Mathematical Methods in Economics (Part I)

Lecture Note

Kai Hao Yang*

09/03/2018

Contents

*Department of Economics, University of Chicago; e-mail: khyang@uchicago.edu

1 Basic Topology and Linear Algebra

1.1 Review of Metric Space and Introduction of Topological Space

Topology is one of the most commonly used mathematical concepts in economic theory, it abstractly defines the concept of *neighborhoods* and thus *continuity*, which are crucial to optimization problems in economics. Metric spaces— which I assume most of you are familiar with—is one of the special cases, under which the concept of *distance* is properly defined. In this section, we will briefly review the definitions and properties of metric spaces and introduce the concept of topology and continuity. Finally, we will look at some widely used related theorems in economics.

1.1.1 Metric Space

Recall that a given a nonempty set X, a *metric* on X is a mapping $d : X \times X \to \mathbb{R}_+$ that describes the "distance" between two points $x, y \in X$. Formally, we say that $d : X \times X \to \mathbb{R}_+$ is a metric if:

1. $d(x, y) \geq 0$, for all $x, y \in X$.
2. $d(x, y) = 0$ if and only if $x = y$.
3. $d(x.y) = d(y, x)$, for all $x, y \in X$
4. $d(x, y) \leq d(x, z) + d(z, y)$, for all $x, y, z \in X$.

For any metric d on X, we say that (X, d) is a *metric space*.

Example 1.1.1.

- (Euclidean distance). Let $X = \mathbb{R}^n$, $d : X \times X \to \mathbb{R}_+$ be defined as

$$d(x, y) := \left(\sum_{k=1}^{n} (x_k - y_k)^2 \right)^{\frac{1}{2}}.$$

  . This is a very commonly seen metric space.

- (L^2 distance). Let X be the collection of functions f such that $\int_0^1 f(x)^2 dx < \infty$ and let $d : X \times X \to \mathbb{R}_+$ be defined as

$$d(f,g) := \left(\int_0^1 (f(x) - g(x))^2 dx \right)^{\frac{1}{2}}.$$

  This is also a very commonly used metric on functional spaces, often denoted by $L^2([0,1])$, we will have more discussions on this in the later sections.

- (Discrete metric). Let X be any nonempty set, define $d : X \times X \to \mathbb{R}_+$ as

$$d(x,y) = \begin{cases} 1, & \text{if } x \neq y \\ 0, & \text{if } x = y \end{cases}.$$

- A well-known measure of "distance" in probability theory and statistics (econometrics as well) is sometimes called the *Kullback-Leibler distance*. Formally, let f and g be two probability density functions[1], the Kullback-Leibler distance between f and g is given by:

$$D(f||g) := \int_{\mathbb{R}} \log \left(\frac{f(x)}{g(x)} \right) f(x) dx.$$

  Notice that this is **NOT** a metric! (which property fails?)

A crucial element in metric spaces is the *open balls*. Given any $x \in X$, any $\delta > 0$, let

$$N_\delta(x) := \{y \in X | d(y,x) < \delta\}$$

be the set of elements in X that are δ-closer to x. We say that $N_\delta(x)$ is an open ball around x with radius δ.

Using open balls as basis, we then have the concept of *open sets*, which are the building blocks of topology.

Definition 1.1.1. Let (x,d) be a metric space, a set $\mathcal{O} \subseteq X$ is *open* if for any $x \in \mathcal{O}$, there exists $\delta > 0$ such that $N_\delta(x) \subset \mathcal{O}$.

[1] We will return to a more formal definition of probability density functions in chapter 4. For now, I assume your familiarity with basic probability theory. In this case, you can imagine f, g as two distinct density function of, say, normal distribution.

The following concept, *closedness*, is closely related to openess and is also crucial to many analysis problems.

Definition 1.1.2. Let (X, d) be a metric space, a set $F \subseteq X$ is *closed* if $X \backslash F$ is open.

Definition 1.1.3. Let (X, d) be a metric space and let $S \subseteq X$ be any set. The *closure* of S is defined as the smallest closed set that contains S. That is:

$$\mathrm{cl}(S) := \bigcap \{F \subset X | F \text{ is closed}, S \subset F\}.$$

Definition 1.1.4. Let (X, d) be a metric space, define the *interior* of a set $S \subseteq X$ as:

$$\mathrm{int}(S) := \{x \in S | \exists\, \delta > 0 \text{ s.t. } N_\delta(x) \subset S\}$$

Remark 1.1.1.

1. If a set S is not closed, it does not mean that it is open. Conversely, a set can be both open and closed. Example?

2. Open sets and closed sets are defined *relatively* to the spaces they are in. A set $S \subset Y \subset X$ may be open (closed) in the metric space (Y, d) but not open (closed) in (X, d). Example?

Sequences, Convergence and the $\varepsilon - N$ criterion

Given a metric space (X, d), it is sometimes very useful to define or characterize concepts by sequences and the notion of convergence. Let X be a nonempty set, we say that $\{x_n\}$ is a *sequence* in X if $x_n \in X$ for all $n \in \mathbb{N}$. Furthermore, we say that $\{x_{n_k}\}$ is a *subsequence* of $\{x_n\}$ if for all $k \in \mathbb{N}$, $x_{n_k} = x_n$ for some $n \in \mathbb{N}$.

Definition 1.1.5. Let (X, d) be a metric space, we say that a sequence $\{x_n\}$ *converges* to $x \in X$, denoted as $\{x_n\} \to x$ or $\lim_{n\to\infty} x_n = x$ if for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ whenever $n > N$

There is a useful characterization of closed sets that involves convergence of sequences.

Proposition 1.1.1. *Let (X, d) be a metric space, a set $F \subseteq X$ is closed if and only if for any $\{x_n\} \subset F$, $x \in X$ such that $\{x_n\} \to x$, we have $x \in F$.*

If $X = \mathbb{R}$ and d is the Euclidean metric and $\{x_n\}$ is a sequence in X, the following two concepts are also useful.[2]

$$\limsup_{n\to\infty} x_n := \lim_{N\to\infty} \sup\{x_n | n > N\}.$$

$$\liminf_{n\to\infty} x_n := \lim_{N\to\infty} \inf\{x_n | n > N\}.$$

It is easy to show that $\lim_{n\to\infty} x_n$ exists if and only if $\limsup_{n\to\infty} x_n = \liminf_{n\to\infty} x_n$.

1.1.2 Other Important Notions and Propetries

Below we will introduce other important notions in metric spaces and summarize their useful properties. Throughout this section, we will fix a metric space (X, d). The first notion we will introduce is *compactness*, which is essential to many economic and mathematical problems.

Definition 1.1.6. A set $S \subseteq X$ is *compact* if for any family of open sets $\mathcal{F}$ such that $S \subset \cup_{O\in\mathcal{F}} O$, there exists a finite subcollection $\{O_k\}_{k=1}^n \subset \mathcal{F}$ such that $S \subset \cup_{k=1}^n O_k$.

Proposition 1.1.2. *A set S is compact if and only if it is* sequentially compact. *That is, for any $\{x_n\} \subset S$ such that $d(x_n, x_0) \leq D$ for some $x_0 \in X$, $D \in \mathbb{R}_+$, there exists a subsequence $\{x_{n_k}\} \subset \{x_n\}$ such that $\{x_{n_k}\} \to x$ for some $x \in S$.*

Proposition 1.1.3. *A set S is closed and bounded if it is compact.*

Notice that if $X = \mathbb{R}^n$ and d is the Euclidean distance, the converse is also true. However, the converse is not always true (counter-example?)

Proposition 1.1.4. *Let $F \subset S$ be a closed subset of a compact set S. Then F is compact.*

The rest of the notions will also be useful in some economic analyses, some of which will be used later in this course.

Definition 1.1.7. A subset $D \subseteq X$ is *dense* in (X, d) if for any $x \in X$, for any $\delta > 0$, there exists $x_0 \in D$ such that $x \in N_\delta(x_0)$.

[2]Existence of limsup is ensured by observing that $\sup\{x_n | n > N\}$ is decreasing and thus the limit exists, as an implication of completeness of real numbers. Similarly for liminf

Definition 1.1.8. A metric space (X, d) is *separable* if there exists a countable subset $D \subset X$ that is dense in X.

Definition 1.1.9. A metric space (X, d) is connected if there does not exist two nonempty open sets $U, V \subset X$ such that $X = U \cup V$ and $U \cap V = \emptyset$.

Proposition 1.1.5. *Let (X, d) be a connected metric space, if set $S \subseteq X$ is open and closed, then $S = X$ or $S = \emptyset$.*

Definition 1.1.10. A sequence $\{x_n\} \subseteq X$ is a *Cauchy sequence* if for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ for all $n, m > N$.

It is easy to verify that any convergent sequence is a Cauchy sequence. However, the converse is not necessarily true, which leads to the next notion in metric spaces.[3]

Definition 1.1.11. A metric space (X, d) is *complete* if all the Cauchy sequences are convergent.

Many of the well-known metric spaces are complete. In particular, the real numbers (with Euclidean distance) are complete. As the theorem of nested interval in real numbers, completeness of a metric space can be characterized by a similar property.

Proposition 1.1.6. *A metric space (X, d) is complete if any only if for any sequence of closed sets $\{F_n\} \subset 2^X$ such that $F_{n+1} \subset F_n$, there exists $x \in X$ such that $\cap_{n=1}^{\infty} F_n = \{x\}$.*

1.1.3 Topological Space and Normed Linear Space

One of the special cases of metric space is called *normed linear space*. Besides the notion of *distance*, as titled, norm linear space has *linear structure* on it and inherits the notion of *norm*.[4]

Definition 1.1.12. Let X be a linear space (defined on the field $\mathbb{R}$), a *norm* on X, denoted by $||\cdot||$ is a nonnegative valued function on X such that

[3]Complete metric spaces have an important implication, which is related to the existence of one of the most frequently used problems in economics. We will return to this in the next section.

[4]Here I assume also your familiarity with simple linear algebra and therefore the notion of *vector space*. We will return to this topic in the second half of this course.

1. $||x|| = 0$ if and only if $x = 0$ for all $x \in X$.

2. $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in X$.

3. $||\alpha x|| = |\alpha|||x||$ for all $x \in X$, $\alpha \in \mathbb{R}$.

For any normed linear space X with norm $||\cdot||$, define $d : X \times X \to \mathbb{R}_+$ by

$$d(x, y) := ||x - y||, \forall x, y \in X.$$

It is then easy to verify that (X, d) is a metric space.

Example 1.1.2.

- The Euclidean space, with the Euclidean norm, is a normed linear space.

- For any $1 < p < \infty$, the space

$$L^p([0,1]) := \left\{ f : [0,1] \to \mathbb{R} \middle| \left(\int_0^1 |f(x)|^p dx \right)^{\frac{1}{p}} < \infty \right\}$$

  is a normed linear space, with the norm being defined by[5]

$$||f||_p := \left(\int_0^1 |f(x)|^p dx \right)^{\frac{1}{p}}.$$

- The set of continuous functions on $[0, 1]$, denoted by $C([0, 1])$ is a normed linear space, where the norm is defined by[6]

$$||f||_\infty := \max_{x \in [0,1]} |f(x)|$$

It is noteworthy that although a set $S \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded, closed and bounded set in $L^p([0, 1])$ or $C([0, 1])$ may not be compact. (counter-example?). However, under the notion of *weak topology*, which we will introduce in chapter 4, closed and bounded sets in these normed linear spaces are indeed compact.

[5]The integral here is the *Lebesgue integral*, which we will discuss in the next chapter.

[6]This is well defined due to the well-known property that continuous functions take maximum on a compact set, which we will also discuss in the next section.

Definition 1.1.13. Let X be a normed linear space, X is called a *Banach space* if it is complete under the induced metric.

Proposition 1.1.7. *$C([0,1])$ and $L^p([0,1])$, $1 < p < \infty$ are Banach spaces.*

On the other hand, there is a generalization of metric space, called the *topological space*. Instead of defining the notion of distance and use open balls to describe open sets, topological spaces define abstractly the notion of open sets without using distance.

Definition 1.1.14. Let X be a nonempty set, a collection of subsets of X, $\mathcal{T}$, is a *topology* if

1. $X, \emptyset \in \mathcal{T}$.
2. For any $\{O_k\}_{k=1}^n \subset \mathcal{T}$, $\cap_{k=1}^n O_k \in \mathcal{T}$.
3. For any $\mathcal{F} \subseteq \mathcal{T}$, $\cup_{O \in \mathcal{F}} O \in \mathcal{T}$.

Given a topology $\mathcal{T}$ on X, we say that $(X, \mathcal{T})$ is a topological space.

From the problem set, we can see that any metric space is a topological space, where the topology is exactly the collection of open sets. Under topological spaces, we can still have the notion of convergence.

Definition 1.1.15. Let $(X, \mathcal{T})$ be a topological space, a sequence $\{x_n\} \subset X$ is said to converge to $x \in X$ if for any $\mathcal{N} \in \mathcal{T}$ such that $x \in \mathcal{N}$, there exists $N \in \mathbb{N}$ such that $x_n \in \mathcal{N}$ whenever $n > N$.

Conceptually, we call a set $\mathcal{N} \in \mathcal{T}$ such that $x \in \mathcal{N}$ a *neighborhood* of x. Neighborhood replaces the role of open ball in metric spaces.

Remark 1.1.2.

1. Topological spaces are relatively less directly used in economics, the reason to introduce topological space is, first of all, for completion, and second of all, it is sometimes easier to work with the notion of topology instead of distances. For example, in chapter 4, we will briefly introduce the convergence of probability measures, although the space we consider is in fact a metric space, it is sometimes more convenient to think in topological terms.

2. Under most of the topological spaces that might appear in economic analyses, the above properties in metric spaces still hold. The precise analogy is beyond the scope of the course and is not of interest.

1.1.4 Application: Preference and Utility Representation

In most of the economic analyses, we base our model on economic agents' (consumers, household, players, agents etc) utility maximization problems. One of the most common critiques (from other disciplines) is that the economic agents may not have an actual "function" in their minds when making decision. However, the utility maximization problems can in fact be derived from a collection of natural axioms on individual's preferences. That is, with a few (relatively) acceptable axioms on individual's *preference*, one can then analyze such individual's choice *as if* they have a utility function to maximize. To elaborate, let X be a topological space (for expositional purpose, you can think of X as a subset of $\mathbb{R}^n$ that describes the set of feasible bundles). Consider a binary relation $\succsim$ on X, which we will call *preference*. Also, define $\succ$, $\sim$ by:

$$x \succ y \iff x \not\precsim y,$$

and

$$x \sim y \iff x \succsim y \text{ and } x \precsim y,$$

which we call *strict preference* and *indifference*. We will impose some minimal assumptions on such binary relation so that it describes a reasonable and *rational* preference relation of an individual.

Definition 1.1.16.

1. A binary relation is *complete* if for any $x, y \in X$, either $x \succsim y$ or $y \succsim x$.

2. A binary relation is *transitive* if for any $x, y, z \in X$, $x \succsim y$ and $y \succsim z$ implies $x \succsim z$.

3. A binary relation is *continuous* if for any $x \in X$, the upper-contour $\succsim (x) := \{y \in X | y \succsim x\}$ and the lower-contour $\precsim (x) := \{y \in X | y \precsim x\}$ are closed in X.

As we can see, preferences are hard to analyze, especially when there are infinitely many alternatives. To address this, we wish to *represent* one's preference by utility functions so that more mathematical tools can be applied.

Definition 1.1.17. A preference $\succsim$ is represented by a utility function $u : X \to \mathbb{R}$ if for any $x, y \in X$, $x \succsim y$ if and only if $u(x) \geq u(y)$.

The following theorem ensures existence of utility representation for reasonable preferences. As such, an economic agent's choice is *as if* he/she is maximizing a "utility function".

Theorem 1.1.1 (Debreu)**.** *Let X be a connected and separable topological space and $\succsim$ be a complete, transitive and continuous preference on X. Then there exists $u : X \to \mathbb{R}$ that represents $\succsim$.*

It is worthwhile to notice that completeness, transitivity and continuity are also necessary for existence of an utility representation. Consider the following example:

Example 1.1.3 (Lexicographical Preference). Let $\succsim$ be a binary relation on $\mathbb{R}^2_+$ which is defined as follows: For any $x, y \in \mathbb{R}^2_+$, if $x = y$, then $x \sim y$. If $x \neq y$, and if $x_1 > y_1$, then $x \succ y$. If $x \neq y$, $x_1 = y_1$ and $x_2 > y_2$, then $x \succ y$.

It can be directly verified that $\succsim$ is complete and transitive. However, consider any $x \in \mathbb{R}^2_+$, by definition, $\{(x_1 - \frac{1}{n}, x_2 + 1)\}_{n\in\mathbb{N}} \subset \prec (x) \subset \precsim (x)$, but $(x_1, x_2 + 1) = \lim_{n\to\infty}(x_1 - \frac{1}{n}, x_2 + 1) \succ x$ and thus $\precsim (x)$ is not closed. This shows that $\succsim$ is not continuous.

Proposition 1.1.8. *The lexicographical preference cannot be represented by any real-valued function.*

Proof. Suppose that $\succsim$ is represented by some real-valued function $u : \mathbb{R}^2_+ \to \mathbb{R}$. Then for any $x_1, y_1 \in \mathbb{R}_+$ with $y_1 > x_1 > 1$, $u(x_1, \cdot)$ and $u(y_1, \cdot)$ must be strictly increasing. Also, $u(x_1, \cdot)$ and $u(y_1, \cdot)$ must be bounded from above and from below. Indeed, if not, then for any $z \in \mathbb{R}_+$, there exists some $\overline{x_2}$, $\underline{x_2}$ such that $u(x_1 + 1, z) < u(x_1, \overline{x_2})$ and $u(x_1 - 1, z) > u(x_1, \underline{x_2})$, a contradiction. Similar for $u(y_1, \cdot)$. Therefore, there exists $m_x, M_x, m_y, M_y \in \mathbb{R}_+$ with $-\infty < m_x < M_x < \infty$, $-\infty < m_y < M_y < \infty$ such that $m_x = \inf_{z\in\mathbb{R}_+} u(x_1, z)$, $M_x = \sup_{z\in\mathbb{R}_+} u(x_1, z)$, $m_y = \inf_{z\in\mathbb{R}_+} u(y_1, z)$, $M_y = \sup_{z\in\mathbb{R}_+} u(y_1, z)$. Moreover, $m_y > M_x$, since if not, there exists $z, z' \in \mathbb{R}_+$ such that $u(y_1, z') < u(x_1, z)$, a contradiction.

Therefore, we may conclude that for any $x \in \mathbb{R}_+$, there exists $m_x, M_x \in \mathbb{R}_+$ such that $u(x,z) \in (m_x, M_x)$ for all $z \in \mathbb{R}_+$ and the collection of intervals $\{(m_x, M_x)\}_{x\in\mathbb{R}_+}$ is disjoint. Furthermore, since the rational numbers $\mathbb{Q}$ are dense in $\mathbb{R}_+$, for each $x \in \mathbb{R}_+$, there exists some $q_x \in \mathbb{Q}$ such that $q_x \in (m_x, M_x)$. Take and fix such q_x for each $x \in \mathbb{R}_+$. It then follows that the map $x \mapsto q_x$ is one-to-one. That is, there exists some one-to-one function $\varphi : \mathbb{R}_+ \to \mathbb{Q}$ and hence the cardinality of $\mathbb{R}_+$ is the same as that of $\mathbb{Q}$, a contradiction. ■

1.2 Function and Correspondence on Topological Space

1.2.1 Continuous Function

Economic analysis involves many optimization problems, from consumer's choice problem to firm's profit maximization to deriving best responses in games. A particular class of functions are essential and pervasive in these type of maximization problems. This section will be devoted to (re)introduce continuous functions and their properties that are most relevant to economic problems. To begin with, we will start with the definition and several characterization of continuous functions.

Definition 1.2.1. Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be two topological spaces. A function $f : X \to Y$ is continuous at $x_0 \in X$ if for any $U \in \mathcal{S}$ such that $f(x_0) \in U$, there exists $O \in \mathcal{T}$ such that $x_0 \in O$ and $f(y) \in U$ for any $y \in O$. Furthermore, f is said to be continuous on X if it is continuous at all $x \in X$.

Proposition 1.2.1. *Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be two topological spaces. A function $f : X \to Y$ is continuous if and only if for any $U \in \mathcal{S}$,*

$$f^{-1}(U) := \{x \in X | f(x) \in U\} \in \mathcal{T}$$

The above definition is the most general form of the definition of continuous function. Below we will further mention two more widely-used characterization of functions in metric spaces.

Proposition 1.2.2. *Let (X, d), (Y, ρ) be two metric spaces and $f : X \to Y$ be a function. The following are equivalent:*

1. *f is continuous at $x_0 \in X$.*

2. *(Sequential characterization) For any $\{x_n\}$ such that $\{x_n\} \to x_0$,*

$$\lim_{n\to\infty} f(x_n) = f(x_0).$$

3. *($\varepsilon - \delta$ criterion) For any $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$d(x, y) < \delta \Rightarrow \rho(f(x), f(x_0)) < \varepsilon.$$

A commonly used property of continuous functions is called the *Intermediate Value Theorem*, which states that a continuous function does not have "gaps" in its range.

