^2/2$ *and* $p(x+u)$, *it is the sum of concave functions of* u *and* x. *Moreover, since the scrap function* $S(x)$ *is strictly concave in* x, *the solution candidate we have obtained is indeed the solution.*

9.8.1 Current Value Formulation

Many control problems in economics have the following structure:

$$\max_{u\in U\subseteq\mathbb{R}} \left\{ \int_{t_0}^{t_1} f(t,x,u)e^{-rt}dt + S(x(t_1))e^{-rt_1} \right\}, \ \dot{x} = g(t,x,u), \ x(t_0) = x_0 \ (*)$$

$$(a) \ x(t_1) = x_1, \ \ (b) \ x(t_1) \geq x_1, \ \text{or} \ \ (c) \ x(t_1) \ \text{free} \ (**)$$

The current value Hamiltonian for the problem is

$$H^c(t,x,u,\lambda) = \lambda_0 f(t,x,u) + \lambda g(t,x,u).$$

Theorem 9.8.2 (Current Value Maximum Principle: Scrap Values) *Suppose that the admissible pair* $(x^*(t), u^*(t))$ *solves problem* $(*)$ *and* $(**)$. *Then, there exist a continuous function* $\lambda(t)$ *and a number* λ_0, *either* 0 *or* 1, *such that, for all* $t \in [t_0, t_1]$, *we have* $(\lambda_0, \lambda(t)) \neq (0,0)$, *and:*

(A) $u = u^*(t)$ *maximizes* $H^c(t, x^*(t), u, \lambda(t))$ *for* $u \in U$

(B) $\dot{\lambda}(t) - r\lambda(t) = -\dfrac{\partial H^c(t, x^*(t), u^*(t), \lambda(t))}{\partial x}$ *whenever* $u^*(t)$ *is continuous*

(C) *The transversality conditions are* :

$$\begin{aligned} &(a') \quad \lambda(t_1) \text{ no condition} \\ &(b') \quad \lambda(t_1) \begin{cases} = \lambda_0 S'(x^*(t_1)) & \text{if } x^*(t_1) > x_1 \\ \geq \lambda_0 S'(x^*(t_1)) & \text{otherwise} \end{cases} \\ &(c') \quad \lambda(t_1) = \lambda_0 S'(x^*(t_1)). \end{aligned}$$

Theorem 9.8.3 (Sufficient Conditions) *The conditions in Theorem 9.8.2 with* $\lambda_0 = 1$ *are sufficient if* U *is convex,* $H^c(t,x,u,\lambda(t))$ *is concave in* (x,u), *and* $S(x)$ *is concave in* x.

Proof The proof is a straightforward extension of Theorem 9.8.1. So, we omit the proof. ■

Example 9.8.2 *Consider the following problem:*

$$\max_{u\in\mathbb{R}}\left\{\int_0^T (x-u^2)e^{-0.1t}dt + ax(T)e^{-0.1T}\right\} \text{ subject to } \dot{x} = -0.4x+u,\ x(0)=1,\ x(T) \text{ free},$$

where a is a positive constant.

We formulate the current value Hamiltonian with $\lambda_0 = 1$:

$$H^c(t,x,u,\lambda) = x - u^2 + \lambda(-0.4x+u).$$

Since H^c consists of the sum of $-u^2+\lambda u$ and $(1-0.4\lambda)x$ and $-u^2+\lambda u$ is a concave function in u and $(1-0.4\lambda)x$ is a concave function in x, H^c is concave in (x,u). Moreover, $S(x) = ax$ is linear so that it is concave in x. Therefore, the conditions in the maximum principle are sufficient.