Proposition 1.2.3 (Intermediate Value Theorem)**.** *For any $-\infty < a < b < \infty$, let $f : [a, b] \to \mathbb{R}$ be a real-valued function. If $f(a)f(b) < 0$, then there exists $c \in (a, b)$ such that $f(c) = 0$.*

A natural generalization of continuity when the functions considered are real valued is called semicontinuity. Conceptually, continuous functions do not "jump", whereas semicontinuous functions are only allowed to "jump" in one direction. The ones that only jump up are called upper-semicontinuous and the ones that only jump down are called lower-semicontinuous.

Definition 1.2.2. Let $(X, \mathcal{T})$ be a topological space and $f : X \to \mathbb{R}$ be a real-valued function. f is said to be *upper (lower)-semicontinuous* at $x_0 \in X$ if for any $\varepsilon > 0$, there exists $O \in \mathcal{T}$ such that $x_0 \in O$ and $f(x) < f(x_0) + \varepsilon$ ($f(x) > f(x_0) - \varepsilon$) for all $x \in O$. Furthermore, f is said to be upper (lower)-semicontinuous on X if it is upper (lower)-semicontinuous at all $x \in X$.

As continuous functions, when (X, d) is a metric space, sequential characterization of semicontinuity is particularly useful.

Proposition 1.2.4. *Let (X, d) be a metric space and $f : X \to \mathbb{R}$ is upper-semicontinuous (lower-semicontinuous) at $x_0 \in X$ if and only if for any $\{x_n\} \subset X$ such that $\{x_n\} \to x_0$,*

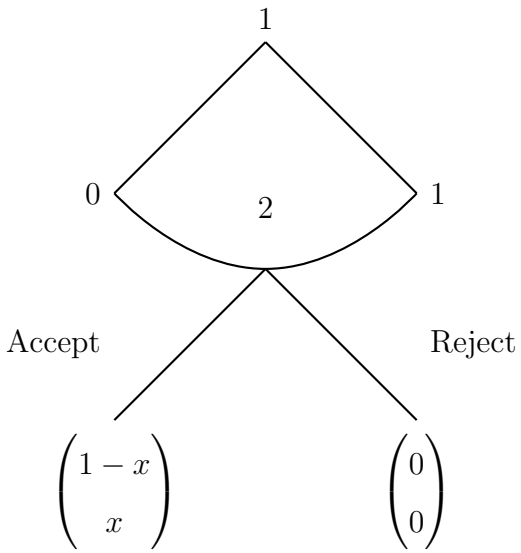
$$\limsup_{n\to\infty} f(x_n) \leq f(x_0) \quad (\liminf_{n\to\infty} f(x_n) \geq f(x_0)).$$

It is clear by definition that a function is continuous if and only if it is both upper and lower semicontinuous. The following properties are essential to economic analyses.

Proposition 1.2.5. *Let $(X, \mathcal{T})$ be a compact topological space and $f : X \to \mathbb{R}$ be a real-valued function. Then there exists $x^* \in X$ such that $f(x^*) \geq (\leq) f(x)$ for all $x \in X$ if f is upper-semicontinuous (lower-semicontinuous). In particular,* $\max_{x \in X} f(x)$ *and* $\min_{x \in X} f(x)$ *exists if f is continuous.*

Proposition 1.2.5 ensures existence of solution if one attempts to maximize or minimize a function f. In most of the economic problems, we need at least semi-continuity of the objective to ensure that optimum exists. As a result, there are many cases under which ties must be broken so that the objective is upper-semicontinuous.

Example 1.2.1. Consider a take-it-or-leave-it bargaining game. In this game, two players are bargaining over a pie of size 1. Player 1 can make an offer $x \in [0, 1]$ and then player 2, after seeing 1's offer, can decide whether to accept of reject. If player 2 accepts the offer, they divide the pie according to $(1 - x, x)$. If player 2 rejects the offer, both players get 0. The game is illustrated below:



We solve the subgame perfect equilibrium by using backward induction. Given any offer $x \in [0, 1]$, it is optimal for player 2 to accept if $x > 0$. On the other hand, if $x = 0$, player 2 is indifferent between accepting and rejecting so the best response can be any randomization between accepting and rejecting.

However, the only tie breaking rule so that player 1's payoff, as a function of x, given 2's strategy is upper-semicontinuous must be that player 2 accepts for sure when $x = 0$. Otherwise player 1's payoff function does not have a maximum. As such, even though there is a continuum of tie breaking rules for player 2 when $x = 0$, the unique subgame perfect equilibrium in this game is that player 1 offers $x = 0$ and player 2 accepts all the offers.

1.2.2 Correspondence

A curcial requirement for a function is that it must be single-valued. That is, for any x in the domain, there must be one and only one value $f(x)$ in the range being assigned. A natural generalization is to relax this requirement and allow multiple values in the range to be assigned to the same x. This is called a *correspondence*.

Definition 1.2.3. Let $(X, \mathcal{T}), (Y, \mathcal{S})$ be two topological spaces, $F : X \Rightarrow Y$ is called a *correspondence* if for any $x \in X$, $F(x) \subseteq Y$.

There are several analogous notions of continuity for correspondences.

Definition 1.2.4. Let $(X, \mathcal{T}), (Y, \mathcal{S})$ be two topological spaces, $F : X \Rightarrow Y$ is *upper-hemicontinuous* at $x_0 \in X$ if for any $U \in \mathcal{S}$ such that $F(x_0) \subseteq U$, there exists $O \in \mathcal{T}$ such that $x_0 \in O$ and

$$F(x) \subseteq U, \forall x \in O.$$

Furthermore, F is said to be upper-hemicontinuous on X if it is upper-hemicontinuous at every $x \in X$.

Definition 1.2.5. Let $(X, \mathcal{T}), (Y, \mathcal{S})$ be two topological spaces, $F : X \Rightarrow Y$ is *lower-hemicontinuous* at $x_0 \in X$ if for any $U \in \mathcal{S}$ such that $F(x_0) \cap U \neq \emptyset$, there exists $O \in \mathcal{T}$ such that $x_0 \in O$ and

$$F(x) \cap U \neq \emptyset, \forall x \in O.$$

Furthermore, F is saide to be lower-hemicontinuous on X if it is lower-hemicontinuous at every $x \in X$.

When X, Y are compact metric spaces, hemicontinuity also has useful sequential characterizations.

Proposition 1.2.6. *Let (X, d), (Y, ρ) be two metric spaces with (Y, ρ) being compact and $F : X \Rightarrow Y$ be a correspondence such that $F(x)$ is closed for all $x \in X$. Then:*[7]

1. *F is upper-hemicontinuous at $x_0 \in X$ if and only if for any $\{x_n\} \subset X$, $\{y_n\} \subset Y$ such that $\{x_n\} \to x_0$, $\{y_n\} \to y \in Y$ and $y_n \in F(x_n)$, for all $n \in \mathbb{N}$, $y \in F(x)$.*

2. *F is lower-hemicontinuous at $x_0 \in X$ if and only if for any $\{x_n\} \subset X$, $y \in Y$ such that $\{x_n\} \to x_0$ and $y \in F(x_0)$, there exists $\{y_n\} \subset Y$ such that $\{y_n\} \to y$ and $y_n \in F(x_n)$ for all $n \in \mathbb{N}$.*

Definition 1.2.6. Let $(X, \mathcal{T}), (Y, \mathcal{S})$ be two topological spaces, $F : X \Rightarrow Y$ is *continuous* at $x_0 \in X$ if it is both upper-hemiconituous and lower-hemicontinuous at x_0. We say that F is continuous on X if it is continuous at every $x_0 \in X$.

1.2.3 Theorem of Maximum

In this section, we will introduce a very widely used mathematical theorem in economic analysis. The baseline problem we are considering is as follows: Let (X, d), (Θ, ρ) be two metric spaces, $f : X \times \Theta \to \mathbb{R}$ be a function, $\Gamma : \Theta \Rightarrow X$ be a correspondence. The problem of interest is

$$\max_{x \in \Gamma(\theta)} f(x, \theta). \tag{1}$$

For each $\theta \in \Theta$, let

$$V(\theta) := \sup_{x \in \Gamma(\theta)} f(x, \theta)$$

and

$$X^*(\theta) := \{x \in \Gamma(\theta) | f(x, \theta) = V(\theta\}.$$

Many economic problems can be translated into (1). We can think of x as a *choice variable* that some economic is choosing to optimize his/her objective f. θ can be regarded as the *parameter* of the environment that the economic agent cannot affect but must take as given. This affect both the agent's objective and *constraint*, which is modeled as Γ. Many

[7]Part 1 is also called *closed graph* property. In fact, the condition that X, Y are compact and $F(x)$ is closed for all $x \in X$ is not necessary for 2 to hold. For our purpose in economic application, most of the time these conditions will be true.

of the economic analyses involves property of V and X^*. That is, it is often informative to know what how *optimal value* and *optimal choices* vary as the environment changes. The Theorem of Maximum ensures that they change continuously if the primitives exhibit property continuity.

Example 1.2.2. A very common example is the canonical consumer's problem. Let $\mathbb{R}^n_+$ be a commodity space, $\mathbf{p} \in \mathbb{R}^n_+$ be a price vector in the market and $m > 0$ be a consumer's income, $u : \mathbb{R}^n_+ \to \mathbb{R}$ be a consumer's utility function that is continuous. A consumer's problem is

$$\max_{\mathbf{x} \in \mathbb{R}^n_+} u(x) \text{ s.t. } \mathbf{p}^\intercal \mathbf{x} \leq m$$

We can think of the budget constraint as a correspondence on $\mathbb{R}^n_+ \times (0, \infty)$, that is define

$$\Gamma(\mathbf{p}, m) := \{\mathbf{x} \in \mathbb{R}^n_+ | \mathbf{p}^\intercal \mathbf{x} \leq m\}.$$

Using θ to denote $(\mathbf{p}, m)$, the consumer's problem can then be rewritten as:

$$\max_{x \in \Gamma(\theta)} u(x),$$

which is exactly a special case of (1) where f does not depend on θ. In this problem, we are often interested in $X^*(\mathbf{p}, m)$. Fix a $m > 0$ regarding this as a correspondence of $\mathbf{p}$, $X^*(\cdot, m)$ is in fact a (Marshallian) *demand correspondence*, which is crucial to many equilibrium analyses.

We will now formally state and prove the Theorem of Maximum.

Theorem 1.2.1 (Theorem of Maximum)**.** *Let (X, d), (Θ, ρ) be two metric spaces and suppose that X is compact.*[8] *Let $f : X \times \Theta \to \mathbb{R}$ be a continuous function, $\Gamma : \Theta \Rightarrow X$ be a compact-valued correspondence that is continuous at $\theta_0 \in \Theta$. For each $\theta \in \Theta$, define*

$$V(\theta) := \max_{x \in \Gamma(\theta)} f(x, \theta)$$

$$X^*(\theta) := \operatorname*{argmax}_{x \in \Gamma(\theta)} f(x, \theta).$$

[8] A slightly more general version does not require compactness of X. The requirement of compactness here is simply for the ease of the statement and proof, in that when X is compact Proposition 1.2.6 ensures equivalence between closed graph property and upper hemicontinuity.

Then $V : \Theta \to \mathbb{R}$ is continuous at θ_0 and $X^* : \Theta \Rightarrow X$ is compact valued and upper-semicontinuous at θ_0.

Proof. By Proposition 1.2.5 $X^*(\theta) \neq \emptyset$ and $V(\theta)$ is well-defined. For any $\theta \in \Theta$, by continuity of f and closedness of $\Gamma(\theta)$, it is easy to verify that $X^*(\theta)$ is closed, and therefore, by Proposition 1.1.4, $X^*(\theta)$ is compact for all $\theta \in \Theta$. Since X is compact, by Proposition 1.2.6, it suffices to show that for any $\{\theta_n\} \subset \Theta$, $\{x_n\} \subset X$ such that $\{\theta_n\} \to \theta_0$, $\{x_n\} \to x \in X$ and $x_n \in X^*(\theta_n)$, $x \in X^*(\theta_0)$. Indeed, if $x \notin X^*(\theta_0)$, by definition, there exists $x' \in \Gamma(\theta_0)$ such that $f(x', \theta_0) > f(x, \theta_0)$. Since $x' \in \Gamma(\theta_0)$ and Γ is lower-hemicontinuous at θ_0, Proposition 1.2.6 ensures that there exists a sequence $\{x'_n\} \subset X$ such that $x'_n \in \Gamma(\theta_n)$ for all $n \in \mathbb{N}$ and $\{x'_n\} \to x'$. In particular, the sequence $\{(x'_n, \theta_n)\}$ converges to (x', θ_0). As f is continuous, pick $\varepsilon \in (0, f(x', \theta_0) - f(x, \theta_0))$, there exists $n \in \mathbb{N}$ such that

$$f(x'_n, \theta_n) > f(x', \theta_0) - \varepsilon/2 > f(x, \theta_0) + \varepsilon/2 > f(x_n, \theta_n),$$

which contradicts to $x_n \in X^*(\theta_n)$. This establishes upper-hemicontinuity of X^* at θ_0.

On the other hand, to show that V is continuous at θ_0, consider any $\{\theta_n\} \subset \Theta$ such that $\{\theta_n\} \to \theta_0$. For each $n \in \mathbb{N}$, take and fix any $x_n \in X^*(\theta_n)$. Notice that $f(x_n, \theta_n) = V(\theta_n)$ for all $n \in \mathbb{N}$. Now consider any subsequence $\{V(\theta_{n_k})\}$, since X is compact, there is a further convergent subsequence $\{x_{n_{kl}}\}$ such that $\{x_{n_{kl}}\} \to x_0$ for some $x_0 \in X$. By upper-hemicontinuity of X^*, $x_0 \in X^*(\theta_0)$, which implies that $V(\theta_0) = f(x_0, \theta_0)$. Together, by continuity of f,

$$\lim_{l \to \infty} V(\theta_{n_{kl}}) = \lim_{l \to \infty} f(x_{n_{kl}}, \theta_{n_{kl}}) = f(x_0, \theta_0) = V(\theta_0).$$

Thus, since for any subsequence $\{V(\theta_{n_k})\}$, there is a further subsequence that converges to $V(\theta_0)$, we may conclude that $\{V(\theta_n)\} \to V(\theta_0)$, which establishes continuity of V. ■

There are many applications of the Theorem of Maximum, both in micro and macro economics.[9] One of the most immediate implications is called the *demand continuity lemma*, which asserts that the demand function is continuous on $\mathbb{R}^n_{++}$. Consider the setting as

[9]Keep this in mind so that you can return to this theorem when the Bellman equation is introduced in the second half of the class.

in Example 1.2.2, assume that $X \subset \mathbb{R}^n_+$ is a compact set of feasible bundles[10] and that $u : X \to \mathbb{R}$ is continuous and strictly quasi-concave.[11]

Corollary 1.2.1 (Demand Continuity Lemma)**.** *Suppose that $X \subset \mathbb{R}^n_+$ is a compact set, $u : X \to \mathbb{R}$ is a continuous and strictly quasi-concave function. Fix any $m > 0$, for any $\mathbf{p} \in \mathbb{R}^n_{++}$, define*

$$X^*(\mathbf{p}, m) := \underset{\mathbf{x} \in X}{\operatorname{argmax}} u(x) \text{ s.t. } \mathbf{p}^\intercal \mathbf{x} \leq m.$$

Then $X^*(\cdot, m)$ is a continuous function on $\mathbb{R}^n_{++}$.

Proof. For any $m > 0$, $\mathbf{p} \in \mathbb{R}^n_{++}$, define $\Gamma(\mathbf{p}, m) := \{x \in X | \mathbf{p}^\intercal x \leq m\}$. Notice that $\Gamma(\cdot, m)$ is compact-valued and continuous at every $\mathbf{p} \in \mathbb{R}^n_{++}$. By Theorem 1.2.1, $X^*(\mathbf{p}, m)$ is upper-hemicontinuous on $\mathbb{R}^n_{++}$. Furthermore, since u is strictly quasi-concave, $X^*(\mathbf{p}, \theta)$ is a singleton for any $\mathbf{p} \in \mathbb{R}^n_{++}$. Continuity then follows from Proposition 1.2.6. ■

1.2.4 Fixed Point Theorem

While the Theorem of Maximum is useful for *optimization* problems, *fixed point theorems* are often used for *equilibrium analysis*. As in most of the contexts in economics, *equilibrium* is defined to be a state that exhibits some sort of self-consistency, which can often be represented by a *fixed point* of a function or a correspondence. This section will introduce some commonly used fixed point theorems in economics.

Definition 1.2.7. Let X be a nonempty set, $f : X \to X$ be a function and $F : X \Rightarrow X$ be a correspondence. x^* is said to be a *fixed point* of f if $x^* = f(x^*)$ and x^* is said to be a *fixed point* of F if $x^* \in F(x^*)$.

Recall that by the Intermediate Value Theorem, if $f : [a, b] \to \mathbb{R}$ is a continuous function and $(f(a) - a)(f(b) - b) < 0$, there exists $x^* \in [a, b]$ such that $x^* = f(x^*)$. In other words, x^* is a fixed point of f. In fact, this can be generalized, by the following fixed point theorem.

Definition 1.2.8. Let X be a linear space, a subset $S \subseteq X$ is *convex* if for any $x, y \in S$ and any $\lambda \in [0, 1]$, $\lambda x + (1 - \lambda)y \in S$.

[10] In fact, we do not need compactness of X, but to apply the version of Theorem of Maximium given by Theorem 1.2.1, we need this assumption.

[11] That is, for any $x, y \in X$, $x \neq y$ and any $\lambda \in (0, 1)$, $u(\lambda x + (1 - \lambda)y) < \max\{u(x), u(y)\}$.

Theorem 1.2.2 (Brouwer's Fixed Point Theorem)**.** *For any $n \in \mathbb{N}$, let $S \subseteq \mathbb{R}^n$ be a nonempty, compact and convex subset of $\mathbb{R}^n$ and $f : S \to S$ be a continuous function. Then there exists $x^* \in S$ such that $x^* = f(x^*)$. That is, f has a fixed point.*

A extension of Brower's fixed point theorem to correspondence is called *Kakutani's fixed point theorem.*

Theorem 1.2.3 (Kakutani's Fixed Theorem)**.** *For any $n \in \mathbb{N}$, let $S \subseteq \mathbb{R}^n$ be a nonempty, compact and convex subset and let $F : S \Rightarrow S$ be a nonempty, closed convex-valued upper-hemicontinuous correspondence. Then there exists $x^* \in S$ such that $x^* \in F(x^*)$. That is, F has a fixed point.*

Example 1.2.3 (Existence of Mixed Strategy Nash Equilibrium). Let G be a (finite) game described by: 1) The set of players $\{1, \ldots, N\}$; 2) Each player's (finite) strategy space $\{S_i\}_{i=1}^n$; and each player's payoff function $u_i : S \to \mathbb{R}$, where $S := \prod_{i=1}^N S_i$. Consider a *mixed extension* of G, denoted as $\bar{G}$, in which the set of players is the same, but each player i has strategy space

$$M_i := \Delta(S_i) := \left\{ \lambda \in \mathbb{R}_+^{|S_i|} \middle| \lambda(s_i) \in [0,1], \forall s_i \in S_i; \sum_{s_i \in S_i} \lambda(s_i) = 1 \right\},$$

which denotes the set of *probability distributions* on S_i. And let the extended payoff function be expected payoffs, that is:

$$\bar{u_i}(\lambda) := \sum_{i=1}^{N} \sum_{s_i \in S_i} u_i(s) \lambda_i(s_i).$$

We say that a *mixed strategy profile* $\lambda = (\lambda_i)_{i=1}^N \in M := \prod_{i=1}^N M_i$ is a *Nash equilibrium* if for any $i \in \{1, \ldots, N\}$, any $\lambda_i' \in M_i$,

$$\bar{u_i}(\lambda) \geq \bar{u_i}(\lambda_i', \lambda_{-i}).$$

Theorem 1.2.4 (Nash)**.** *Let G be a finite game, then the mixed extension of G, $\bar{G}$, has a Nash equilibrium.*

Proof. For any $i \in \{1, \ldots, N\}$, define the *best response correspondence* by:

$$\beta_i(\lambda_{-i}) := \operatorname*{argmax}_{\lambda_i' \in M_i} \bar{u_i}(\lambda_i', \lambda_{-i}).$$

By construction, since $\bar{u}_i$ is linear in λ, it is continuous and thus, by compactness of M_i and Theorem 1.2.1, β_i is nonempty and compact valued and is upper-hemicontinuous on $M_{-i} := \prod_{j \neq i} M_j$. Moreover, since $\bar{u}_i$ is linear in λ, β_i is convex valued as well.

Let $\beta : M \Rightarrow M$ be defined as:

$$\beta(\lambda) := \begin{pmatrix} \beta_1(\lambda_{-1}) \\ \vdots \\ \beta_n(\lambda_{-n}) \end{pmatrix}.$$

β is then a nonempty, compact and convex valued upper-hemicontinuous correspondence. By Kakutani's fixed point theorem, there exsits $\lambda^* \in M$ such that $\lambda^* \in \beta(\lambda^*)$. As desired. ■

Although Brouwer's fixed point theorem and Kakutani's fixed point theorem are valid only when the domain is subset of $\mathbb{R}^n$, there are generalizations so that the fixed points are ensure to exist when the domain is infinite dimensional.

Theorem 1.2.5 (Schauder's Fixed Point Theorem)**.** *Let X be a locally convex topological vector space*[12] *and $S \subseteq X$ be a nonempty, compact and convex subset of X. Then for any continuous function $f : S \to S$, there exists $x^* \in S$ such that $x^* = f(x^*)$*

Theorem 1.2.6 (Fan-Glicksberg Fixed Point Theorem)**.** *Let X be a locally convex topological vector space and $S \subseteq X$ be a nonempty, compact and convex subset. Then for any nonempty, compact and convex-valued correspondence $F : S \Rightarrow S$ that is upper-hemicontinuous, there exists $x^* \in S$ such that $x^* \in F(x^*)$.*

Unlike the fixed theorems above, the next import fixed point theorem does not rely on the continuity of the function. Instead, it depends on the *contracting* property.

Definition 1.2.9. Let (X, d) be a metric space, a function $f : X \to X$ is a *contraction* (with modulus r) if there exists $r \in [0, 1)$ such that for any $x, y \in X$, $d(f(x), f(y)) \leq r d(x, y)$.

Theorem 1.2.7 (Banach Contraction Principle)**.** *Let (X, d) be a complete metric space and $f : X \to X$ be a contraction on X. Then there exists a unique $x^* \in X$ such that $x^* = f(x^*)$*

[12] We did not introduce the concept of topological vector space and local convexity. For most of the economic applications, it is sufficient to think of X as a Banach space

Proof. Take and fix any $x_0 \in X$. Define recursively a sequence $\{x_n\}_{n=0}^{\infty}$ by:

$$x_{n+1} := f(x_n).$$

Since f is a contraction, for each $n \in \mathbb{N}$,

$$d(x_{n+1}, x_n) = d(f(x_n), x_n) = d(f(f(x_{n-1})), f(x_{n-1})) \leq r d(f(x_{n-1}), x_{n-1}),$$

for some $r \in [0, 1)$ Inductively, for each $n \in \mathbb{N}$,

$$d(x_{n+1}, x_n) \leq r^n d(f(x_0), x_0).$$

Therefore, for any $n, m \in \mathbb{N}$ with $n < m$,

$$\begin{aligned} d(x_n, x_m) &\leq \sum_{j=n+1}^{m} d(x_j, x_{j-1}) \\ &\leq \sum_{j=n}^{m-1} r^j d(f(x_0), x_0) \\ &\leq \frac{r^n}{1-r} d(f(x_0), x_0) \end{aligned}$$

As $r < 1$, $\{x_n\}$ is a Cauchy sequence. Since (X, d) is complete, $\{x_n\} \to x^*$ for some $x^* \in X$. Moreover,

$$d(f(x^*), x^*) = \lim_{n \to \infty} d(f(x_n), x_n) = 0$$

and thus $x^* = f(x^*)$.

For uniqueness, suppose that $x^* = f(x^*)$ and $y^* = f(y^*)$ some some $x^*, y^* \in X$. Then

$$d(x^*, y^*) = d(f(x^*), f(y^*)) \leq r d(x^*, y^*),$$

which implies that $d(x^*, y^*) = 0$ and hence $x^* = y^*$. ■

Example 1.2.4 (Bellman Equation). In many economic problems, especially when dynamics are considered, the following type of problem (or more complicated versions) are often of interest: Given $x_0 \in X$

$$\max_{\{x_t\}_{t=0}^{\infty} \subset X} \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}), \text{ s.t. } x_t \in \Gamma(x_{t-1}), \forall t \in \mathbb{N}, \tag{2}$$

where $\beta \in (0,1)$, $X \subset \mathbb{R}$ is a compact metric space, F is bounded and continuous and $\Gamma : X \Rightarrow X$ is a compact-valued and continuous correspondence. We can interpret (2) as follows: A forward-looking economic agent is optimizing the life-time objective, with the same objective function (say, utility) being the same each period and β being the discount factor. The problem is complicated because in each period, current choice x_t may enter into current period flow payoff and the next period flow payoff (since F depends on both x_t and x_{t+1}). Moreover, current choice x_t may affect the feasible choices in the future (since $x_t \in \Gamma(x_{t-1})$). It is well-known that such problem, under the conditions provided above, has a solution.[13]. Let $V^*(x_0)$ denote the solution, as a function of the initial $x_0 \in X$.

Now observe that at the beginning of period $t \in \mathbb{N}$, if we take x_{t-1} as given, the maximization problem for the future is exactly the same as (2) with x_0 being replaced by x_{t-1}. Inductively, we can write $V^*(x_0)$ as:

$$V^*(x_0) = \max_{x \in \Gamma(x_0)} [F(x_0, x) + \beta V^*(x)]$$

As such, if we let $W : C(X) \to C(X)$ be an function defined by:

$$W(V)[x] := \max_{y \in \Gamma(x)} [F(x, y) + \beta V(y)], \forall x \in X,$$

by the Banach Contraction Principle, if W is a contraction, since by Proposition 1.1.7, $C(X)$ is a complete metric space, there exists a unique fixed point. That is, there exists a unique V such that:

$$V(x) = \max_{y \in \Gamma(x)} [F(x, y) + \beta V(y)], \tag{3}$$

which then implies that $V = V^*$. This then gives a nice characterization of the solution of the problem (2). We call (3) the *Bellman Equation*.

It is well known that when $\beta \in (0,1)$, W is indeed a contraction with modulus β. Thus, there must be a unique solution to (3).

1.3 Exercises

1. Show that the *Kullback-Leibler distance* is not a valid metric.

[13] You will see a lot more about this in the second half of the course

2. Give example to show that a set can be both open and closed and that a set can be neither open nor closed.

3. Given an example to show that a set can be open in one metric space but not open in another.

4. Show that $C([0,1])$ is complete.

5. Let (X,d) be a metric space. Show that the collection of open sets in X is a topology.

6. (Characterization of Continuity of Preference) Let $\succeq$ be a preference relation on $\mathbb{R}^n_+$ that is complete and transitive. Consider the following axiom:

$$\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+ | \mathbf{x} \succeq \mathbf{y}\} \text{ is closed.} \tag{C}$$

   Show that (C) holds if and only if $\succeq$ is continuous. (Hint: This is a little hard, you might need to notice that $\mathbb{R}^n_+$ is a connected metric space)

7. Prove Proposition 1.2.2.

8. Consider a *Bertrand competition game*. That is, there are two firms, both of them are facing a market with perfectly inelastic demand and firm i has a marginal cost of production c_i. Suppose that the firms are competing by setting lower prices. That is, if $p_i < p_j$, then firm i wins the whole market and gets profit $(p_i - c_i)$ and firm j loses and gets zero. Suppose also that whenever there is a tie, market is divided in half. That is when $p_1 = p_2$, then firm i gets profit $\frac{1}{2}(p_i - c_i)$ for $i \in \{1,2\}$. Show that:

   (a) When $c_1 = c_2 = c > 0$, there exists a unique pure strategy Nash equilibrium, under which $p_1 = p_2 = c$.

   (b) When $0 < c_1 < c_2$, there is no pure strategy equilibrium. What went wrong? Is there any assumption you can make to get around with this?

9. (Blackwell's Test) Let X be compact topological space. Suppose that $W : C(X) \to C(X)$ is a self map on $C(X)$, the collection of continuous functions, under the norm $||f||_\infty := \max_{x \in X} |f(x)|$. Show that W is a contraction if:

(a) W is increasing. That is, for any $f, g \in X$, $f \geq g$, $W(f) \geq W(g)$.

(b) For any $f \in X$, $\alpha \in \mathbb{R}_+$,

$$W(f + \alpha) \leq W(f) + \beta\alpha,$$

for some $\beta \in (0, 1)$. Furthermore, use this to show that the Bellman operator (3) is a contraction.

2 Integration and Differentiation

In many economic problems, we consider maximization problems and the changes of endogenous solutions as the environment changes. Differentiation is a very useful mathematical tool for achieving these goals. On the other hand, integration deals with the notions of aggregation and average, which is also essential in many economic problems. Moreover, differentiation and integration are closely related. In this chapter, we will introduce a powerful and widely used method of integration and thus the associated concepts of differentiation. We will start with the definition and properties of *Lebesgue integral* and then introduce differentiation and its relationship with integration. Finally, we conclude this chapter by briefly applying these concepts in a simple *mechanism design* problem.

2.1 Introduction to Lebesgue Measure

2.1.1 Construction and Basic Properties of Lebesgue Measure

Before introducing the *Lebesgue integral*, we first introduce the notion of *Lebesgue measure* on $\mathbb{R}$.[14] In particular, we will introduce the *Lebesgue measure* on $\mathbb{R}$. Conceptually, Lebesgue measure is a function define on a particular subset of the power set of $\mathbb{R}$ that inscribers the "size" of a given set. Obviously, for any interval $(a, b) \subset \mathbb{R}$, the most "natural" way to define the size for it by using its length, $b - a$. In fact, the same idea can be applied to many other sets.

[14] We will return to a more general theory of measure in chapter 4, for now we will only focus on the the Lebesgue measure.

Formally, given any subset $E \subset \mathbb{R}$, we may define a *outer measure* of E by approximating it with the total length of the covering intervals. Formally, for any $E \subset \mathbb{R}$, define

$$\lambda^*(E) := \inf \left\{ \sum_{k=1}^{\infty} l(I_k) \middle| \bigcup_{k=1}^{\infty} I_k \supseteq E;\ I_k \text{ is an interval } \forall k \right\},$$

where $l(I) := \sup I - \inf I$ is the length of an interval I. By construction, it is not hard to verify that the outer measure λ^* is *monotone* and *countably subadditive*. That is for any $A, B \subseteq \mathbb{R}$ with $A \subset B$, $\lambda^*(A) \leq \lambda^*(B)$ and for any disjoint collection of sets $\{E_k\}_{k=1}^{\infty} \subset \mathbb{R}$,

$$\lambda^* \left(\bigcup_{k=1}^{\infty} E_k \right) \leq \sum_{k=1}^{\infty} \lambda^*(E_k).$$

The outer measure, although intuitive, has a fundamental unattractive property that makes it arguably not a proper measure of the "size" of a set. That is, although it is countable subadditive, it is possible that for two disjoint sets $A, B \subset \mathbb{R}$,

$$\lambda^*(A \cup B) < \lambda^*(A) + \lambda^*(B).$$

In other words, there exists two disjoint sets under which the "size" of the union of these two sets is strictly smaller than the sum of individual "sizes" under the outer measure. To address this problem, we consider only a subclass of the subset in $\mathbb{R}$ so that the issue above does not occur, which we will name *measurable sets*.

Definition 2.1.1. A set $E \subseteq \mathbb{R}$ is (Lebesgue) *measurable* if, for any $A \subseteq \mathbb{R}$,

$$\lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap E^C).$$

As such, for any disjoint *measurable* sets $A, B \subset \mathbb{R}$,

$$\lambda^*(A \cup B) = \lambda^*([A \cup B] \cap A) + \lambda^*([A \cup B] \cap A^C) = \lambda^*(A) + \lambda^*(B)$$

and then we have—after induction and some manipulations of the set-theoretic algebra — finite addititivity when restricting the outer measure on measurable sets.

The collection of measurable sets have some important properties in measure theory, as the collection of open sets are crucial in topology.

Proposition 2.1.1. *Let $\mathcal{M}$ be the collection of (Lebesgue) measurable sets in $\mathbb{R}$. $\mathcal{M}$ has following properties:*

1. $\mathbb{R} \in \mathcal{M}$.

2. For any $E \in \mathcal{M}$, $E^C \in \mathcal{M}$.

3. For any countable collection of sets $\{E_k\}_{k=1}^{\infty} \subset \mathcal{M}$, $\cup_{k=1}^{\infty} E_k \in \mathcal{M}$.

In other words, the collection $\mathcal{M}$ has some algebraic properties, including that it contains whole space and empty set, it closed under taking complements and countable unions. In fact, $\mathcal{M}$ belongs to a family called *σ-algebra*.

Definition 2.1.2. Let $\mathcal{F}$ be a collection of sets in $\mathbb{R}$, we say that $\mathcal{F}$ is a *σ-algebra* if:

1. $\mathbb{R} \in \mathcal{F}$.

2. For any $E \in \mathcal{M}$, $E^C \in \mathcal{F}$.

3. For any countable collection of sets $\{E_k\}_{k=1}^{\infty} \subset \mathcal{F}$, $\cup_{k=1}^{\infty} E_k \in \mathcal{F}$.

Proposition 2.1.1 can then be stated as the collection of Lebesgue measurable sets is a σ-algebra. Moreover, by definition, σ-algebra is closed under countable intersection and union operations and thus countable unions and intersections of σ-algebras are still σ-algebra.

Relatedly, it is not difficult to show that every interval is measurable.

Proposition 2.1.2. *For any interval* $I \subset \mathbb{R}$, $I \in \mathcal{M}$.

An immediate Corollary from Proposition 2.1.1 and Proposition 2.1.2 is that all the open sets are (Lebesgue) measurable, as all open sets in $\mathbb{R}$ can be represented as countable union of disjoint intervals. In other words, as a σ-algebra, $\mathcal{M}$ contains all the open sets. On the other hand, since intersections of σ-algebras are still σ-algebra, there exists the *smallest* (in set inclusion term) σ-algebra that contains all the open sets, we call this σ-algebra the *Borel* algebra and denote it by $\mathcal{B}$. Together, since $\mathcal{M}$ contains all the open sets and it is a σ-algebra, $\mathcal{B} \subseteq \mathcal{M}$. That is, any set in the Borel algebra is (Lebesgue) measurable. As we can see, Borel algebra is closely related to the topological structure of $\mathbb{R}$ and is a crucial linkage between measure theory and topology. We will return to this point in chapter 4.

With the collection $\mathcal{M}$, we can now properly define the *Lebesgue measure* on $\mathbb{R}$ as the restriction of the outer measure on $\mathcal{M}$. That is, the *Lebesgue measure* $\lambda : \mathcal{M} \to \mathbb{R}_+$ is defined as

$$\lambda := \lambda^*|_{\mathcal{M}}.$$

Notice that when defined on $\mathcal{M}$, by monotonicity of λ^* and Proposition 2.1.1, together with finite additive of $\lambda^*|_{\mathcal{M}}$, we have that for any disjoint collection $\{E_k\}_{k=1}^{\infty} \subset \mathcal{M}$,

$$\lambda\left(\bigcup_{k=1}^{\infty} E_k\right) \geq \lambda\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} \lambda(E_k),$$

for all $n \in \mathbb{N}$ and thus

$$\lambda\left(\bigcup_{k=1}^{\infty} E_k\right) \geq \lim_{n\to\infty} \sum_{k=1}^{n} \lambda(E_k) = \sum_{k=1}^{\infty} \lambda(E_k).$$

Therefore, together with countable subadditive of λ^*, we have

$$\lambda\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \lambda(E_k),$$

which (finally) established the desired property of countable additivity.

A crucial implication of countable additivity of λ is called *continuity* of measure.

Proposition 2.1.3. *The Lebesgue measure λ has the following properties:*

1. *(Continuity from below) For any $\{E_k\}_{k=1}^{\infty} \subset \mathcal{M}$ such that $E_{k+1} \supseteq E_k$,*

$$\lambda\left(\bigcup_{k=1}^{\infty} E_k\right) = \lim_{k\to\infty} \lambda(E_k).$$

2. *(Continuity from above) For any $\{F_k\}_{k=1}^{\infty} \subset \mathcal{M}$ such that $F_{k+1} \subseteq F_k$, $\lambda(F_1) < \infty$*

$$\lambda\left(\bigcap_{k=1}^{\infty} F_k\right) = \lim_{k\to\infty} \lambda(F_k).$$

3. *(Countable Subadditivity) For any $\{E_k\}_{k=1}^{\infty} \subset \mathcal{M}$ (need not to be disjoint),*

$$\lambda\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} \lambda(E_k).$$

With the Lebusgue measure, given any $E \subseteq \mathbb{R}$ we often say that a property holds *(Lebesgue) almost everywhere* provided that this property holds on E except for a subset of Lebesgue measure zero. For instance, it is not hard to show that any countable set on $\mathbb{R}$ is (Lebesgue) measurable and has Lebesgue measure zero. As a result, we may say that an increasing function $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ almost everywhere, as it is well known that increasing functions have at most countably many jumps.

The last property we want to emphasize in this section is called the *Borel-Cantelli Lemma*

Theorem 2.1.1 (Borel-Cantelli Lemma)**.** *For any $\{E_k\}_{k=1}^{\infty} \subset \mathcal{M}$ such that $\sum_{k=1}^{\infty} \lambda(E_k) < \infty$, almost all $x \in \mathbb{R}$ belongs to at most finitely many sets in the collection $\{E_k\}_{k=1}^{\infty}$*

Proof. By countable additivity,

$$\lambda\left(\bigcup_{k=m}^{\infty} E_k\right) \leq \sum_{k=m}^{\infty} \lambda(E_k), \, \forall m \in \mathbb{N}$$

and therefore by continuity of measure,

$$\lambda\left(\bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} E_k\right) = \lim_{m\to\infty} \lambda\left(\bigcup_{k=m}^{\infty} E_k\right) \leq \lim_{m\to\infty} \sum_{k=m}^{\infty} \lambda(E_k) = 0$$

■

Remark 2.1.1. In fact, there are many notion of "sizes" in mathematics, Lebesgue measure (or measures, in general) is only one of the ways to define "size" of a set. Various definitions might give a very different descriptions of the size of a set. For instance, it is well-known that the *Cantor set* is uncountable, which is considerably "large" in terms of set theory, in the sense that it has the same cardinality as $\mathbb{R}$. However, it is easy to show that the Cantor set has Lebesgue measure zero!

2.1.2 Measurable Functions

Just as continuous function is closely related to the topological structure on its domain and range, *measurable functions* are defined connect two measurable spaces. In this section, we will consider real-valued functions on $\mathbb{R}$. We will return to this in chapter 4.

Recall that we denote $\mathcal{M}$ by the collection of (Lebesgue) measurable sets on $\mathbb{R}$. Sometimes we refer $(\mathbb{R}, \mathcal{M})$ as a *measurable space*. Henceforth, throughout the chapter, we will take

and fix a measurable subset of $\mathbb{R}$, denoted by X and will let $\mathcal{M}$ denote the collection of (Lebesgue) measurable sets in X. Consider a function $f : \mathbb{R} \to \mathbb{R}$.

Definition 2.1.3. A function $f : X \to \mathbb{R}$ is (Lebesgue) *measurable* if for any Boral set $E \in \mathcal{B}$,

$$f^{-1}(E) := \{x \in \mathbb{R} | f(x) \in E\}$$

is Lebesgue measurable.

From Proposition 1.2.1, we know that a function $f : X \to \mathbb{R}$ is continuous on X if and only if for any open set in $\mathbb{R}$, its *pre-image* is also open in X. Analogously, we say that f is measurable if for any Borel set in $\mathbb{R}$, its pre-image is (Lebesgue) measurable. Using the definition of Borel algebra and the property that $\sigma-$ algebra is closed under complements and countable unions, the following characterization is useful:

Proposition 2.1.4. *A function $f : X \to \mathbb{R}$ is (Lebesgue) measurable if and only if for any $c \in \mathbb{R}$,*

$$\{x \in X | f(x) > c\} \in \mathcal{M}.$$

Notice that since σ-algebra is closed under complements, the "$>$" above can be replaced by "$<$", "$\geq$", "$\leq$".

With Proposition 2.1.4 and Proposition 1.2.1, it is clear that all the continuous functions on X are measurable. Furthermore, using the characterization in Proposition 2.1.4, and the fact that all open sets are in the Borel algebra, it immediately follows that a composition between a measurable function and a continuous function is again measurable:

Proposition 2.1.5. *Let $f : X \to \mathbb{R}$ be a measurable function and $g : X \to \mathbb{R}$ be a continuous function. Then the composition, $f \circ g$ is measurable.*

A nice property of measurable functions is that they can be "approximated" by a family of relatively more tractable functions, which we call *simple functions*. Before introducing the property formally, we need to first define our notion of approximation

Definition 2.1.4. Let $\{f_n\}$ be a sequence of real-valued functions on X and $f : X \to \mathbb{R}$ be a real-valued function, the sequence $\{f_n\}$ converges to f on X *pointwisely* if for any $x \in X$,

$$\lim_{n\to\infty} f_n(x) = f(x).$$

Furthermore, we say that $\{f_n\}$ converges to f pointewisely (Lebesgue) *almost-everywhere* if the exists a set $E \subseteq X$, with $\lambda(X) = \lambda(E)$, such that

$$\lim_{n\to\infty} f_n(x) = f(x).$$

An immediate observation, by using Proposition 2.1.1, is that measurability is preserved under pointwise convergence.