Because H^c is concave in u and $u\in\mathbb{R}$, the maximum of H^c with respect to u is characterized as its first-order condition:

$$(1)\quad \frac{\partial H^c(t,x^*(t),u^*(t),\lambda(t))}{\partial u} = -2u^*(t)+\lambda(t) = 0.$$

By the maximum principle, we also have the following condition: $\dot{\lambda} - r\lambda = -\partial H^c/\partial x$. So, we have

$$\dot{\lambda}(t) - 0.1\lambda(t) = -\partial H^c/\partial x = -1 + 0.4\lambda(t).$$

The above differential equation can be rewritten as follows:

$$\dot{\lambda} - 0.5\lambda = -1.$$

By the formula of first-order linear differential equations, we obtain

$$(2)\quad \lambda(t) = Ce^{0.5t} + \frac{-1}{-0.5} = Ce^{0.5t} + 2,$$

where C is a constant.

Since $x(T)$ is free and $S(x) = ax$, by the transversality condition $\lambda(t_1) = \lambda_0 S'(x^*(t_1))$, we have

$$(3)\quad \lambda(T) = a.$$

Combining (2) and (3) together, we have

$$\lambda(t) = (a-2)e^{-0.5(T-t)} + 2.$$

It follows from (1) that $u^*(t) = \lambda(t)/2$. Plugging $u^*(t) = \lambda(t)/2$ into the differential equation $\dot{x} = -0.4x+u$,

$$\dot{x} + 0.4x = \lambda(t)/2$$

By the formula of first-order linear differential equations, we have

$$x^*(t) = Ce^{-0.4t} + e^{-0.4t}\int e^{0.4t}\lambda(t)/2dt,$$

where C is a constant. Since $\lambda(t) = (a-2)e^{-0.5(T-t)} + 2$, we solve the above differential equation as follows:

$$\begin{aligned} x^*(t) &= Ce^{-0.4t} + \frac{e^{-0.4t}}{2}\int \left\{(a-2)e^{-0.5T+0.9t} + 2e^{0.4t}\right\} dt \\ &= Ce^{-0.4t} + \frac{(a-2)e^{-0.4t-0.5T}}{2}\int e^{0.9t}dt + e^{-0.4t}\int e^{0.4t}dt \\ &= Ce^{-0.4t} + \frac{(a-2)e^{-0.4t-0.5T}}{2}\cdot\frac{e^{0.9t}}{0.9} + e^{-0.4t}\frac{e^{0.4t}}{0.4} \\ &= Ce^{-0.4t} + \frac{5(a-2)}{9}e^{-0.5(T-t)} + \frac{5}{2}. \end{aligned}$$

Since $x^*(0) = 0$, we can pin down the value of C:

$$C = -\frac{5(a-2)}{9}e^{-0.5T} - \frac{5}{2}.$$

Thus, the solution to this problem is obtained:

$$\begin{aligned} \lambda(t) &= (a-2)e^{-0.5(T-t)} + 2 \\ u^*(t) &= \lambda(t)/2 \\ x^*(t) &= Ce^{-0.4t} + \frac{5(a-2)}{9}e^{-0.5(T-t)} + \frac{5}{2}, \end{aligned}$$

where

$$C = -\frac{5(a-2)}{9}e^{-0.5T} - \frac{5}{2}.$$

9.9 Infinite Horizon

A typical infinite horizon optimal control problem in economics takes the following form:

$$\begin{aligned} &\max \int_{t_0}^{\infty} f(t, x(t), u(t))e^{-rt}dt, \\ &\text{subject to} \quad \dot{x}(t) = g(t, x(t), u(t)),\ x(t_0) = x_0,\ u(t) \in U \quad (1) \end{aligned}$$

Many problems do impose the constraint

$$\lim_{t\to\infty} x(t) \geq x_1 \quad (x_1 \text{ is a fixed number}) \qquad (2)$$

The pair $(x(t), u(t))$ is *admissible* if it satisfies $\dot{x}(t) = g(t, x(t), u(t))$, $x(t_0) = x_0$, $u(t) \in U$, along with (2) when that is imposed.