Proposition 2.1.6. *Let $\{f_n\}$ be a sequence of real-valued measurable functions on X. Suppose that $\{f_n\}$ converges to $f : X \to \mathbb{R}$ pointwisely. Then f is measurable.*

Definition 2.1.5. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a real-valued function on $\mathbb{R}$, we say that φ is a *simple function* if it takes at most finitely many values $\{a_k\}_{k=1}^n$, for some $n \in \mathbb{N}$ and for each $a_k \in \mathbb{R}$,

$$A_k := \varphi^{-1}(a_k) = \{x \in \mathbb{R} | \varphi(x) = a_k\} \in \mathcal{M}$$

With the definitions above, we can now introduce the approximation result.

Proposition 2.1.7 (Simple Approximation Theorem)**.** *Let $f : X \to \mathbb{R}$ be a real-valued measurable function. Then there exists a sequence of simple functions $\{\varphi_n\}$ with $|\varphi_n| \leq |f|$ for all $n \in \mathbb{N}$ such that $\{\varphi_n\}$ converges to f pointwisely. If, furthermore, $f \geq 0$, then such $\{\varphi_n\}$ can be taken so that $0 \leq \varphi_n \leq \varphi_{n+1}$ for all $n \in \mathbb{N}$.*

2.2 Integral

2.2.1 Construction and Basic Properties of Lebesgue Integral

Our goal in this section is to formally define the *Lebesgue integral*. Before we start, we will have a quick review on the integral that you are very familiar with: *Riemann integral*

The intuition of Riemann integral is to approximate an area "below" the graph of a function by "chopping" up the *domain* and compute the sum of the rectangular areas. Formally, let $f : [a, b] \to \mathbb{R}$ be a real-valued function on some interval $[a, b]$. We let p denote a *partition* on $[a, b]$. Namely,

$$p = \{x_0, \ldots, x_n | a = x_0 < x_1 < \ldots < x_n = b\}$$

for some finite $n \in \mathbb{N}$. Given a partition p on $[a, b]$, define respectively

$$L(f, p) := \sum_{k=1}^{n} m_k(x_k - x_{k-1})$$
$$U(f, p) := \sum_{k=1}^{n} M_k(x_k - x_{k-1}),$$

where $m_k : \inf_{x \in (x_{k-1}, x_k)} f(x)$ and $M_k := \sup_{x \in (x_{k-1}, x_k)}$ for all $k \in \{1, \ldots, n\}$ and let

$$\overline{\int_a^b} f(x)dx := \sup\{L(f, p) | p \text{ is a partition of } [a, b]\}$$
$$\underline{\int_a^b} f(x)dx := \inf\{U(f, p) | p \text{ is a partition of } [a, b]\}.$$

Definition 2.2.1. A function $f : [a, b] \to \mathbb{R}$ is *Riemann integrable* if

$$\overline{\int_a^b} f(x)dx = \underline{\int_a^b} f(x)dx.$$

In this case, define the *Riemann integral of* f as

$$\int_a^b f(x)dx := \overline{\int_a^b} f(x)dx.$$

A well-known result is that if a function f is Riemann integrable, it can be approximated by a family of *step functions*. Formally, a function $\varphi : [a, b] \to \mathbb{R}$ is a *step function* if it takes form of

$$\varphi(x) = \sum_{k=1}^{n} c_k \mathbf{1}\{x \in (x_{k-1}, x_k)\},$$

for some $\{c_k\}_{k=1}^n \subset \mathbb{R}$ and some partition $p = \{x_k\}_{k=1}^n$. By definition, all step functions are Riemann integrable and the integral is

$$\int_a^b \varphi(x)dx = \sum_{k=1}^{n} c_k(x_k - x_{k-1}).$$

Proposition 2.2.1. *Let $f : [a, b] \to \mathbb{R}$ be a Riemann integrable function. Then there exists a sequence of step functions $\{\varphi_n\}$ such that*

$$\lim_{n \to \infty} \int_a^b \varphi_n(x)dx = \int_a^b f(x)dx.$$

Although simple and intuitive, there are many functions that are not Riemann integrable. Consider the following example:

Example 2.2.1 (Dirichlet's Function). Let $f : [0,1] \to [0,1]$ be defined as

$$f(x) := \mathbf{1}\{x \in \mathbb{Q}\}.$$

It can be shown that f is not Riemann integrable.

In fact, there is a characterization for Riemann integrable functions.

Proposition 2.2.2. *A real-valued function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if f is continuous Lebesgue-almost everywhere on $[a,b]$.*

To address this limitation of Riemann integrable, Lebesgue integral is then developed. Conceptually, as Riemann integral approximate the area below the graph of a function by "chopping" the *domain*, Lebesgue integral approximates the same thing by "chopping" the *range*. Lebesgue integral can exist for functions that behaves discontinuously on it's domain, as in Example 2.2.1. We now start construct Lebesgue integral formally. From the spirit of Proposition 2.2.1, we wish to "approximate" the integral by a class of relatively more tractable functions. While the Riemann integral uses step functions, we will not use *simple functions*.

Fix a measurable set $X \subseteq \mathbb{R}$, let $\varphi : X \to \mathbb{R}$ be a simple function. By definition, we can write φ as

$$\varphi(x) = \sum_{k=1}^{n} a_k \mathbf{1}\{x \in E_k\},$$

for some $n \in \mathbb{N}$, where $\{a_k\}_{k=1}^n$ is a collection of distinct real numbers and $\{E_k\}_{k=1}^n$ is a collection of disjoint measurable subsets in X. Now define the Lebesgue integral for simple functions as:

$$\int_X \varphi d\lambda := \sum_{k=1}^{n} a_k \lambda(E_k).$$

Now, by the simple approximation theorem, for any positive measurable real-valued function f on X, there exists a sequence of simple functions $\{\varphi_n\}$ with $\varphi_n \leq f$ for all $n \in \mathbb{N}$ such that $\{\varphi_n\}$ converges to f pointwisely and $0 \leq \varphi_n \leq \varphi_{n+1}$ for all $n \in \mathbb{N}$. Notice that for each $n \in \mathbb{N}$,

$$\int_X \varphi_n d\lambda = \sum_{k=1}^{m(n)} a_k^n \mathbf{1}\{x \in E_k^n\} < \infty$$

is well-defined. Furthermore, since $0 \leq \varphi_n \leq \varphi_{n+1}$, the sequence $\{\int_X \varphi_n d\lambda\}$ is increasing and thus its limit exists (could be ∞). As such, we can then define the Lebesgue integral for nonnegative measurable functions as

$$\int_X f d\lambda := \lim_{n\to\infty} \int_X \varphi_n d\lambda.$$

Finally, for any measurable function $f : X \to \mathbb{R}$, let $f^+ := \max\{f(x), 0\}$, $f^- := -\min\{f(x), 0\}$. Since f is measurable, f^+ and f^- are also measruable. and both are nonnegative. Moreover,

$$f = f^+ - f^-.$$

Therefore, for any measurable function $f : X \to \mathbb{R}$, if

$$\int_X f^+ d\lambda < \infty \text{ or } \int_X f^- d\lambda < \infty$$

we can properly define the Lebesgue integral for measurable functions as:

$$\int_X f d\lambda := \int_X f^+ d\lambda - \int_X f^- d\lambda.$$

With the constructions above, we say that a measurable function is *Lebesgue integrable* if

$$\int_X |f| d\lambda = \int_X (f^+ + f^-) d\lambda < \infty$$

and define its Lebesgue integral as

$$\int_X f d\lambda := \int_X f^+ d\lambda - \int_X f^- d\lambda.$$

Several properties of Lebesgue integral can be derived from its definition and properties of Lebesgue measure.

Proposition 2.2.3. *Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be two real-value functions that are Lebesgue integrable. Then:*

1. *(Monotonicity)*
$$f \geq g \Rightarrow \int_X f d\lambda \geq \int_X g d\lambda.$$
2. *(Linearity) For any $\alpha, \beta \in \mathbb{R}$, $\alpha f + \beta g$ is integrable and*
$$\int_X (\alpha f + \beta g) d\lambda = \alpha \int_X f d\lambda + \beta \int_X g d\lambda.$$

3. *(Characterization of integrability) f is integrable if and only if for any $\varepsilon > 0$, there exists $\delta > 0$ such that*
$$\int_A f d\lambda < \varepsilon$$
for any $A \in \mathcal{M}$ with $\lambda(A) < \delta$.

4. *(Countable additivity) For any disjoint collection $\{E_n\} \subset \mathcal{M}$,*
$$\int_{\cup_{n=1}^{\infty} E_n} f d\lambda = \sum_{n=1}^{\infty} \int_{E_n} f d\lambda.$$

5. *(Almost everywhere equivalence) For any $A \in \mathcal{M}$ with $\lambda(A) = 0$,*
$$\int_{X \setminus A} f d\lambda = \int_X f d\lambda.$$

6. *(Converse of almost everywhere equivalence) If $f \geq 0$, then*
$$\int_X f d\lambda = 0 \Rightarrow f \equiv 0 \text{ almost everywhere .}$$

As a Corollary of Proposition 2.2.3, Dirichlet's function is Lebesgue integrable and its integral is zero, as
$$\int_{[0,1]} f d\lambda = \int_{[0,1]\setminus\mathbb{Q}} f d\lambda = 0$$

The following Proposition, together with the previous observation and Example 2.2.1, show that Lebesgue integral is strictly better than Riemann integral.

Proposition 2.2.4. *Let $f : [a, b] \to \mathbb{R}$ be a Riemann integrable function. Then f is Lebesgue integrable and*
$$\int_a^b f(x)dx = \int_{[a,b]} f d\lambda$$

Remark 2.2.1. As noted in Proposition 2.2.4, by 5. in Proposition 2.2.3, since $\lambda(\{a\}) = \lambda(\{b\}) = 0$ for any $a, b \in \mathbb{R}$. For any Riemann integrable function $f : [a, b] \to \mathbb{R}$,
$$\int_a^b f(x)dx = \int_{[a,b]} f d\lambda = \int_{(a,b]} f d\lambda = \int_{[a,b)} f d\lambda = \int_{(a,b)} f d\lambda.$$
Therefore, there is no confusion to interchange the notations between Riemann integral and Lebesgue integral when a function is Riemann integral. In the rest of this note, we will use

these notations interchangeably when there is no confusion. Furthermore, it is sometimes informative to write

$$\int_X f d\lambda = \int_X f(x)dx = \int_X f(x)\lambda(dx).$$

As λ is a measure constructed from length of intervals, it is natural to regard $\lambda(dx)$ as dx.

Remark 2.2.2. At this point, we can revisit the normed linear space $L^p([0,1])$ before. When introducing this space, we did not explicitly discussed the integration we are using. In fact, all the integrals used when defining $L^p([0,1])$ and its norm should be regarded as *Lebesgue integrals*. In fact, for any measurable set $X \subseteq \mathbb{R}$, any $p \in [1,\infty)$ we can define

$$L^p(X) := \left\{ f : X \to \mathbb{R} \middle| f \text{ measurable and } \left(\int_X |f|^p d\lambda \right)^{\frac{1}{p}} < \infty \right\}$$

and the norm $||\cdot||_p$ is given by:

$$||f||_p := \left(\int_X |f|^p d\lambda \right)^{\frac{1}{p}}.$$

Notice that by this definition, $L^1(X)$ is exactly the collection of integrable functions on X.

Under L^p spaces, a useful inequality is called *Hölder inequality*, which states that for any $f \in L^p(X), g \in L^q(X)$ with $p \in (1,\infty)$ and $1/p + 1/q = 1$, we have:

$$\int_X |fg| d\lambda \leq ||f||_p \cdot ||g||_q, \tag{4}$$

of which a special case when $p = 2$ is called the *Cauchy-Schwartz inequality*.

2.2.2 Convergence Theorems

One of the common questions related to integration is when can we change the order between integration and limits. For instance, let $X \sim U[0,1]$ be a random variable with uniform distribution, it is well known that for any measurable function $f : [0,1] \to \mathbb{R}$, the *expectation* of $f(X)$ is given by:

$$\mathbb{E}[f(X)] = \int_0^1 f(x)dx.$$

Now the question is, suppose that $\{f_n\}$ converges to some $f : [0,1] \to \mathbb{R}$. When can we say that $\mathbb{E}[f_n(X)]$ converges to $\mathbb{E}[f(X)]$. For example, we may think of $f_n(x)$ as the output of firm $x \in [0,1]$ in the economy at day $n \in \mathbb{N}$ and there are $[0,1]$ many firms in the economy.

Suppose that for each firm x, its profit eventually converges to its long-run output $f(x)$. Can we say that the aggregate output $\int_0^1 f_n(x)dx$ converges as well? In this section we will introduce three of the most widely used convergence theorems that ensures the validity of changing the order between integral and limits.

The following Lemma can sometimes be useful and is essential to the proofs of the following theorems

Lemma 2.2.1 (Fatou's Lemma)**.** *For any sequence of nonnegavie measurable functions* $\{f_n\}$ *on* X *such that* $\{f_n\}$ *converges pointwisely almost everywhere for some* $f : X \to \mathbb{R}$*,* f *is integrable and*

$$\int_X f d\lambda \leq \liminf_{n\to\infty} \int_X f_n d\lambda$$

Theorem 2.2.1 (Monotone Convergence Theorem)**.** *For any sequence of nonnegative measurable functions* $\{f_n\}$ *such that* $0 \leq f_n \leq f_{n+1}$ *for all* $n \in \mathbb{N}$ *and that* $\{f_n\}$ *converges poitnwisely almost everywhere to some* $f : X \to \mathbb{R}$*,* f *is integrable and*

$$\lim_{n\to\infty} \int_X f_n d\lambda = \int_X f d\lambda$$

Theorem 2.2.2 (Dominance Convergence Theorem)**.** *For any sequence of measurable functions* $\{f_n\}$ *that converges pointwisely almost everywhere to some* $f : X \to \mathbb{R}$*, if there exists an integrable function* $g : X \to \mathbb{R}$ *such that* $|f_n| \leq |g|$ *for all* $n \in \mathbb{N}$*, then* f *is integrable*

$$\lim_{n\to\infty} \int_X f_n d\lambda = \int_X f d\lambda.$$

The next convergence theorem requires an additional concept called *uniform integrability*

Definition 2.2.2. A collection of real-valued measurable functions on X is said to be *uniformly integrable* if for any $\varepsilon > 0$, there exists $\delta > 0$ such that whenever $A \subseteq X$ is measurable and $\lambda(A) < \delta$,

$$\int_X f d\lambda < \varepsilon$$

for all $f \in \mathcal{F}$.

Definition 2.2.3. A collection of real-valued measurable functions on X is said to be *tight* if for any $\varepsilon > 0$, there exists $X_0 \subset X$ with $\lambda(X_0) < \infty$ and

$$\int_{X\setminus X_0} f d\lambda < \varepsilon.$$

Theorem 2.2.3 (Vitali's Convergence Theorem)**.** *Let $\{f_n\}$ be a sequence of real-valued measurable functions that is uniformly integrable such that $\{f_n\}$ converges to some $f : X \to \mathbb{R}$ pointwisely almost everywhere. Then f is integrable and*

$$\lim_{n\to\infty} \int_X f_n d\lambda = \int_X f d\lambda$$

if $\lambda(X) < \infty$ or $\{f_n\}$ is tight.

Remark 2.2.3. As noted before, in infinite-dimensional normed linear spaces, we no longer have the property that a set is compact if and only if it is bounded and closed. However, in many economic problems, we might still encounter maximization problems in which the choice variable is infinite dimensional. It is then more difficult to apply Proposition 1.2.5 for existences of solution, in that it is less obvious to verify whether the domain of the objective function is compact. Nevertheless, there are several well-known subsets in infinite-dimensional spaces that are indeed compact. We will introduce one of them here, by using the dominant convergent theorem.

Let $\mathcal{F}$ be a class of *increasing functions* that are uniformly bounded on $[0, 1]$. That is, for any $f \in \mathcal{F}$, $|f| \leq M$ for some $M \in \mathbb{R}$. The following theorem is critical to our result.

Theorem 2.2.4 (Helley's selection theorem)**.** *For any sequence $\{f_n\} \subset \mathcal{F}$, there exists a subsequence $\{f_{n_k}\}$ such that $\{f_{n_k}\}$ converges pointwisely to some $f \in \mathcal{F}$.*

As such, for any bounded sequences $\{f_n\} \subset \mathcal{F}$, take the pointwise convergent subsequence $\{f_{n_k}\}$. Since $\mathcal{F}$ is uniformly bounded, by the dominant convergent theorem, for any $p \in [1, \infty)$,

$$\lim_{k\to\infty} \int_0^1 |f_{n_k}(x) - f(x)|^p dx = 0.$$

As such, there exists a convergent subsequence $\{f_{n_k}\}$ that converges to some $f \in \mathcal{F}$ under the $||\cdot||_p$ norm. That is, the subset $\mathcal{F} \subset L^p([0, 1])$ is compact under the $||\cdot||_p$ norm for all $p \in [1, \infty)$.

2.2.3 Operational Rules

Although we have rigorously defined Lebesgue integral, it is often difficult to compute integration of particular functions analytically. We now review some of the important operational rules that you are familiar with that will be useful under some economic contexts.

Proposition 2.2.5 (Integration by parts)**.** *Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be two integrable functions. Suppose that f and g are differentiable on X and f', g' are integrable. Then*

$$\int_a^b f(x)g'(x)dx = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x)dx.$$

Example 2.2.2 (Stochastic dominance and characterization)*.* Many economic analyses involves uncertainty, sometimes we wish to compare two different lotteries in a meaningful way. *Stochastic dominance* order is a commonly used method. For instance, when making investment decisions, the notion of stochastic dominance allows us to (partially) rank assets that have different distribution of returns. We will discuss and characterize two notions of stochastic dominance—first order and second order. Let X and Y be two random variables on $\mathbb{R}$ and let F and G denote the *cumulative distribution function*, or *CDF*, of X and Y, respectively, that have a common support $[a, b]$.[15]

Definition 2.2.4. We say that F *first-order stochastic dominates* G if for all $x \in \mathbb{R}$, $F(x) \leq G(x)$.

It is will be more intuitive to rewrite the condition into $1 - F(x) \geq 1 - G(x)$ for all $x \in \mathbb{R}$ when F first-order stochastic dominates G. Recall that for each $x \in \mathbb{R}$, $1 - F(x)$ and $1 - G(x)$ are the *probabilities* that the random variables X and Y are greater than x, respectively. Therefore, X first order stochastic dominates Y means that for each possible value x, the probability that X has a realization greater than x is larger than that of Y.

Definition 2.2.5. We say that F *second-order stochastic dominates* G if for all $x \in \mathbb{R}$, $\int_a^x F(t)dt \leq \int_a^x G(t)dt$.

Intuitively, while first-order stochastic dominance describes the notion of ranking in *levels*, second order stochastic dominance describes the notion of *dispersion*. The following characterizations reflects this intuition and is a consequence of integration by parts.

Proposition 2.2.6. *Let F and G be two cumulative distribution functions of random variables X, Y that admits densities f and g, respectively. Then F first order stochastic domi-*

[15]Recall that for a random variable X, the CDF of X is given by $F(x) := \mathbb{P}(X \leq x)$.

nates G if and only if for any increasing function u that is differentiable,[16]

$$\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$$

Proof. Recall that since F and G have densities f and g,

$$\mathbb{E}[u(X)] - \mathbb{E}[u(Y)] = \int_a^b u(x)f(x)dx - \int_a^b u(x)g(x)dx = \int_a^b u(x)(f(x) - g(x))dx.$$

Using integration by parts, we have:

$$\begin{aligned}
&\int_a^b u(x)(f(x) - g(x))dx \\
=&u(b)(F(b) - G(b)) - u(a)(F(a) - G(a)) - \int_a^b u'(x)(F(x) - G(x))dx \\
=&\int_a^b u'(x)(G(x) - F(x))dx,
\end{aligned}$$

where the second equality follows from the assumption that F and G have common support $[a, b]$ and thus $F(b) = G(b) = 1$, $F(a) = G(a) = 0$. Now suppose that F first order stochastic dominates G. Then for any increasing u, since $u' \geq 0$,

$$\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)] = \int_a^b u'(x)(G(x) - F(x))dx \geq 0.$$

Conversely, suppose that for any increasing u,

$$\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)] = \int_a^b u'(x)(G(x) - F(x))dx \geq 0$$

and suppose contrarily that $F(x_0) > G(x_0)$ for some $x_0 \in [a, b]$. Since F and G are continuous, there exists an interval I on which $F > G$. Now define $v'(x) := \mathbf{1}\{x \in I\}$ and let $v(x) := \int_0^x v'(t)dt$. Clearly v is increasing and yet

$$\int_a^b v'(x)(G(x) - F(x))dx = \int_I v'(x)(G(x) - F(x))dx < 0,$$

a contradiction. ■

[16] In fact, differentiability of u is a redundant assumption. As we will see later in this chapter, increasing functions are differentiable almost everywhere. Also, existence of density is not necessary, as we will see in chapter 4 when general integrations are introduced.