Theorem 9.9.1 (Sufficient Conditions with an Infinite Horizon) *Suppose that an admissible pair $(x^*(t), u^*(t))$ for problem (1), with or without terminal condition (2), satisfies the following conditions for some $\lambda(t)$ for all $t \geq t_0$, with $\lambda_0 = 1$:*

(a) $u^*(t)$ maximizes $H^c(t, x^*(t), u, \lambda(t))$ with respect to $u \in U$;
(b) $\dot{\lambda}(t) - r\lambda = -\partial H^c(t, x^*(t), u^*(t), \lambda(t))/\partial x$;
(c) $H^c(t, x, u, \lambda(t))$ is concave in (x, u);
(d) $\lim_{t\to\infty} \lambda(t)e^{-rt}[x(t) - x^*(t)] \geq 0$ for all admissible $x(t)$.

Then, $(x^*(t), u^*(t))$ is a solution to Problem (1).

Proof: For any admissible pair $(x(t), u(t))$ and for all $t \geq t_0$, define

$$\begin{aligned} D_u(t) &= \int_{t_0}^{t} \underbrace{f(\tau, x^*(\tau), u^*(\tau))}_{=f^*} e^{-r\tau} d\tau - \int_{t_0}^{t} \underbrace{f(\tau, x(\tau), u(\tau))}_{=f} e^{-r\tau} d\tau \\ &= \int_{t_0}^{t} (f^* - f) e^{-r\tau} d\tau \end{aligned}$$

Using the similar simplified notation,

$$\begin{aligned} f^* &= (H^c)^* - \lambda g^* \underbrace{=}_{\dot{x}^* = g^*} (H^c)^* - \lambda \dot{x}^*, \\ f &= H^c - \lambda g \underbrace{=}_{\dot{x} = g} H^c - \lambda \dot{x}. \end{aligned}$$

So,

$$D_u(t) = \int_{t_0}^{t} \left[(H^c)^* - H^c\right] e^{-r\tau} d\tau + \int_{t_0}^{t} \lambda e^{-r\tau} (\dot{x} - \dot{x}^*) d\tau.$$

By concavity of H^c with respect to (x, u), one has

$$H^c - (H^c)^* \leq \frac{\partial (H^c)^*}{\partial x}(x - x^*) + \frac{\partial (H^c)^*}{\partial u}(u - u^*).$$

This is equivalent to

$$\begin{aligned} (H^c)^* - H^c &\geq -\frac{\partial (H^c)^*}{\partial x}(x - x^*) + \frac{\partial (H^c)^*}{\partial u}(u^* - u) \\ &= (\dot{\lambda} - r\lambda)(x - x^*) + \frac{\partial (H^c)^*}{\partial u}(u^* - u). \end{aligned}$$

So,

$$D_u(t) \geq \int_{t_0}^{t} e^{-r\tau} \left[(\dot{\lambda} - r\lambda)(x - x^*) + \lambda(\dot{x} - \dot{x}^*)\right] d\tau + \int_{t_0}^{t} \frac{\partial (H^c)^*}{\partial u}(u^* - u) e^{-r\tau} d\tau$$

As we have already shown elsewhere,

$$(a) \Rightarrow \frac{\partial (H^c)^*}{\partial u}(u^* - u) \geq 0 \Rightarrow \int_{t_0}^{t} \frac{\partial (H^c)^*}{\partial u}(u^* - u) e^{-r\tau} d\tau \geq 0.$$

Thus,

$$\begin{aligned} D_u(t) &\geq \int_{t_0}^{t} \frac{d}{d\tau}\left[e^{-r\tau}\lambda(\tau)(x(\tau)-x^*(\tau))\right]d\tau \\ &= \left[e^{-r\tau}\lambda(\tau)(x(\tau)-x^*(\tau))\right]_{t_0}^{t} \\ &= e^{-rt}\lambda(t)(x(t)-x^*(t)) \ \ (\because \ x(t_0)=x^*(t_0)=x_0) \end{aligned}$$

Then,

$$D_u(\infty) \geq \lim_{t\to\infty} e^{-rt}\lambda(t)(x(t)-x^*(t)) \underbrace{\geq}_{\because (d)} 0$$

Noting

$$D_u(\infty) = \int_{t_0}^{\infty} (f^* - f)e^{-r\tau}d\tau \geq 0,$$

we conclude that $(x^*(t), u^*(t))$ is a solution to Problem (1). ■

Lemma 9.9.1 (A condition guaranteeing (d)) *Let $(x^*(t), u^*(t))$ be an admissible pair for Problem (1) satisfying (a), (b), and (c) in the sufficiency result. Suppose that $\lim_{t\to\infty} x(t) \geq x_1$ for any admissible $x(t)$. Assume further that the following three conditions hold:*

(A) $\lim_{t\to\infty} \lambda(t)e^{-rt}(x_1 - x^*(t)) \geq 0$;
(B) $\exists M \in \mathbb{R}_+$ such that $|\lambda(t)e^{-rt}| \leq M$ for all $t \geq t_0$;
(C) $\exists t' \in \mathbb{R}$ such that $\lambda(t) \geq 0$ for all $t \geq t'$.

Then, (d) $\lim_{t\to\infty} \lambda(t)e^{-rt}[x(t) - x^*(t)] \geq 0$ for all admissible $x(t)$.

Proof: We define $\xi(t)$ and $\zeta(t)$ as follows:

$$\begin{aligned} \xi(t) &\equiv \lambda(t)e^{-rt}(x(t) - x_1), \\ \zeta(t) &\equiv \lambda(t)e^{-rt}(x_1 - x^*(t)). \end{aligned}$$

By construction, we also have $\xi(t) + \zeta(t) = \lambda(t)e^{-rt}[x(t) - x^*(t)]$. Then, satisfying Condition (d) boils down to proving $\lim_{t\to\infty} \xi(t)+\zeta(t) \geq 0$. Because of (A), it suffices to show that $\lim_{t\to\infty} \xi(t) \geq 0$. Since $\lim_{t\to\infty} x(t) \geq x_1$, our argument is reduced to considering the following two cases: $\lim_{t\to\infty} x(t) = x_1$ or $\lim_{t\to\infty} x(t) > x_1$. So, if the former case applies, because of (B), we have $\lim_{t\to\infty} \xi(t) = 0$. If the latter case applies, there exists t' large enough so that $x(t) > x_1$ for all $t > t'$. Then, because of (C), we have $\lambda(t)e^{-rt}(x(t) - x_1)$ tends to a number, which is greater than or equal to 0, as $t \to \infty$. So, in this case, we also have $\lim_{t\to\infty} \xi(t) \geq 0$. We thus conclude that if (A) through (C) are all satisfied, then (d) holds. ■

Remark: If it is additionally required that $x(t) \geq x_1$ for all t, then it suffices to check conditions (A) and (C). This result is referred to as the *Malinvaud transversality condition*.

Example 9.9.1 *Consider the problem*

$$\max \int_0^\infty -u^2 e^{-rt} dt,\ \dot{x} = ue^{-at},\ x(0) = 0,\ \lim_{t\to\infty} x(t) \geq K,\ u \in \mathbb{R},$$

where r, a, and K are positive constants with $a > r/2$.

Set $H^c = -u^2 + \lambda u e^{-at}$ as the current value Hamiltonian. We compute the Hessian matrix of H^c:

$$H(x,u) = \begin{pmatrix} \partial^2 H^c/\partial x^2 & \partial^2 H^c/\partial u \partial x \\ \partial^2 H^c/\partial x \partial u & \partial^2 H^c/\partial u^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}.$$

Since $H(x,u)$ is negative semidefinite for all (x,u), H^c is concave in (x,u). We find

$$\begin{aligned} \frac{\partial H^c}{\partial x} &= 0, \\ \frac{\partial H^c}{\partial u} &= -2u + \lambda e^{-at}. \end{aligned}$$

Since every $u \in \mathbb{R}$ is an interior point,

$$u^*(t) \in \arg\max H^c(t, x^*(t), u, \lambda(t)) \Rightarrow u^*(t) = \frac{1}{2}\lambda e^{-at}.$$

Next, since the differential equation below is separable, we obtain

$$\dot{\lambda} - r\lambda = -\partial H^c/\partial x = 0 \Rightarrow \lambda(t) = Ae^{rt},$$

where A is a constant. Thus,

$$u^*(t) = \frac{1}{2} A e^{(r-a)t}.$$

Since $\lambda(t) = Ae^{rt}$, we have

$$\lambda(t)e^{-rt} = A \Rightarrow |\lambda(t)e^{-rt}| \leq A \ \textit{for all } t \geq t_0.$$

Thus, (B) in the lemma holds.