Proposition 2.2.7. *Let F and G be two cumulative distribution functions of random variables X, Y that admits densities f and g, respectively. Then F second order stochastic dominates G if and only if for any increasing concave function u that is twice differentiable,*

$$\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)].$$

Proof. Using definition of expectations and using integration by parts twice, we have:

$$\begin{aligned}
&\mathbb{E}[u(X)] - \mathbb{E}[u(Y)] \\
&= \int_a^b u(x)(f(x) - g(x))dx \\
&= u(b)(F(b) - G(b)) - u(a)(F(a) - G(a)) - \int_a^b u'(x)(F(x) - G(x))dx \\
&\qquad\qquad \text{(Integration by parts)} \\
&= -u'(b)\left(\int_a^b F(x)dx - \int_a^b G(x)dx\right) + \int_a^b u''(x)\left(\int_a^x F(t)dt - \int_a^x G(t)dt\right)dx \\
&\qquad\qquad \text{(Integration by parts again)}
\end{aligned}$$

Now suppose that F second-order stochastic dominates G. Then for any increasing and concave u, $u' \geq 0$ and $u'' \leq 0$. Therefore,

$$\mathbb{E}[u(X)] - \mathbb{E}[u(Y)] = -u'(b)\left(\int_a^b F(x)dx - \int_a^b G(x)dx\right) + \int_a^b u''(x)\left(\int_a^x F(t)dt - \int_a^x G(t)dt\right)dx \geq 0.$$

Conversely, suppose that for any increasing and concave u,

$$\mathbb{E}[u(X)] - \mathbb{E}[u(Y)] = -u'(b)\left(\int_a^b F(x)dx - \int_a^b G(x)dx\right) + \int_a^b u''(x)\left(\int_a^x F(t)dt - \int_a^x G(t)dt\right)dx \geq 0$$

and suppose that $\int_a^{x_0} F(t)dt > \int_a^{x_0} G(t)dt$ for some $x_0 \in [a, b]$. If $x_0 = b$, we can then take $u(x) := (x - a)/(b - a)$. Then

$$-u'(b)\left(\int_a^b F(x)dx - \int_a^b G(x)dx\right) + \int_a^b u''(x)\left(\int_a^x F(t)dt - \int_a^x G(t)dt\right)dx < 0,$$

a contradiction. If $x_0 < b$, and $\int_a^b F(t)dt \leq \int_a^b G(t)dt$, then there exists an interval I such that $\int_a^x F(t)dt > \int_a^x G(t)dt$ for all $x \in I$ and $b \notin \text{cl}(I)$. Define $v''(x) := -\mathbf{1}\{x \in I\}$, $v'(x) := \int_a^x -v''(t)dt$, $v(x) := \int_a^x v'(t)dt$. Then v is increasing and concave and $v'(b) = 0$. Thus,

$$-v'(b)\left(\int_a^b F(x)dx - \int_a^b G(x)dx\right) + \int_a^b v''(x)\left(\int_a^x F(t)dt - \int_a^x G(t)dt\right)dx < 0,$$

a contradiction. ∎

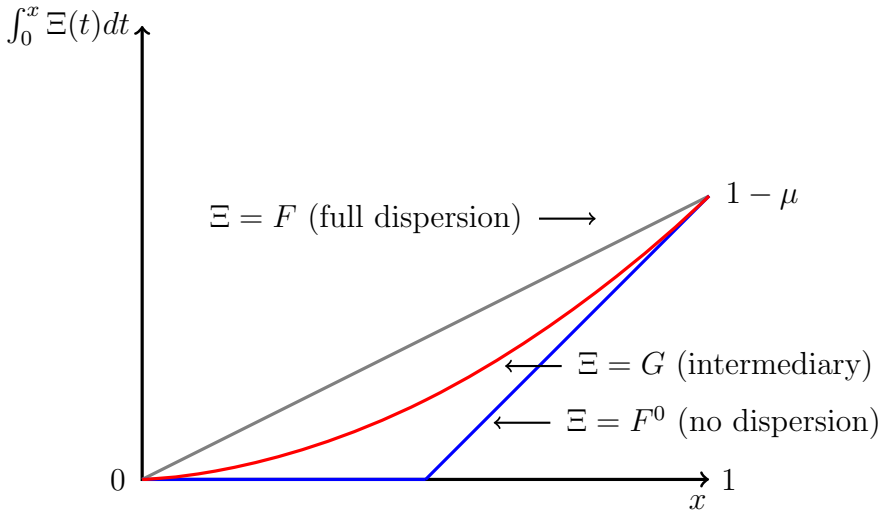
Proposition 2.2.6 and Proposition 2.2.7 are intuitive to interpret. A distribution F first-order stochastic dominates G if and only if for any decision maker who prefers more X, she is better-off under lottery F. A distribution F second-order stochastic dominates G if and only if for any *risk averse* decision maker who prefers more X, she is better-off under lottery G. In fact, if we consider a relation under which F second-order stochastic dominates G and yet they have the same mean, which is referred as G is a *mean preserving spread* of F, Proposition 2.2.7 still holds, which is exactly the description of risk-aversion. Using integration by parts, we know that

$$\mathbb{E}[X] = \int_a^b f(x)dx = b - \int_a^b F(t)dt.$$

Therefore, if G is a mean preserving spread of F with mean μ, we know that

$$\int_a^x F(t)dt \leq \int_a^x G(t)dt$$

for all $x \in [a, b]$ and the equality holds when $x = b$. As such, it is easy to visualize the collection of distributions that are a mean preserving spread of μ. As shown in the below figure.



In this figure, the horizontal axis is x and the vertical axis is integral of a CDF, $\int_a^x G(t)dt$. Since all CDFs are increasing, we know that each CDF must corresponds to a convex function. Furthermore, the degenerate distribution that has a mass 1 at μ is the convex function at the bottom, as this corresponds to a CDF $F^0(x) = \mathbf{1}\{x < \mu\}$. The most-dispersed CDF F is

the one that puts probability μ on 1 and probability $1-\mu$ on 0, which corresponds to the top convex function. As noted above, the collection of distributions that are a mean preserving spread of μ corresponds to all the convex functions on the graph that are below the bottom curve and below the top curve and are sharing the same end points, and we can visualize the ranking given by second-order stochastic dominance by just examining the levels of these convex functions.

So far, we have been dealing with integration on $\mathbb{R}$. In fact, the approach that we used to construct Lebesgue integral is more general. In particular, it can also be used to construct Lebesgue integral on $\mathbb{R}^n$ for any $n \in \mathbb{N}$. Rather than using intervals and length of interval to construct measures, we can use *cubes*—product of intervals—and *volume* of cubes to construct measures on multi-(but finite) dimensional space. With this construction, we then have the Lebesgue measure on $\mathbb{R}^n$ and the Lebesgue integrals on $\mathbb{R}^n$ can be defined analogously. We let λ^n denote the *Lebesgue measure* on $\mathbb{R}^n$ and denote the Lebesgue integral on $\mathbb{R}^n$ as

$$\int_X f d\lambda^n.$$

for integrable $f : X \to \mathbb{R}$ and $X \subset \mathbb{R}^n$ that is measurable.

Nevertheless, we know few about computing integrals on $\mathbb{R}^n$. Fortunately, the following theorem allows us to calculate and integral on $\mathbb{R}^n$ *separately*.

Theorem 2.2.5 (Fubini's Theorem)**.** *Let $X \subseteq \mathbb{R}^m$ and $Y \subseteq \mathbb{R}^k$ be two measurable subsets. Suppose that $f : X \times Y \to \mathbb{R}$ is Lebesgue measurable. Then for (Lebesgue) almost all $x \in X$, the function $f(x, \cdot) : Y \to \mathbb{R}$ is Lebesgue integrable over Y and*

$$\int_{X\times Y} f d\lambda^{m+k} = \int_X \left(\int_Y f(x, y)\lambda^k(dy) \right) \lambda^m(dx).$$

Fubini's theorem allows us to decompose a multidimensional integration into many one-dimensional integrations, which we are more familiar with. We will use Fubini's theorem in an example at the end of this chapter. For now, let us illustrate the computation by the follwoing example.

Example 2.2.3. Let X, Y be two independent random variables with (marginal) densities f and g, respectively. Suppose that X and Y have support on $\mathbb{R}$ By independence, we know

that the joint density is simply fg. We wish to derive the CDF of the random variable $Z := X + Y$, denoted by H. By definition, $H(z) = \mathbb{P}(X + Y \leq z)$ for all $z \in \mathbb{R}$. Therefore,

$$H(z) = \int_{\{(x,y)\in\mathbb{R}^2 | x+y\leq z\}} f(x)g(y)\lambda^2(dx, dy).$$

By Fubini's theorem, we can write the integral separately as a double integral and each of them is on $\mathbb{R}$. That is:

$$H(z) = \int_{\{(x,y)\in\mathbb{R}^2 | x+y\leq z\}} f(x)g(y)\lambda^2(dx, dy) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f(x)dx \right) dy.$$

Let F denote the CDF of f, we than have:

$$H(z) = \int_{\mathbb{R}} F(z - y)g(y)dy = \mathbb{E}[F(z - Y)]$$

for all $z \in \mathbb{R}$.

2.3 Differentiation

While integration deals with aggregating a function, differentiations allow us to examine a function's local behavior. As you are already familiar with, differentiation is closely related to the notion of *tangency* geometrically and captures the local changes of a function. Furthermore, differentiation and integration are closely related, linked by the *Fundamental Theorem of Calculus*. In this section, we will introduce the notion of differentiation and the related differentiability results. Finally, we will introduce the Fundamental Theorem of Calculus. Throughout this section, we will still take and fix a measurable set $X \subseteq \mathbb{R}$ and assume that X has interior points.

2.3.1 Definition and Differentiability

Let $f : X \to \mathbb{R}$ be a real-valued function, for any $x \in \text{int}(X)$, we define the *upper* and *lower* derivatives of f at x as:

$$\overline{D}f(x) := \lim_{\delta\to 0} \left[\sup_{t\in[-\delta,\delta]} \frac{f(x + t) - f(x)}{t} \right]$$

and

$$\underline{D}f(x) := \lim_{\delta\to 0} \left[\inf_{t\in[-\delta,\delta]} \frac{f(x + t) - f(x)}{t} \right]$$

Notice that both $\overline{D}(f)(x)$ and $\underline{D}f(x)$ exist (could be ∞ or $-\infty$) We say that f is *differentiable* at $x \in \text{int}(X)$ if $\overline{D}f(x) = \underline{D}f(x)$ and define the *derivative* of f at x as:

$$f'(x) := \overline{D}f(x).$$

Furthermore, we say that f is differentiable on $\text{int}(X)$ if it is differentiable at all $x \in \text{int}(X)$, and that f is twice differentiable at $x \in \text{int}(X)$ if its derivative is differentiable at x. It is then clear that for any $f : X \to \mathbb{R}$ that is differentiable at $x \in \text{int}(X)$,

$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}.$$

As such, the derivative of f at x is the limiting slope of line segments that crosses $(x, f(x))$, and is the slope of the *tangent* line.

Although many functions fail to be differentiable, the following Propositions give some sufficient conditions for (almost-everywhere) differentiability.

Proposition 2.3.1. *Let $f : [a, b] \to \mathbb{R}$ be an increasing function. Then f is differentiable (Lebesgue) almost everywhere.*

Proposition 2.3.2. *Let $f : [a, b] \to \mathbb{R}$ be a convex function. Then f is differentiable Lebesgue almost everywhere.*

However, (almost everywhere) differentiability is sometimes not enough for connecting differentiation and integration. In particular, there might be cases when f is differentiable but f' is not integrable. Or when f is differentiable (Lebesgue) almost everywhere on $[a, b]$ but $f(x) \neq \int_a^x f'(t)dt$. Below are two examples:

Example 2.3.1. Let $f : [0, 1] \to \mathbb{R}$ be defined by

$$\begin{cases} x \sin(1/x), & \text{if } x \in (0, 1] \\ 0, & \text{if } x = 0 \end{cases}.$$

It can be shown that f is differentiable on $(0, 1)$ but its derivative f' is not integrable.

Example 2.3.2 (Cantor function). Let $X = [0, 1]$, we first start by defining $D_{0,1} := (1/3, 2/3)$, which is the middle third interval of X. Remove $D_{0,1}$ from X. Now define the middle third

intervals $D_{1,1} : (1/9, 2/9)$ and $D_{1,2} := (7/9, 8/9)$ of the two parts of $X \backslash D_{0,1}$. Inductively, we have removed an open set

$$D := \bigcup_{n=1}^{\infty} \bigcup_{k=1}^{2^n} D_{n,k}.$$

Let $C := X \backslash D$. C is called the *Cantor set*. As remarked before, C is in fact uncountable but has Lebesgue measure zero.

Now we will construct a special function on X. Let a function $f : X \to [0, 1]$ be defined as follows: $f_0(x) := x$ for all $x \in X$. For each $n \in \mathbb{N}$ define f_n as

$$f_n(x) := \begin{cases} \frac{1}{2} f_{n-1}(3x), & \text{if } x \in [0, \frac{1}{3}] \\ \frac{1}{2}, & \text{if } x \in [\frac{1}{3}, \frac{2}{3}] \\ \frac{1}{2} + \frac{1}{2} f_{n-1}(3x - 2), & \text{if } x \in [\frac{2}{3}, 1] \end{cases}.$$

It can be shown that $\{f_n\}$ converges pointwisely to some function $f : X \to [0, 1]$ and this function is increasing, continuous and takes a constant value on every $D_{n,k}$, with $f(0) = 0$ and $f(1) = 1$. Since $\lambda(C) = 0$ and f is a constant on all $D_{n,k}$. f is differentiable almost everywhere and $f' \equiv 0$ almost every where, however, it is clear that $f(x) \neq \int_0^x f'(t)dt$ for all $x \in (0, 1)$.

2.3.2 Absolute Continuity and the Fundamental Theorem of Calculus

The issues with Example 2.3.1 and Example 2.3.2 is that these functions, although differentiable (almost everywhere), varies "too much". To avoid these issues, the functions we consider has to be continuous enough. We will now introduce the notion of *absolute continuity*.

Definition 2.3.1. Let $f : [a, b] \to \mathbb{R}$ be a real-valued function. f is said to be *absolute continuous* if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for every finite disjoint collection of intervals $\{(a_k, b_k)\}_{k=1}^n$ such that $\sum_{k=1}^n |b_k - a_k| < \delta$,

$$\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \varepsilon$$

The following notion is closely related to absolute continuity and is sometimes easier to verify.

Definition 2.3.2. Let $f : [a, b] \to \mathbb{R}$ be a real-valued function. f is *Lipschitz continuous* on $[a, b]$ if there exits a $K \in \mathbb{R}_+$ such that for any $x, y \in [a, b]$,

$$|f(x) - f(y)| \leq K|x - y|.$$

It is clear from definitions that a function is absolutely continuous if it is Lipschitz continuous. Furthermore, it can be shown that convex functions on an interval (a, b) is Lipschitz continuous on any $[c, d] \subset (a, b)$. With absolute continuity, we are now ensured with the Fundamental Theorem of Calculus.

Theorem 2.3.1 (Fundamental Theorem of Calculus)**.** *Let $f : [a, b] \to \mathbb{R}$ be a real-valued absolutely continuous function. Then f is differentiable almost everywhere and its derivative f', is integrable on $[a, b]$ and*

$$f(b) - f(a) = \int_a^b f'(x)dx$$

Notice that in Theorem 2.3.1, the interval $[a, b]$ is arbitrary. An immediate corollary of it can be derived if we fix an interval $[a, b]$ and take any $x \in (a, b]$ and apply the theorem. After rearranging, we have:

$$f(x) = f(a) + \int_a^x f'(t)dt,$$

which is now the canonical form of the Fundamental Theorem of Calculus.

Now we can revisit Example 2.3.1 and Example 2.3.2. In both examples, the function f are not absolutely continuous, which is the reason why the Fundamental Theorem of Calculus fails. In fact, we notice that the Cantor function is continuous and increasing, and therefore is a valid CDF for some random variable C with support $[0, 1]$. However, although being almost everywhere differentiable,

$$\mathbb{P}(C \leq x) := f(x) \neq \int_0^x f'(t)dt.$$

This tells us that even if a CDF is continuous and differentiable almost everywhere, if it fails to be *absolutely continuous*, it's derivative is **NOT** it's density. In fact, such distributions do not have a densities at all! We will return to this subject in chapter 4.

2.3.3 Differentiation of Functions on $\mathbb{R}^n$

So far we have been studying differentiation for functions on $\mathbb{R}$. Similar concepts can be extended to functions on multi-dimensional spaces. We will now examine the case when the domain is finite-dimensional. Differentiations of infinite-dimensional functions are beyond the scope of this course, you will see some special cases in the second part. To begin with, let us start with the notion of *partial derivative*. As before, we take and fix a subset $X \subset \mathbb{R}^n$ and suppose that $\text{int}(X) \neq \emptyset$.

Definition 2.3.3. Let $f : X \to \mathbb{R}$ be a real-valued function. We say that f has a *partial derivative* with respect to the i-th component at $\mathbf{x}^0 = (x_i^0)_{i=1}^n \in \text{int}(X)$ if

$$\lim_{\delta \to 0} \frac{f(x_1^0, \ldots, x_i^0 + \delta, \ldots, x_n^0) - f(\mathbf{x}^0)}{\delta}$$

exists. In this case, we denote the limit as

$$\left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}^0} := \lim_{\delta \to 0} \frac{f(x_1^0, \ldots, x_i^0 + \delta, \ldots, x_n^0) - f(x^0)}{\delta}$$

Just as Fubini's theorem allows us to regard integral of multivariate function as separated integrals on each one-dimensional subspace, the notion of partial derivative is to *fix* other variables and examine the one-dimensional derivative of a function on a projected space. As a shorthand, we sometimes denote the i-th partial derivative of a function f at $x^0 \in X$ as $f_i(x^0)$. Also, we use

$$\nabla f(\mathbf{x}^0) := \begin{pmatrix} \left.\frac{\partial f}{\partial x_1}\right|_{\mathbf{x}=\mathbf{x}^0} \\ \vdots \\ \left.\frac{\partial f}{\partial x_n}\right|_{\mathbf{x}=\mathbf{x}^0} \end{pmatrix}$$

to denote the vector of derivatives of all components, provided that they all exist, which is called the *gradient vector* of f at x^0.

With multivariate functions, there is one more complication when thinking about the changes as the variables change. In additional to the changes in each dimension when other variables are held fixed, we can also discuss the change of the function when the variables change in a particular *direction*. That is, it is sometimes useful to examine how $f(\mathbf{x})$ changes when $\mathbf{x}$ changes to $\mathbf{x} + r\mathbf{v}$ for some $r \in \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$. To this end, we define the notion of *directional derivative* as follows:

Definition 2.3.4. Let $f : X \to \mathbb{R}$ be a real-valued function. For any $\mathbf{x}^0 \in \text{int}(X)$ and any $\mathbf{v} \in \mathbb{R}^n$, the value:

$$Df_{\mathbf{v}}(\mathbf{x}^0) := \lim_{\delta \to 0} \frac{f(\mathbf{x}^0 + \delta \mathbf{v}) - f(\mathbf{x}^0)}{\delta},$$

provided that it exists, is called the *directional derivative* toward $\mathbf{v}$ of f at $\mathbf{x}^0$.

As a remark, with multivariate functions, existence of gradient vector does not imply differentiability and existence of Hessian matrix does not imply twice differentiability. We hereby define differentiability for multivariate functions.

Definition 2.3.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function. f is *differentiable* at $\mathbf{x}^0$ if there exists a linear functional $L : \mathbb{R}^n \to \mathbb{R}$ such that

$$\lim_{\mathbf{v} \to 0} \frac{|f(\mathbf{x}^0 + \mathbf{v}) - f(\mathbf{x}^0) - L(\mathbf{v})|}{||\mathbf{v}||}$$

.

If f is differentiable at $\mathbf{x}^0$, then the gradient vector exists at $\mathbf{x}^0$ and the linear functional is precisely $\mathbf{v} \mapsto \nabla f(\mathbf{x}^0)^\intercal \mathbf{v}$. However, it is possible that the gradient vector exists at $\mathbf{x}^0$ but f is not differentiable at $\mathbf{x}^0$. We also say that f is twice-differentiable if all the partial derivatives are differentiable.

As a notation, hereafter, we let $C^k(X)$ denote the collection of functions that are k-times differentiable on $X \subseteq \mathbb{R}^n$ and each partial derivatives are continuous.

Using the chain rule of one-dimensional derivative, it is easy to verify that

$$Df_{\mathbf{v}}(\mathbf{x}) = \nabla f(\mathbf{x})^\intercal \mathbf{v},$$

for all $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{x} \in \text{int}(X)$ if f is differentiable. That is, the directional derivative is the inner product of the gradient vector and the direction.

Two immediate properties about the gradient vector can then be derived from the above observation.

Proposition 2.3.3. *Let $f : X \to \mathbb{R}$ be a real-valued function and suppose that the directional derivative of f toward every direction exists at $\mathbf{x} \in \text{int}(X)$ and that gradient vector of f exists at $\mathbf{x}$. Then*

$$\left\{ \frac{1}{||\nabla f(\mathbf{x})||} \nabla f(\mathbf{x}) \right\} = \underset{\{\mathbf{v} \in \mathbb{R}^n | ||\mathbf{v}|| = 1\}}{\text{argmax}} |Df_{\mathbf{v}}(\mathbf{x})|,$$

where $||\cdot||$ *is the Euclidean norm.*

Proof. Using the above observation and by the Cauchy Schwartz inequality, for any $\mathbf{v} \in \mathbb{R}^n$ with $||\mathbf{v}|| = 1$,

$$|Df_{\mathbf{v}}(\mathbf{x})| = |\nabla f(\mathbf{x})^\intercal \mathbf{v}| \leq ||\nabla f(\mathbf{x})||||\mathbf{v}|| = ||\nabla f(\mathbf{x})||$$

and the equality holds if and only if $\mathbf{v} = 1/||\nabla f(\mathbf{x})||\nabla f(\mathbf{x})$. ■

In words, Proposition 2.3.3 says that the direction of the gradient vector is the direction toward which the function changes most rapidly. The next property, related, is more geometric and says that the gradient vector is perpendicular to the level curve of the function.

Proposition 2.3.4. *Let* $f : X \to \mathbb{R}$ *be a real-valued function and suppose that the gradient vector exists for all* $\mathbf{x} \in \text{int}(X)$*. For each* $r \in f(X)$*, let* $\Gamma_r := \{\mathbf{x} \in \mathbb{R} | f(\mathbf{x}) = r\}$ *be the level curve of* f *at* r*. Then*

$$\nabla f \perp \Gamma_r$$

at every $\mathbf{x} \in \text{int}(X)$ *for every* $r \in f(X)$*.*

The operations of partial derivative is the same as the operations of derivative of univairate functions. As univariate functions, we can regard each partial derivative as a function and compute the *second-order* derivative (or even higher orders), prodived that they exist. As convention, we will denote the second-order derivative of f with respect to its i-th and j-th component as

$$\frac{\partial^2 f}{\partial x_i \partial x_j}$$

if $i \neq j$ or

$$\frac{\partial^2 f}{\partial^2 x_i^2}$$

if $i = j$.

The following theorem allows to simplify the calculation of all the partial derivatives.

Theorem 2.3.2 (Young's Theorem)**.** *Let* $f : X \to \mathbb{R}$ *be a real-valued function and it's second-order derivative exist and is continuous for any pair* $i, j \in \{1, \ldots, n\}$ *on an open set* $U \subseteq X$*. Then for all* $i, j \in \{1, \ldots, n\}$*, for all* $\mathbf{y} \in U$

$$\left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\mathbf{y}} = \left.\frac{\partial^2 f}{\partial x_j \partial x_i}\right|_{\mathbf{x}=\mathbf{y}}.$$

In most of the economic problems that second order derivatives are involved, this condition will hold. Thus, hereafter, if not otherwise stated, we will assume that second order derivatives are interchangeable. With Young's theorem, we may collect all the cross second-order derivatives of a function and this will form a symmetric matrix. We call this *Hessian matrix*. Formally,

$$H_f(\mathbf{x}^*) := \begin{pmatrix} \left.\frac{\partial^2 f}{\partial^2 x_1}\right|_{\mathbf{x}=\mathbf{x}^*} & \cdots & \left.\frac{\partial^2 f}{\partial x_1 \partial x_n}\right|_{\mathbf{x}=\mathbf{x}^*} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 f}{\partial x_n \partial x_1}\right|_{\mathbf{x}=\mathbf{x}^*} & \cdots & \left.\frac{\partial^2 f}{\partial^2 x_n}\right|_{\mathbf{x}=\mathbf{x}^*} \end{pmatrix}$$

2.4 Application: Mechanism Design—Monopolistic Screening

While many economic theories study the society and human interactions by setting up models and examining the behaviors of individuals and outcomes in equilibrium. Mechanism design is a class of problems that takes one step forward—with the understanding of how individuals will response under a given environment via equilibrium analyses, mechanism design further considers how to set up proper environments to induced desired outcomes. In some classical mechanism design problems, the mathematical tools introduced above turn out to be very useful. We will introduce one of the simplest mechanism design problem to illustrate the usage of these tools. The discussions below is a simplified version of Myerson (1981).

Consider a monopolist who wishes to sell an indivisible good to a consumer. Being a monopolist, we assume that she can set up any selling scheme in order to maximize revenue (with the production cost being normalized to zero or assumed to be sunk). However, the monopolist does not have complete information about the environment. Specifically, suppose that the consumer has a quasi-linear preference, given the probability of trade p, and the amount of money the consumer has to pay, t, the consumer's utility is:

$$pv - t,$$

where $v \in [0, 1]$ denotes the consumer's *valuation* of the good. We assume that the consumer knows exactly the value of v but the monopolist does not. Instead, she only knows that the value is distributed according to a CDF F that has full support on $[0, 1]$ and a density f. Since the monopolist can set up *any* selling scheme, she can propose any *mechanism*,

which constitutes of a nonempty set M, a set of strategy for the consumer to use and two functions $p : M \to [0,1]$ and $t : M \to \mathbb{R}$, that specify the probability of trade and the amount of payment based on the strategy that the consumer chooses. We also assume that the consumer can always choose to not participate in the mechanism, in this case, he gets a utility from his outside option, normalized to 0. As such, given any mechanism (M, p, t), the consumer solves the problem:

$$\max_{m \in M} (p(m)v - t(m), 0)^+.$$

If the problem has a solution, we let $\sigma^*(v) \in \text{argmax}_{m \in M}(p(m)v - t(m))^+$ be an optimal choice of the consumer that the monopolist prefers the most. The monopolist's goal is to choose a mechanism (M, p, t) to maximize

$$\mathbb{E}_F[t(\sigma^*(v))].$$

Observe that, since the monopolist can choose *any* selling scheme, she can always set the strategy space as $M = [0,1]$ and ask the consumer to *report* his value. Therefore, suppose that $(M, \tilde{p}, \tilde{t})$ is a mechanism that maximizes revenue, with consumer's optimal choice being σ^*, $([0,1], p, t)$ is still a mechanism that attains the same revenue, where $p := \tilde{p} \circ \sigma^*$ and $t := \tilde{t} \circ \sigma^*$. Furthermore, under this mechanism, reporting the true value is always an optimal choice of the consumer. Indeed, for any $v, v' \in [0,1]$

$$\begin{aligned} p(v')v - t(v') &= \tilde{p}(\sigma^*(v'))v - \tilde{t}(\sigma^*(v')) \\ &\leq \tilde{p}(\sigma^*(v))v - \tilde{t}(\sigma^*(v)) \\ &= p(v)v - t(v), \end{aligned}$$

where the inequality follows from the fact the σ^* is optimal. Consequently, it is without loss for the monopolist to restrict attention on mechanisms of form $([0,1], p, t)$ under which reporting true value is an optimal choice for the consumer, where $p : [0,1] \to [0,1]$, $t : [0,1] \to \mathbb{R}$. This is called the *incentive compatible direct mechanisms*. Together, the monopolist's

problem can be simplified to

$$\max_{(p,t):[0,1]^2\to[0,1]\times\mathbb{R}} \int_0^1 t(v)f(v)dv$$

$$\text{s.t. } p(v)v - t(v) \geq p(v')v - t(v'),\ \forall v, v' \in [0,1] \tag{IC}$$

$$p(v)v - t(v) \geq 0,\ \forall v \in [0,1]. \tag{IR}$$

This is a well-defined mathematical problem. This simplification procedure is called the *revelation principle*.

We can now begin to solve the monopolist's problem. We will first examine the condition (IC). It turns out that, due to quasi-linearity, the condition (IC) alone reduces the problem to just with a choice variable p, up to a constant, which is stated in the following Proposition.

Proposition 2.4.1. *A direct mechanism (p,t) satisfies* (IC) *if and only if:*

1. $t(v) = t(0) + p(v)v - \int_0^v p(x)dx$.
2. p *is increasing.*

Proof. For sufficiency, suppose that (p,t) is a mechanism that satisfies conditions 1 and 2. Then for any $0 \leq v < v' \leq 1$,

$$\begin{aligned}
p(v')v - t(v') &= p(v')v - t(0) - p(v')v' + \int_0^{v'} p(x)dx \\
&= -p(v')(v'-v) + \int_v^{v'} p(x)dx + \int_0^v p(x)dx - t(0) \\
&= \int_v^{v'} (p(x) - p(v'))dx + \left(\int_0^v p(x)dx - t(0)\right) \\
&= \int_v^{v'} (p(x) - p(v'))dx + (p(v)v - t(v)) \\
&\leq p(v)v - t(v),
\end{aligned}$$

where the inequality follows from monotonicity of p.

For necessity, suppose that (p,t) satisfies (IC), let $\Pi(v) : p(v)v - t(v)$. Then for any $0 \leq v < v' \leq 1$,

$$\begin{aligned}
\Pi(v) = p(v)v - t(v) &\geq p(v')v - t(v') \\
&= p(v')v' - t(v') - p(v')(v'-v) \\
&= \Pi(v') - p(v')(v'-v)
\end{aligned}$$

and

$$\begin{aligned}\Pi(v') = p(v')v' - t(v') &\geq p(v)v' - t(v) \\ &= p(v)v - t(v) + p(v)(v' - v) \\ &= \Pi(v) + p(v)(v' - v).\end{aligned}$$

Together,

$$p(v) \leq \frac{\Pi(v') - \Pi(v))}{v' - v} \leq p(v'), \tag{5}$$

from which monotonicity of p immediately follows. Moreover, since $p(v) \in [0, 1]$ for all $v \in [0, 1]$, (5) implies that

$$|\Pi(v') - \Pi(v)| \leq |v' - v|, \forall v, v' \in [0, 1].$$

That is, Π is Lipschitz continuous and thus is absolutely continuous, by the Fundamental Theorem of Calculus, Π is differentiable almost everywhere on $[0, 1]$ and its derivative at $v \in [0, 1]$, whenever exists, is

$$\lim_{v' \downarrow v} \frac{\Pi(v') - \Pi(v)}{v' - v} = \lim_{v' \uparrow v} \frac{\Pi(v) - \Pi(v')}{v - v'},$$

which, by (5), is exactly $p(v)$. Again, by the Fundamental Theorem of Calculus,

$$p(v)v - t(v) = \Pi(v) = \Pi(0) + \int_0^v p(x)dx.$$

Finally, use $\Pi(0) = -t(0)$, we have

$$t(v) = t(0) + p(v)v - \int_0^v p(x)dx,$$

as desired. ■

With Proposition 2.4.1, the monopolist's problem can be further simplified as

$$\max_{p:[0,1]\to[0,1]} \int_0^1 \left(p(v)v - \int_0^v p(x)dx \right) f(v)dv + t(0). \text{ s.t. } p \text{ is increasing, } t(0) \leq 0$$

By (IR), and by the observation that Π is increasing, it is optimal to set $t(0) = 0$. Morevover,

by Fubini's theorem,

$$\begin{aligned}\int_0^1 \left(\int_0^v p(x)dx\right) f(v)dv &= \int_0^1 \int_0^v p(x)f(v)dxdv \\ &= \int_0^1 \left(\int_x^1 f(v)\right) p(x)dx \\ &= \int_0^1 p(x)(1-F(x))dx \\ &= \int_0^1 p(v)\left(\frac{1-F(v)}{f(v)}\right) f(v)dv\end{aligned}$$

Together, the problem becomes:

$$\max_{p:[0,1]\to[0,1]} \int_0^1 p(v)\left(v - \frac{1-F(v)}{f(v)}\right) f(v)dv \text{ s.t. } p \text{ is increasing,} \tag{6}$$

which is now an optimization problem with linear (and hence continuous) objective. Furthermore, as observed in Remark 2.2.3, the collection of uniformly bounded increasing functions on $[0,1]$ is compact under the $L^1([0,1])$ norm. Therefore, the solution to (6) must exist. We will return to a sharper characterization of the solutions in next chapter.

2.5 Exercises

1. Prove Proposition 2.1.3.

2. Prove Proposition 2.1.4.

3. Recall that a sequence of real-valued functions $\{f_n\}$ is said to converge pointwisely to some function $f : [a,b] \to \mathbb{R}$ if for any $x \in [a,b]$, $\{f_n(x)\} \to f$. We say that $\{f_n\}$ converges to f uniformly if for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that whenever $n > N$, $|f_n(x) - f(x)| < \varepsilon$ for all $x \in [a,b]$. Give an example to show that pointwise convergent does not imply uniform convergent.

4. (Chebychev's Inequality). Let $f \geq 0$ be a measurable function on $X \in \mathcal{M}$, $X \subset \mathbb{R}$. Show that for any $c > 0$,

$$\lambda(\{x \in X | f(x) \geq c\}) \leq \frac{1}{c}\int_X f d\lambda.$$

5. (Convergence in Measure). Let $\{f_n\}$, f be measurable functions on some measurable set $X \subset \mathbb{R}$. Suppose that f_n and f are finite almost everywhere on X. We say that $\{f_n\}$ *converges in measure* to f if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \lambda(\{x \in X || f_n(x) - f(x)| > \epsilon\}) = 0.$$

Show that $\{f_n\} \to f$ under the L^p norm, $p \in [1, \infty)$ implies $\{f_n\}$ converges to f in measure.

6. Show that an increasing function $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ except at at most countably many points.

7. Consider again the monopolistic screening problem in section 2.4. Another version of this model is that the monopolist can produce a good with different quality $x \in \mathbb{R}_+$, with a production cost c. Assume that the buyer's utility from buying a good with quality x is $u(x, v)$. Thus, payoff is given by

$$u(x, v) - t.$$

Assume also that $u \in C^2(\mathbb{R}^2_{++})$ and that $u_1, u_2, u_{12} > 0$. Also, assume that there exists $K \in \mathbb{N}$ such that $u_2 < K$.

Suppose that the seller has full bargaining power and thus can choose any selling mechanism. By the same argument as the revelation principle in section 2.4, it is without loss of generality to restrict attention on incentive compatible and individually rational direct mechanisms. That is, (x, t) such that $x : [0, 1] \to \mathbb{R}_+$, $t : [0, 1] \to \mathbb{R}$,

$$u(x(v), v) - t(v) \geq u(x(v'), v) - t(v'),\ \forall v, v' \in [0, 1] \tag{IC}$$

$$u(x(v), v) - t(v) \geq 0,\ \forall v \in [0, 1]. \tag{IR}$$

Show that (x, t) is incentive compatible (IC holds) if and only if:

(a) $t(v) = t(0) + u(x(v), v) - \int_0^v u_2(x(z), z)dz$.

(b) x is increasing.

Use this to conclude that the seller's expected revenue under any incentive compatible and individual rational direct mechanism is not better than

$$\sup_{x:[0,1]\to\mathbb{R}} \int_0^1 \left(u(x(v), v) - u_1(x(v), v)\frac{1 - F(v)}{f(v)} - c(x(v)) \right) f(v)dv.$$

$$\text{s.t. } x \text{ is increasing}$$

3 Optimization and Comparative Statics

In this chapter we will introduce some widely used method for solving optimization problems in economics and some mathematical tools for studying comparative static problems. Recall that from Proposition 1.2.5, we know that for any upper-semicontinuous function defined on a compact set, it must have a solution. However, we do not have many mathematical tools that can help us find what exactly the solutions are. Using Calculus, we can examine the local behaviors of a function and approximate a function linearly. This allows us to find maxima and minima *locally*. If, furthermore, the objective has some sort of concavity/convexity, such local maxima/minima are also global. This is often called the *first order approach*. On the other hand, when the function is not always differentiable but is concave/convex, some results in *convex* analysis can also help us find some maxima/minima. We will first introduce the first-order approach and some concepts in convex analysis. Finally, we will look at some methods that enable us to do comparative static analyses.

3.1 First-order approach

3.1.1 Motivation: First-Order Approach with Univariate Functions

We start with first order approach with functions on $\mathbb{R}$. Let $[a, b] \subset \mathbb{R}$ be an closed interval. Consider a continuous, twice differentiable function $f : \mathbb{R} \to \mathbb{R}$. Our goal is to characterize the solution of

$$\max_{x\in[a,b]} f(x).$$

We say that $x^* \in [a, b]$ is a *local maximum* of f on $[a, b]$ if there exists a neighborhood of x^*, $O \subset [a, b]$ such that $f(x^*) \geq f(x)$ for all $x \in O$. Observe that if $x^* \in (a, b)$ and $f'(x^*) > 0$

or $f'(x^*) < 0$, then x^* cannot be a local maximum. Indeed, suppose that $f(x^*) > 0$. Then for any $\delta > 0$, as

$$0 < f'(x^*) = \lim_{x' \downarrow x^*} \frac{f(x') - f(x^*)}{x' - x^*},$$

there exists $x' \in (x^*, x^* + \delta)$ such that $f(x') > f(x^*)$. Analogous argument holds when $f'(x^*) < 0$. As such, we know that for any $x^* \in (a, b)$, a necessary for x^* to be a local maximum, and hence a global maximum, is that $f'(x^*) = 0$. On the other hand, if $x^* = a$, since elements in $\mathbb{R} \backslash [a, b]$ are not allowed to be selected, it is possible for x^* to be a local maximum even if $f'(a) < 0$. Similarly, it is possible for $x^* = b$ to be a maximum even when $f'(b) > 0$. We can summarize the observations above by the following Proposition.

Proposition 3.1.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Then $x^* \in [a, b]$ is a solution of*

$$\max_{x \in [a,b]} f(x)$$

only if

$$f'(x^*)(x^* - a) = 0 \text{ and } f'(x^*) \leq 0;$$
$$\text{or } f(x^*)(b - x^*) = 0 \text{ and } f'(x^*) \geq 0.$$

The necessary condition above is also called the *first-order Khun-Tucker* conditions. It is essentially saying that any local maximum must has zero derivative. The complication arrives when considering boundaries of the choice set, in such cases we allow the derivative to have negative (positive) signs. The multiplication terms ensure that non zero derivatives only occurs at boundaries and the inequalities ensure that the direction is correct. This is sometimes called the *complementary slackness* condition.

However, this condition is not sufficient. Clearly, if $f(x) = (x - 1/2)^2$ for all $x \in [0, 1]$, then $f'(1/2) = 0$ but we know that f is *minimized* at $1/2$. To ensure (local) maximization, we need to examine the local behaviors of f around x^* even when $f'(x^*) = 0$. Suppose that $f'(x^*) = 0$ at some $x^* \in (a, b)$ and that $f'(x) > 0$ for all $x^* - \delta < x < x^*$ and $f'(x) < 0$ for all $x^* + \delta > x > x^*$ for some $\delta > 0$. Then we know that $f(x^*) > f(x)$ for all $x \in (x^* - \delta, x^* + \delta)$. As f is twice differentiable, this is equivalent to $f''(x^*) < 0$. As such, the sufficiency of first order condition is ensured.

Proposition 3.1.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a twice differentiable function with continuous second-order derivative and suppose that $f'(x^*) = 0$ for some $x^* \in [a, b]$. Then x^* is a local maximum of f on $[a, b]$ if $f''(x^*) < 0$. Furthermore, if $f'(a) > 0$ ($f'(b) > 0$), then a (b) is local maximum*

It is possible that local maximum is not global maximum. However, if $f'' < 0$ on $[a, b]$, or equivalently, if f is strictly concave, then there is a unique local maximum. This condition is called the *second-order condition*.

3.1.2 Unconstrained First-Order Kuhn-Tucker Condition

In this section, we will formally derive the first-order Kuhn-Tucker condition for unconstrained problems with multivairate functions. With the results in this section, the observations in the previous section is then a special case. We start with the most essential theorem.

Theorem 3.1.1 (Mean Value theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function such that $\nabla f(\mathbf{x})$ exists for all $\mathbf{x} \in \mathbb{R}^n$. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{z})^\intercal (\mathbf{x} - \mathbf{y}),$$

for some $\mathbf{z} = t\mathbf{x} + (1 - t)\mathbf{y}$ with $t \in [0, 1]$.

Mean value theorem allows us to approximate the function f *linearly* by using an *affine* function. By the same spirit, we can also approximate f by using a *quadratic* function if f is twice differentiable.

Theorem 3.1.2 (Generalized Mean Value Theorem)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-value function such that $H_f(\mathbf{x})$ exists for all $\mathbf{x} \in \mathbb{R}^n$. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$f(\mathbf{x}) = f(\mathbf{y}) + \nabla f(\mathbf{y})^\intercal (\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^\intercal H_f(\mathbf{z})(\mathbf{x} - \mathbf{y}),$$

for some $\mathbf{z} = t\mathbf{x} + (1 - t)\mathbf{y}$ with $t \in [0, 1]$.

From the mean value theorem, it is then immediate that for $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}^*$ is a maximum only if:

$$\nabla f(\mathbf{x}^*) = 0.$$

Since if not, since ∇f is continuous, there exists an open ball $O := N_\delta(\mathbf{x}^*)$ such that $\nabla f(\mathbf{x}) \neq 0$ for any $\mathbf{x} \in O$. We may then take a direction $\mathbf{v}$ such that $\nabla f(\mathbf{x})^\intercal \mathbf{v} > 0$. Thus, for $\mathbf{y} := \mathbf{x}^* + \lambda \mathbf{v}$, for $\lambda > 0$ small enough so that $\mathbf{y} \in O$,

$$f(\mathbf{x}^*) - f(\mathbf{y}) = \nabla \lambda f(\mathbf{z})^\intercal \mathbf{v} > 0,$$

since $t\mathbf{x} + (1-t)\mathbf{y} \in O$ for all $t \in [0,1]$.