Plugging $u = (1/2)Ae^{(r-a)t}$ into $\dot{x} = ue^{-at}$, we obtain

$$\dot{x}^* = \frac{1}{2} A e^{(r-2a)t}.$$

Solving the differential equation, we obtain

$$x^* = C + \int \frac{1}{2} A e^{(r-2a)t} dt \Rightarrow x^* = C + \frac{1}{2(r-2a)} A e^{(r-2a)t},$$

where C is a constant. Noting $x^*(0) = 0$, we have $C = -A/2(r-2a)$ so that

$$x^*(t) = \frac{A}{2(2a-r)} \left(1 - e^{(r-2a)t}\right).$$

As $a > r/2$, we have

$$\lim_{t\to\infty} x^*(t) = \frac{A}{2(2a-r)} \underbrace{\Rightarrow}_{\lim_{t\to\infty} x(t)\geq K} A \geq 2K(2a-r) \Rightarrow A > 0.$$

So, (C) in the lemma holds.

We compute

$$\begin{aligned} \lambda(t)e^{-rt}(K - x^*(t)) &= Ae^{rt}e^{-rt}\left[K - \frac{A}{2(2a-r)}\left(1 - e^{(r-2a)t}\right)\right] \\ &\to A\left[K - \frac{A}{2(2a-r)}\right] \quad \textit{as } t \to \infty \end{aligned}$$

Setting $A = 2K(2a-r)$, we obtain

$$\lim_{t\to\infty} \lambda(t)e^{-rt}(K - x^*(t)) = 0 \Rightarrow \textit{(A) in the lemma hold.}$$

By Lemma 9.9.1, (d) holds. Thus, by Theorem 9.9.1, we have found the solution to Problem (1).

We note a necessary condition for the maximum principle for an infinite horizon.

Theorem 9.9.2 (Necessary Condition for an Infinite Horizon) *Suppose that $(x^*(t), u^*(t))$ is a solution to Problem (1), with no condition on the limiting behavior of $x(t)$ as $t \to \infty$. Suppose that $\int_{t_0}^{\infty} |f(t, x(t), u(t))|dt < \infty$ for all admissible $(x(t), u(t))$. Assume further that there exist positive constants A and k with $r > k$ such that $|\partial f(t, x, u^*(t))/\partial x| \leq A$ and $\partial g(t, x, u^*(t))/\partial x \leq k$ for all x. Then, there exists a continuous function $\lambda(t)$ such that, with $\lambda_0 = 1$,*

$$H^c(t, x^*(t), u, \lambda(t)) \leq H^c(t, x^*(t), u^*(t), \lambda(t)) \ \textit{ for all } u \in U$$

and such that the function $\lambda(t)$ equals $\lim_{T\to\infty} \lambda(t, T)$, where $\lambda(t, T)$ is the solution of

$$\dot{\lambda} - r\lambda = -\partial H^c(t, x^*(t), u^*(t), \lambda)/\partial x, \ \lambda(T, T) = 0.$$

Bibliography

[1] Axler, A., *Linear Algebra Done Right*, Second Edition, (1997), Springer-Verlag.

[2] Jehle, G-A and P-J. Reny, *Advanced Microeconomic Theory*, Third Edition, (2011), Pearson.

[3] Shilov, G.E., *Linear Algebra*, English Edition Translated and Edited by Richard A. Silverman, (1977), Dover Publications.

[4] Simon, C.P. and L. Blume, *Mathematics for Economists*, (1994), W.W Norton & Company.

[5] Sydsaeter, K., P. Hammond, A. Seierstad, and A. Strøm, *Further Mathematics for Economic Analysis*, Second Edition, (2008), Prentice Hall.