Similar to the case with univariate functions, when the choice set is restricted to a subset of $\mathbb{R}^n$, we need to take first order conditions of the boundaries into account. The next theorem summarizes the necessary condition of maximum of a multivariate function when the choice set is a cube. For other bounded sets, it is more convenient to use the constraint optimization methods that will be introduced in the next section.

Proposition 3.1.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function such that $\nabla f(\mathbf{x})$ exists for all $\mathbf{x} \in \mathbb{R}^n$. Let $X := \prod_{k=1}^n [a_k, b_k]$ and $\mathbf{a} := (a_k)_{k=1}^n$, $\mathbf{b} := (b_k)_{k=1}^n$. Then $\mathbf{x}^* \in X$ is a solution of*

$$\max_{\mathbf{x} \in X} f(\mathbf{x})$$

only if

$$\nabla f(\mathbf{x}^*) \circ (\mathbf{x} - \mathbf{a}) = 0, \text{ and } \nabla f(\mathbf{x}^*) \leq 0$$
$$\text{or } \nabla f(\mathbf{x}^*) \circ (\mathbf{b} - \mathbf{x}) = 0, \text{ and } \nabla f(\mathbf{x}^*) \leq 0,$$

where $\circ$ is the component-wise product and the order is the component-wise order.

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, we say that $\mathbf{x}^*$ is a *local maximum* of f on $X := \prod_{k=1}^n [a_k, b_k]$ if there exists a neighborhood of $\mathbf{x}^*$, $O \subset X$ that is open in X, such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all $\mathbf{x} \in O$.

As in univariate case, the second-order derivatives help us to ensure sufficiency of the first order condition. We can derive this from the generalized mean valued theorem. Indeed, let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that $H_f(\mathbf{x})$ exists for all $\mathbf{x} \in \mathbb{R}^n$, if $\nabla f(\mathbf{x}^*) = 0$, then, by Theorem 3.1.2, for any $\mathbf{y} \in \mathbb{R}$,

$$f(\mathbf{x}^*) - f(\mathbf{y}) = -\frac{1}{2}(\mathbf{x}^* - \mathbf{y})^\intercal H_f(\mathbf{z})(\mathbf{x}^* - \mathbf{y}),$$

for some $\mathbf{z} = t\mathbf{x}^* + (1-t)\mathbf{y}$ with $t \in [0,1]$. Therefore, if the Hessian matrix is such that

$$\frac{1}{2}(\mathbf{x}^* - \mathbf{y})^\intercal H_f(\mathbf{z})(\mathbf{x}^* - \mathbf{y}) \leq 0,$$

we then know that $\mathbf{x}^*$ is a maximum.

To this end, we need to examine the structure of the Hessian matrix. Below we will introduce the relevant notion

Definition 3.1.1. Let $A = (a_{ij})_{i,j=1}^n$ be a symmetric $n \times n$ real matrix. We say that A is *negative (positive) definite* if for any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq 0$,

$$\mathbf{x}^\intercal A\mathbf{x} < (>)0.$$

We say that A is *negative (positive) semi-definite* if the inequality is not strict.

A convenient characterization of negative definite matrix is given below.

Proposition 3.1.4. *A symmetric $n \times n$ real matrix $A = (a_{ij})_{i,j=1}^n$ is negative definite if and only if*

$$(-1)^k \det(A_k) > 0, \, \forall k \in \{1, \ldots, n\}$$

where

$$A_k := (a_{ij})_{i,j=1}^n$$

is called the $k-th$ leading principal minor of A. Furthermore, A is positive definite if and only if

$$\det(A_k) > 0, \, \forall k \in \{1, \ldots, n\}.$$

Just as when f is univariate, the Hessian matrix is closely related to concavity of f.

Proposition 3.1.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function such that $H_f(\mathbf{x})$ exists for all $\mathbf{x} \in \mathbb{R}^n$. f is strictly concave (convex) if and only if $H_f(\mathbf{x})$ is negative definite for all $\mathbf{x} \in \mathbb{R}^n$*

Remark 3.1.1. By Proposition 3.1.5, there might be a function f of which $\partial^2 f / \partial^2 x_i < 0$ for all $i \in \{1, \ldots, n\}$ but f is not concave. That is, second-order derivative of each component is not enough for characterizing convexity.

With the notion of negative definite matrices, we can now give a sufficient condition for maxima.

Proposition 3.1.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function such that $H_f(\mathbf{x})$ exists for all $\mathbf{x} \in \mathbb{R}^n$. Suppose that $\nabla f(\mathbf{x}^*) = 0$. Then $\mathbf{x}^*$ is a local maximum of f on $X := \prod_{k=1}^n [a_k, b_k]$ if $H_f(\mathbf{x}^*)$ is negative definite. Furthermore, if f is concave, then x^* is a global maximum.*

Remark 3.1.2. Unlike the case of univairate function, there might be cases under which $\nabla f(\mathbf{x}^*) = 0$ but $H_f(\mathbf{x}^*)$ is neither positive nor negative definite. In this case, we will say the $\mathbf{x}^*$ is a *saddle point* of f and the first order condition is not sufficient for maxima.

Example 3.1.1 (Profit Maximization Problem). Consider a firm who has to decide its production plan and maximize profit. Suppose that the the firm has a technology that allows it to produce outputs by labor and capital. Formally, let $F : \mathbb{R}^2_+ \to \mathbb{R}_+$ be the firm's *production function* that is twice differentiable. That is, given the amount of labor, $L \geq 0$ and capital $K \geq 0$, the firm can produce $F(L, K)$ units. To rule out corner solution, we assume that $\lim_{L\to 0} F_L(L, K) = \lim_{K\to 0} F_K(L, K) = \infty$ for all $L, K \in \mathbb{R}_+$. Suppose that the firm is a *price taker* and therefore must take market prices as given when making decisions. Let $p > 0$ denote the price of output, $w > 0$ denote the wage for a unit of labor and $r > 0$ denote the rental rate for a unit of capital. The firm's problem is then given by:

$$\max_{(L,K)\in\mathbb{R}^2_+} pF(L, K) - wL - rK$$

Using the first order condition, we know that L^* and K^* is a solution to this problem only if

$$F_L(L^*, K^*) = \frac{w}{p}, \text{ and } F_K(L^*, K^*) = \frac{r}{p}.$$

This is a familiar economic formula which states that marginal productivity of labor (capital) must be the real wage (rental rate) at optimum. Furthermore, the Hessian matrix of F is given by

$$H_F(L, K) = \begin{pmatrix} F_{LL}(L, K) & F_{LK}(L, K) \\ F_{KL}(L, K) & F_{KK}(L, K) \end{pmatrix}.$$

If we further assume that the technology does not have strong substuitability/complementarity so that

$$\det(H_F(L, K)) = F_{LL}(L, K)F_{KK}(L, K) - F_{KL}(L, K)^2 > 0,$$

for all (L, K) and that the technology is exhibiting diminishing marginal productivity of labor and capital so that $F_{LL} < 0$ and $F_{KK} < 0$ by Proposition 3.1.5, F is strictly concave and thus first order condition is sufficient. Consequently, the solution to the firm's problem can be characterize by the system:

$$F_L(L^*, K^*) = \frac{w}{p}, \text{ and } F_K(L^*, K^*) = \frac{r}{p}.$$

We notice that in Example 3.1.1, we need to assume that the determinant of the Hessian matrix of F is positive, since otherwise the second order condition will not hold. In fact, there is a particular class of functions that is common in economic but fail to satisfy this condition. We will introduce this class of functions briefly here.

Definition 3.1.2. Let $X \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}$ such that for any $\mathbf{x} \in X$ and any $r \in (0, \infty)$, $r\mathbf{x} \in X$. Let $f : X \to \mathbb{R}$ be a real-valued function. f is said to be *homogeneous of degree* $k \in \mathbb{N}$ if for any $r \in (0, \infty)$,

$$f(r\mathbf{x}) = r^k f(\mathbf{x}).$$

Two properties of homogeneous functions can be derived by examining the definition.

Proposition 3.1.7. *Let $X \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}$ such that for any $\mathbf{x} \in X$ and any $r \in (0, \infty)$, $r\mathbf{x} \in X$ and that $\operatorname{int}(X) \neq \emptyset$ and let $f : X \to \mathbb{R}$ be a function that is homogeneous of degree $k \in \mathbb{N}$. Suppose that $\nabla f(\mathbf{x})$ exists for all $\mathbf{x} \in \operatorname{int}(X)$. Then f_i is homogeneous of degree $k - 1$ on $\operatorname{int}(X)$.*

Proposition 3.1.8 (Euler's Theorem)**.** *Let $X \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}$ such that for any $\mathbf{x} \in X$ and any $r \in (0, \infty)$, $r\mathbf{x} \in X$ and that $\operatorname{int}(X) \neq \emptyset$ and let $f : X \to \mathbb{R}$ be a function that is homogeneous of degree $k \in \mathbb{N}$. Suppose that $\nabla f(\mathbf{x})$ exists for all $\mathbf{x} \in \operatorname{int}(X)$. Then*

$$\nabla f(\mathbf{x})^\intercal \mathbf{x} = k f(\mathbf{x}).$$

It is commonly seen in economic models that the production function of a firm is *constant return to scale*—when scaling up inputs outputs will be scaled up by the same rate, which is exactly the functions that are homogeneous of degree 1. Suppose that $f : X \to \mathbb{R}$ is

homogeneous of degree 1 and $H_f(\mathbf{x})$ exists and each component is continuous. Applying Euler's theorem, we know that for all $\mathbf{x} \in \text{int}(X)$,

$$\mathbf{x}^\intercal \nabla f(\mathbf{x}) = f(\mathbf{x}).$$

Taking gradient vectors on both sides, we have:

$$\nabla f(\mathbf{x}) + H_f(\mathbf{x})\mathbf{x} = \nabla f(\mathbf{x})$$

and thus for all $\mathbf{x} \in \text{int}(X)$,

$$H_f(\mathbf{x})\mathbf{x} = 0,$$

which means that for any $\mathbf{x} \in \text{int}(X)$, the null space of $H_f(\mathbf{x})$ is not $\{0\}$ and therefore,

$$\det(H_f(\mathbf{x})) = 0, \forall \mathbf{x} \in \mathbb{R}^n.$$

As a result, not only the second order condition does not hold, since $\det(H_f(\mathbf{x})) = 0$ for all $\mathbf{x} \in \text{int}(X)$, the system

$$\nabla f(\mathbf{x}) = \mathbf{w}$$

has infinitely many solutions for any $\mathbf{w} \in \mathbb{R}^n$. For example, as in Example 3.1.1, if $F(L, K) = L^\alpha K^{1-\alpha}$. The solution to the first order condition is in fact a ray:

$$\left\{ (L, K) \in \mathbb{R}^+ \middle| \frac{L}{K} = \frac{\alpha w}{(1-\alpha) r} \right\}.$$

3.1.3 Constrained Maximization and the Lagrange Method

In many economic problems, the feasible set for a maximization problem is often restricted in a more complicated way. For instance, as in Example 1.2.2, the consumer can only choose the bundle that he/she is affordable to maximize utility. The set of feasible bundle is given by an inequality:

$$\Gamma(\mathbf{p}, m) := \{\mathbf{x} \in \mathbb{R}^n_+ | \mathbf{p}^\intercal \mathbf{x} \leq m\}.$$

More generally, we are sometimes interested in problems that take the following form:

$$\begin{aligned} &\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \\ &\text{s.t. } g_j(\mathbf{x}) \leq b_j, \forall j \in \{1, \ldots, m\} \\ &\qquad h_l(\mathbf{x}) = c_l, \forall l \in \{1, \ldots, k\} \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $\{g_j\}_{j=1}^m$, $\{b_j\}_{j=1}^m$, $\{h_l\}_{l=1}^k$ and $\{c_l\}_{l=1}^k$ describe the *feasible set*.

For this class of problem, there is a useful method, called the *Lagrange method*, which allows us to relate the constrained problems to the unconstrained ones. The exact proof of this method is relatively tedious and complicated, but the intuition is easy to understand. We will state the characterization first and illustrate the intuition.

Proposition 3.1.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$, $\{g_j : \mathbb{R}^n \to \mathbb{R}\}_{j=1}^n$, $\{h_l : \mathbb{R}^n \to \mathbb{R}\}_{l=1}^k$ be functions in $C^1(\mathbb{R}^n)$ and let $\{b_j\}_{j=1}^m$, $\{c_l\}_{l=1}^k$ be real numbers. Then $\mathbf{x}^*$ is a solution of*

$$\begin{aligned} &\max_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \\ &s.t.\ g_j(\mathbf{x}) \le b_j,\ \forall j \in \{1,\dots,m\} \\ &\quad h_l(\mathbf{x}) = c_l,\ \forall l \in \{1,\dots,k\} \end{aligned} \tag{7}$$

only if there exists $\boldsymbol{\lambda} = (\lambda_j)_{j=1}^m$, $\boldsymbol{\mu} = (\mu_l)_{l=1}^k$ such that

$$\nabla \mathcal{L}(\mathbf{x}^*; \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0 \tag{FOC}$$

$$\lambda_j(g_j(\mathbf{x}^*) - b_j) = 0,\ \forall j \in \{1,\dots,m\} \tag{CS1}$$

$$h_l(\mathbf{x}^*) = c_l,\ \forall l \in \{1,\dots,k\} \tag{Ceq}$$

$$\boldsymbol{\lambda} \ge 0 \tag{CS2}$$

$$g_j(\mathbf{x}^*) \le b_j,\ \forall j \in \{1,\dots,m\}, \tag{Cineq}$$

where

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\mathbf{x}) - \sum_{j=1}^m \lambda_j(g_j(\mathbf{x}) - b_j) - \sum_{l=1}^k \mu_j(h_l(\mathbf{x}) - c_l), \forall \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}^k.$$

Proposition 3.1.9 is in fact easy to interpret (and hence memorize!). First of all, we call the function

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\mathbf{x}) - \sum_{j=1}^m \lambda_j(g_j(\mathbf{x}) - b_j) - \sum_{l=1}^k \mu_j(h_l(\mathbf{x}) - c_l), \forall \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}^k$$

The *Lagrangian*, which incorporates the *constraints* of problem (7) into its objective linearly. The constants $\{\lambda_j\}_{j=1}^m$, $\{\mu_l\}_{l=1}^k$ are called the *Lagrange multipliers*. We can think of the terms

$\sum_{j=1}^{m} \lambda_j(g_j(\mathbf{x}) - b_j)$ and $\sum_{l=1}^{k} \mu_j(h_l(\mathbf{x}) - c_l)$ as the *punishment* of violating the constraint. With this interpretation, Proposition 3.1.9 is essentially saying that the first order condition of the *unconstrained* problem:

$$\max_{\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}} \mathcal{L}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu})$$

is (almost) the same necessary condition of the constrained problem (7), which corresponds to the condition (FOC). The conditions (Ceq) and (Cineq) are simply the original feasibility constraints. The condition (CS1) and (CS2) are the *complementary slackness* conditions analogous to univairate problems, which requires that the inequality constraints must either be *binding*, (i.e. $g_j(\mathbf{x}^*) = b_j$) or the *marginal change* of moving away from the optimal must be zero (i.e. $\lambda_j = 0$) and, if the constraint is binding, the marginal change of moving away from $\mathbf{x}^*$ must has the correct direction (i.e. $\lambda_j \geq 0$).

We now consider a special case to illustrate the intuition of Proposition 3.1.9. Specifically, consider a problem :

$$\max_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \text{ s.t. } h(\mathbf{x}) = c.$$

The constraint $h(\mathbf{x}) = c$ requires the feasible set to be a *manifold*, or equivalently, on the c level curve of the function h. We denote this manifold as Γ. Now suppose that at some point $\mathbf{x}^*$ on this manifold, $\nabla f(\mathbf{x}^*) \nparallel \nabla h(\mathbf{x}^*)$. Then we can take a vector $\mathbf{v}^* \in \mathbb{R}^n$ such that $\mathbf{v}^* \perp \nabla h(\mathbf{x}^*)$. By Proposition 2.3.4, we know that $\nabla h(\mathbf{x}^*)$ is perpendicular to the manifold Γ and therefore, the hyperplane $\{\mathbf{x} \in \mathbb{R}^n | \mathbf{x}^\intercal \nabla h(\mathbf{x}^*) = c\}$ is *tangent* to the manifold Γ at $\mathbf{x}^*$. As a result, points $\mathbf{z} \in \Gamma$ around $\mathbf{x}^*$ can be approximated by $\mathbf{x}^* + \lambda \mathbf{v}^*$ for some $\lambda > 0$. Now notice that since $\nabla f(\mathbf{x}^*) \nparallel \nabla h(\mathbf{x}^*)$, $\nabla f(\mathbf{x}^*)^\intercal v^* \neq 0$. Without loss, we assume that $\nabla f(\mathbf{x}^*)^\intercal \mathbf{v}^* > 0$. Since the directional derivative of f at $\mathbf{x}^*$ toward $\mathbf{v}^*$ is exactly $\nabla f(\mathbf{x}^*)^\intercal \mathbf{v}^*$, we may conclude that there exists $\bar{\lambda} > 0$ such that $f(\mathbf{x}^* + \lambda \mathbf{v}^*) > f(\mathbf{x}^*)$ for all $\lambda \in (0, \bar{\lambda})$. Together, as there exists some $\mathbf{z} \in \Gamma$ that is close enough to $\mathbf{x}^*$ such that $f(\mathbf{z})$ is close to $f(\mathbf{x}^* + \lambda \mathbf{v}^*)$ for some λ and $\lambda < \bar{\lambda}$, we can conclude that $f(\mathbf{z}) > f(\mathbf{x}^*)$ for some $\mathbf{z} \in \Gamma$. As a result, we know that $\mathbf{x}^*$ is a maximum only if $\nabla f(\mathbf{x}^*) \parallel h(\mathbf{x}^*)$, or equivalently, the level curve of f must be tangent to the manifold Γ at the solution. This means that there exists $\lambda \in \mathbb{R}, \lambda \neq 0$ such that

$$\nabla f(\mathbf{x}^*) = \lambda \nabla h(\mathbf{x}^*),$$

which is exactly the condition

$$\nabla \mathcal{L}(\mathbf{x}; \lambda) = 0.$$

Similar to unconstrained problems, first order conditions given by Proposition 3.1.9 is not always sufficient. To ensure sufficiency, we need to examine the second-order conditions. The following Proposition summarizes such condition.

Proposition 3.1.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$, $\{g_j : \mathbb{R}^n \to \mathbb{R}\}_{j=1}^n$, $\{h_l : \mathbb{R}^n \to \mathbb{R}\}_{l=1}^k$ be real-valued functions such that the gradient vector exists at all $\mathbf{x} \in \mathbb{R}^n$ and let $\{b_j\}_{j=1}^m$, $\{c_l\}_{l=1}^k$ be real numbers. Then $\mathbf{x}^*$ is a local maximum of the problem*

$$\begin{aligned} &\max_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \\ &s.t.\ g_j(\mathbf{x}) \leq b_j,\ \forall j \in \{1, \ldots, m\} \\ &\quad h_l(\mathbf{x}) = c_l,\ \forall l \in \{1, \ldots, k\} \end{aligned}$$

if there exists $\boldsymbol{\lambda} = (\lambda_j)_{j=1}^m$, $\boldsymbol{\mu} = (\mu_l)_{l=1}^k$ such that

$$\begin{aligned} &\nabla \mathcal{L}(\mathbf{x}^*; \boldsymbol{\lambda}, \boldsymbol{\mu}) = 0 \\ &\lambda_j(g_j(\mathbf{x}^*) - b_j) = 0,\ \forall j \in \{1, \ldots, m\} \\ &h_l(\mathbf{x}^*) = c_l,\ \forall l \in \{1, \ldots, k\} \\ &\boldsymbol{\lambda} \geq 0 \\ &g_j(\mathbf{x}^*) \leq b_j,\ \forall j \in \{1, \ldots, m\}, \end{aligned}$$

where

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\mathbf{x}) - \sum_{j=1}^m \lambda_j(g_j(\mathbf{x}) - b_j) - \sum_{l=1}^k \mu_j(h_l(\mathbf{x}) - c_l), \forall \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}^k$$

and for $J := \{j \in \{1, \ldots, m\} | g_j(\mathbf{x}^*) = b_j\}$,

$$H_{\mathcal{L}_J}(\mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

is negative positive definite on $N(J_\mathbf{g}) \cup N(J_\mathbf{h})$. That is, for any $\mathbf{v} \in \mathbb{R}^n$ such that $J_\mathbf{g}(\mathbf{x}^*)\mathbf{v} = 0$, or $J_\mathbf{h}(\mathbf{x}^*)\mathbf{v} = 0$,

$$\mathbf{v}^\intercal H_{\mathcal{L}_J}(\mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu})\mathbf{v} < 0,$$

where

$$\mathcal{L}_J(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\mathbf{x}) - \sum_{j \in J} \lambda_j (g_j(\mathbf{x}) - b_j) - \sum_{l=1}^{k} \mu_j (h_l(\mathbf{x}) - c_l), \forall \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^{|J|}, \boldsymbol{\mu} \in \mathbb{R}^k.$$

Moreover, this condition holds if for any $2(|J| + k) + 1 \leq j \leq n + |J| + k$,

$$(-1)^j det(H_j) < 0,$$

where H_j *is the* j*-th leading principle of* $H_{\mathcal{L}}(\mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu})$.

That is, the Hessian matrix of the Lagrangian that incorporates only the bindining constraints is the relevant second order condition. Explicitly, let $\mathbf{g} := (g_j)_{j \in J}$ and $\mathbf{h} := (h_j)_{j=1}^k$ and let

$$J_{\mathbf{g}}(\mathbf{x}^*) := \left(\frac{\partial g_j}{\partial x_i} \bigg|_{\mathbf{x}=\mathbf{x}^*} \right)_{i \in \{1,\ldots,n\}, j \in J}; \quad J_{\mathbf{h}}(\mathbf{x}^*) := \left(\frac{\partial h_k}{\partial x_i} \bigg|_{\mathbf{x}=\mathbf{x}^*} \right)_{i \in \{1,\ldots,n\}, l \in \{1,\ldots,m\}}$$

denote the *Jacobian matrix* of $\mathbf{g}$ and $\mathbf{h}$ at $\mathbf{x}^*$, respectively, and let

$$H_{\mathcal{L}|_{\mathbb{R}^n}}(\mathbf{x}^*; \boldsymbol{\lambda}, \boldsymbol{\mu}) := \left(\frac{\partial^2 \mathcal{L}(\cdot, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial x_i \partial x_j} \bigg|_{\mathbf{x}=\mathbf{x}^*} \right)_{i,j \in \{1,\ldots,n\}}$$

denote the Hessian matrix of $\mathcal{L}(\cdot, \boldsymbol{\lambda}, \boldsymbol{\mu})$ holding $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ fixed. This Hessian matrix can then be represented as:

$$H_{\mathcal{L}}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{0}_{|J| \times n} & \mathbf{0}_{k \times n} & -J_{\mathbf{g}}(\mathbf{x}^*) \\ \mathbf{0}_{k \times n} & \mathbf{0}_{|J| \times n} & -J_{\mathbf{h}}(\mathbf{x}^*) \\ -J_{\mathbf{g}}(\mathbf{x}^*)^\intercal & -J_{\mathbf{h}}(\mathbf{x}^*)^\intercal & H_{\mathcal{L}|_{\mathbb{R}^n}}(\mathbf{x}^*; \boldsymbol{\lambda}, \boldsymbol{\mu}). \end{pmatrix}$$

This is often called the *bordered Hessian matrix*.

Example 3.1.2 (Consumer's Problem Revisited). We now have a tool to solve explicitly for the consumer's problem introduced in Example 1.2.2. Suppose that a consumer has a twice differentiable utility function $u : \mathbb{R}^n \to \mathbb{R}$ that is strictly increasing in each argument and quasi-concave. We wish to characterization the solution of the problem:

$$\max_{\mathbf{x} \in \mathbb{R}^n_+} u(\mathbf{x}) \text{ s.t. } \mathbf{p}^\intercal \mathbf{x} \leq m,$$

for some $\mathbf{p} \in \mathbb{R}^n_{++}$ and some $m > 0$. Since u is strictly increasing in each argument, $m > 0$ and $\mathbf{p} \in \mathbb{R}^n_{++}$, we know that if $\mathbf{x}^*$ is a solution, we must have $\mathbf{x}^* \in \mathbb{R}^n_{++}$ and $\mathbf{p}'\mathbf{x} = m$. Therefore, the first order condition given by Proposition 3.1.9 can be simplified to

$$\nabla u(\mathbf{x}^*) = \lambda \mathbf{p},$$

for some $\lambda > 0$. That is, for each $i, j \in \{1, \ldots, n\}$,

$$\frac{u_i(\mathbf{x}^*)}{u_j(\mathbf{x}^*)} = \frac{p_i}{p_j},$$

which is the familiar condition saying that marginal rate of substitution must be equal to relative price. We now check the second order condition. For any $\mathbf{x} \in \mathbb{R}^n$, $\lambda \geq 0$, the bordered Hessian matrix is given by:

$$H_{\mathcal{L}}(\mathbf{x}; \lambda) = \begin{pmatrix} 0 & -\mathbf{p}^\intercal \\ -\mathbf{p} & H_u(\mathbf{x}) \end{pmatrix}$$

which, by quasi-concavity of u, must have negative determinant. Therefore, the first order condition is necessary and sufficient, which is equivalent to solving the system:

$$\nabla u(\mathbf{x}) = \lambda \mathbf{p}$$

$$\mathbf{p}^\intercal \mathbf{x} = m$$

For example, for Cobb-Douglas utility $u(\mathbf{x}) = \prod_{i=1}^n x_i^{\alpha_i}$, with $\sum_{i=1}^n \alpha_i = 1$ solving the above system gives

$$x_i^* = \frac{\alpha_i m}{p_i}, \forall i \in \{1, \ldots, n\}.$$

Two remarks about this example:

1. The assumption of concavity of u is not necessary. In fact, we only need u to be quasi-concave. It turns out that u is quasi-concave if any only if the border Hessian for this problem is negative definite at the solution $\mathbf{x}^*$.

2. Cobb-Douglas utility does not satisfy the condition that u is strictly increasing in each argument, since when $x_i = 0$ for some i, $u(\mathbf{x})$ is always zero. However, it can be shown that the constraint must still be binding and thus the characterization is still valid.

3.1.4 First Order Approach in Infinite Dimensional Problem: Euler Equation

In some economic problems, mostly when *dynamics* are involved, the choice variable is no longer a finite dimensional variable. Instead, we sometimes need to choose over functions to maximize an objective. One of the most simplest problem is of form:

$$\max_{x\in AC([a,b])} \int_a^b \Lambda(x(t), x'(t))dt \text{ s.t. } x(a) = \alpha, x(b) = \beta,$$

where $AC([a, b])$ denotes the set of absolutely continuous functions on $[a, b]$. It turns out that the first order approach is still very useful in this type of problem. We introduce the most basic result below, you will see more about these problems in the second part of the course.

Proposition 3.1.11 (Euler Equation: Necessity)**.** *Let $\Lambda : \mathbb{R}^2 \to \mathbb{R}$ be a real-valued function such that H_Λ exists on $\mathbb{R}^n$. An absolute continuous function $x^* \in AC([a, b])$ is a solution of*

$$\max_{x\in AC([a,b])} \int_a^b \Lambda(x(t), x'(t))dt \text{ s.t. } x(a) = \alpha, x(b) = \beta$$

only if

$$\Lambda_1(x^*(t), x^{'*}(t)) - \frac{d}{dt}\Lambda_2(x^*(t), x^{'*}(t)) = 0 \tag{8}$$

for almost all $t \in [a, b]$.

Proof. Suppose that x^* is a solution of

$$\max_{x\in AC([a,b])} \int_a^b \Lambda(x(t), x'(t))dt \text{ s.t. } x(a) = \alpha, x(b) = \beta.$$

For any $x \in AC([a, b])$, let

$$V(x) := \int_a^b \Lambda(x(t), x'(t))dt.$$

Then for any $h \in AC([a, b])$ such that $h(a) = h(b) = 0$, let $\Gamma_h(\varepsilon) := V(x^* + \varepsilon h)$ for any $\varepsilon \in \mathbb{R}$. Since $x^* + \varepsilon h$ is still absolutely continuous on $[a, b]$ and $x^*(a) + \varepsilon h(a) = \alpha$, $x^*(b) + \varepsilon h(b) = \beta$, we must have

$$\Gamma_h(0) = \max_{\varepsilon\in\mathbb{R}} \Gamma_h(\varepsilon),$$

for all $h \in AC([a, b])$ with $h(a) = h(b) = 0$. By assumption, Γ is differentiable and thus it must be that

$$\Gamma_h'(0) = 0.$$

Evaluating Γ'_h, we have

$$\Gamma'_h(\varepsilon) = \int_a^b [\Lambda_1(x^*(t) + \varepsilon h(t), x'^*(t) + \varepsilon h'(t))h(t) + \Lambda_2(x^*(t) + \varepsilon h(t), x'^*(t) + \varepsilon h'(t))h'(t)]dt.$$

By integration by parts,

$$\int_a^b \Lambda_2(x^*(t) + \varepsilon h(t), x'^*(t) + \varepsilon h'(t))h'(t)dt = -\int_a^b \frac{d}{dt}\Lambda_2(x^*(t) + \varepsilon h(t), x'^*(t) + \varepsilon h'(t))h(t)dt,$$

since $h(a) = h(b) = 0$. Together,

$$\Gamma'_h(\varepsilon) = \int_a^b \left[\Lambda_1(x^*(t) + \varepsilon h(t), x'^*(t) + \varepsilon h'(t)) - \frac{d}{dt}\Lambda_2(x^*(t) + \varepsilon h(t), x'^*(t) + \varepsilon h'(t))\right] h(t)dt.$$

Thus,

$$0 = \Gamma'_h(0) = \int_a^b \left[\Lambda_1(x^*(t)), x'^*(t)) - \frac{d}{dt}\Lambda_2(x^*(t), x'^*(t))\right] h(t)dt,$$

for any $h \in AC([a, b])$ with $h(a) = h(b) = 0$. Therefore,

$$\Lambda_1(x^*(t), x'^*(t)) - \frac{d}{dt}\Lambda_2(x^*(t), x'^*(t)) = 0$$

for almost all $t \in [a, b]$, as desired. ■

By (8), to solve for the candidates for the solutions of the problem

$$\max_{x \in AC([a,b])} \int_a^b \Lambda(x(t), x'(t))dt \text{ s.t. } x(a) = \alpha, x(b) = \beta,$$

it is equivalent to solve a boundary value problem of a second-order differential equation, which is sometimes a much more tractable problem and the solution can be characterized.

3.2 Convex Analysis

3.2.1 Fundamental Properties of Convex Set and Convex Functions

Apart form first order approach, sometimes there is another set of methods for solving an optimization problem when the problem exhibits enough of *convexity*. This section explores some of the commonly used methods that rely on convexity of the problems. We begin this section with introducing some fundamental properties of convex sets and convex functions, which will be useful for economic analyses.

Definition 3.2.1. Let X be a linear space, a set $S \subset X$ is *convex* if for any $x, y \in X$, any $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in X$.

Definition 3.2.2. Let X be a linear space. A real-valued function $f : X \to \mathbb{R}$ if *convex* if for any $x, y \in X$, $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$. Furthermore, a function $f : X \to \mathbb{R}$ is *concave* if $-f$ is convex.

We first consider the case when $X = \mathbb{R}$. In this case, convex functions on subsets of $\mathbb{R}$ have some useful properties. Specifically, we will examine the degree of smoothness implied by convexity.

Lemma 3.2.1 (Chordal Slope Lemma)**.** *Let $(a,b) \subseteq \mathbb{R}$ be an open interval ans suppose that $f : (a,b) \to \mathbb{R}$ is convex. Then for any $x, y, z \in (a,b)$ with $x < y < z$,*

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y}.$$

In particular, fix any $x_0 \in (a,b)$, the function $x \mapsto \frac{f(x) - f(x_0)}{x - x_0}$ is increasing.

Proof. Consider any $x, y, z \in (a,b)$ such that $x < y < z$. Notice that

$$y = \frac{y - x}{z - x} z + \frac{z - y}{z - x} x.$$

By convexity of f, since $(y-x)/(z-x) \in (0,1)$ and $1 - (y-x)/(z-x) = (z-y)/(z-x)$,

$$f(y) \leq \frac{y - x}{z - x} f(z) + \frac{z - y}{z - x} f(x).$$

Rearranging, we have

$$f(y) \leq \frac{y - x}{z - x}(f(z) - f(x)) + f(x),$$

or

$$f(y) \leq \frac{z - y}{z - x}(f(x) - f(z)) + f(z)$$

Together, we have

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y},$$

as desired. ∎

A direct consequence of Lemma 3.2.1 is that

$$f'(x^-) := \lim_{h \downarrow 0} \frac{f(x) - f(x-h)}{h} \leq \lim_{h \downarrow 0} \frac{f(x+h) - f(x)}{h} =: f'(x^+).$$

Therefore, given any convex function $f : (a, b) \to \mathbb{R}$, for any $x, y \in (a, b)$ with $y > x$, using Lemma 3.2.1, we have

$$f'(x^-) \leq f'(x^+) \leq \frac{f(y) - f(x)}{y - x} \leq f'(y^-) \leq f'(y^+), \tag{9}$$

and therefore, $|f'(x^-)| < \infty$ and $|f'(x^+)| < \infty$ for all $x \in (a, b)$. (9) has several useful implications.

Proposition 3.2.1. *Let $f : (a, b) \to \mathbb{R}$ be a convex function. Then for any $[\underline{x}, \bar{x}] \subset (a, b)$, f is Lipschitz, and hence absolutely continuous, on $[\underline{x}, \bar{x}]$*

Proof. Fix any $\underline{x}, \bar{x} \in (a, b)$ with $\underline{x} < \bar{x}$. Using (9), let $K := \max\{|f'(\bar{x}^-)|, |f'(\underline{x}^+)|\} < \infty$, we then have

$$|f(x) - f(y)| \leq K|x - y|, \ \forall x, y \in [\underline{x}, \bar{x}].$$

■

Consequently, for any convex function on an interval $[a, b]$, the Fundamental Theorem of Calculus holds. That is, for any $x \in (a, b)$,

$$f(x) = f(a^+) + \int_a^x f(z)dz.$$

Since a convex function f is absolutely continuous on any $[\underline{x}, \bar{x}] \subset (a, b)$, f is differentiable almost everywhere on (a, b). In fact, f is differentiable except at countably mant points.

Proposition 3.2.2. *Let $f : (a, b) \to \mathbb{R}$ be a convex function. Then f is differentiable except at countably many points and the derivative f' is increasing.*

The next property is an essential characterization of convex function on $\mathbb{R}$.

Proposition 3.2.3. *Let $f : (a, b) \to \mathbb{R}$ be a real-valued function. f is convex if and only if it is a pointwise supremum of a class of affine functions.*

Proof. For necessity, let $\mathcal{L}$ be any class of affine functions. For any $x \in (a,b)$, define

$$f(x) := \sup_{L\in\mathcal{L}} L(x).$$

Then for any $x_1, x_2 \in (a,b)$, any $\lambda \in [0,1]$, since for any $L \in \mathcal{L}$, $L(\lambda x_1 + (1-\lambda)x_2) = \lambda L(x_1) + (1-\lambda)L(x_2)$,

$$\begin{aligned} f(\lambda x_1 + (1-\lambda)x_2) =& \sup_{L\in\mathcal{L}} L(\lambda x_1 + (1-\lambda)x_2) \\ \leq& \lambda \sup_{L\in\mathcal{L}} L(x_1) + (1-\lambda)\sup_{L\in\mathcal{L}} L(x_2) \\ =& \lambda f(x_1) + (1-\lambda) f(x_2). \end{aligned}$$

Conversely, for any $x_0 \in (a,b)$, by (9), the set $[f'(x_0^-), f'(x_0^+)]$ is nonempty. Take any $m \in [f'(x_0^-), f'(x_0^+)]$, by (9) again, we have

$$\frac{f(x)-f(x_0)}{x-x_0} \geq m,\ \forall x \in (x_0, b) \text{ and } \frac{f(x_0)-f(x)}{x_0-x} \leq m,\ \forall x \in (a, x_0).$$

Therefore, $f(x) \geq m(x-x_0)+f(x_0)$ for all $x \in (a,b)$. We may define a class of affine function $\mathcal{L}$ as $\mathcal{L} := \{L : (a,b) \to \mathbb{R} | L \leq f,\ L \text{ is an affine function}\}$. Clearly, the function $L_0(x) := m(x-x_0)+f(x_0)$ is in $\mathcal{L}$. Moreover, $L_0(x_0) = f(x_0)$ and therefore $f(x_0) = \sup_{L\in\mathcal{L}} L(x_0)$. Since $x_0 \in (a,b)$ is arbitrary, this completes the proof. ■

As a remark, in the proof of Proposition 3.2.3, we showed that for any $x_0 \in (a,b)$, there exists $m \in \mathbb{R}$ such that

$$f(x) \geq m(x-x_0) + f(x_0),\ \forall x \in (a,b).$$

Such m is called a *subgradient* of the convex function f at x_0 and the set $[f'(x_0^-), f'(x_0^+)]$ is called the *subdiffernetial* at of f at x_0. More generally, let $U \subseteq \mathbb{R}^n$ be a convex open set in $\mathbb{R}^n$ and let $f : U \to \mathbb{R}$ be a convex function. $m \in \mathbb{R}^n$ is a *subgradient* of f at $\mathbf{x}^0 \in U$ if for any $\mathbf{x} \in U$,

$$f(\mathbf{x}) \geq m^\intercal(\mathbf{x}-\mathbf{x}^0) + f(\mathbf{x}^0).$$

Notice that in the proof of Proposition 3.2.3, we have shown that if $n = 1$, for any $x_0 \in U$, a subgradient of f at x_0 must exists.

Proposition 3.2.3 has an immediate implication, also known as *Jensen's inequality*. We will return to this in chapter 4 after we formally define expectations of random variables.

We now begin exploring some methods for solving optimization problems by exploiting the convexity of the problem. The first result is an alternative of Proposition 1.2.5 that ensures existence of solutions when the objective is concave. Recall that Proposition 1.2.5 asserts that upper-semicontinuous functions defined on a compact set has maximum. However, when considering infinite dimensional space, compactness are harder to be ensured. The next Proposition states that in L^p spaces even if the domain is not compact, when the objective function is concave, maximum exists.

Proposition 3.2.4. *Let $X \subseteq \mathbb{R}$ be a measurable set. For any $p \in (1, \infty)$, for any closed, bounded and convex subset $C \subset L^p(X)$ and any continuous and concave functional $T : L^p(X) \to \mathbb{R}$,*

$$\operatorname*{argmax}_{f \in C} T(f) \neq \emptyset.$$

On the other hand, although first-order approach is widely used and is extremely useful for solving optimization problems, sometimes this approach might fail. For instance, when the problem is infinite dimensional or when the objective is not strictly convex/concave. Nevertheless, as long as the feasible set is convex and the objective is linear, there is another approach that can be used to characterize the solutions. We first illustrate the result by a simple example.

Example 3.2.1 (Consumer's Problem with Perfect Substitute Preference). Consider a specified version of Example 1.2.2, with $n = 2$ and $\mathbf{p} = (1, p)$, $p > 0$ and $u(x_1, x_2) = x_1 + x_2$. The consumer's problem is then

$$\max_{(x_1, x_2) \in \mathbb{R}^2_+} x_1 + x_2 \text{ s.t. } x_1 + p x_2 \leq m.$$

As you have seen in undergraduate economics, this problem cannot be solved by setting first-order condition to zero (and hence the Kuhn-Tucker condition as in Proposition 3.1.9 must be considered). It is well-known that the solution is given by

$$(x_1^*(p), x_2^*(p)) = \begin{cases} \{(m, 0)\}, & \text{if } p > 1 \\ \{(0, \frac{m}{p})\}, & \text{if } p < 1 \\ \{(x_1, x_2) \in \mathbb{R}^2_+ | x_1 + p x_2 = m\}, & \text{if } p = 1 \end{cases}.$$

Now notice that if we visualize the feasible set $\{(x_1, x_2) \in \mathbb{R}^2_+\}$ as a triangle on the first orthant, regardless the value of p, there must be at least *one* solution that is at the "corner" of this triangle. This observation is in fact much more general. We will now explore this result.

Definition 3.2.3. Let X be a linear space and $A \subset X$ be a set in X. The *convex hull* of A, $\text{co}(A)$ is defined as:

$$\text{co}(A) := \left\{ x \in X \middle| x = \sum_{i=1}^{n} \lambda_i x_i, \text{ for some } n \in \mathbb{N}, \{x_i\}_{i=1}^n \subseteq A, \{\lambda_i\}_{i=1}^n \subset [0,1] \text{ with } \sum_{i=1}^{n} \lambda_i = 1 \right\}$$

That is, the convex hull of a set A is the collection of elements in X that can be generated by finite convex combinations of the elements in A. It is not difficult to show that for any $A \subset X$, $\text{co}(A)$ is the *smallest* convex set in X that contains A, under the order of set inclusion.

If, furthermore, X has some topological structure, say X is a normed linear space, [17] we say that the *closed convex hull* of a set $A \subset X$ is the smallest closed convex set that contains A. It can also be shown that $\text{cl}(\text{co}(A))$ is exactly the closed convex hull of A.

A crucial element in convex sets is call the *extreme points.*

Definition 3.2.4. Let X be a linear space and $S \subseteq X$ be a convex set. We say that $x \in S$ is an *extreme point* if x cannot be written as a convex combination of two distinct points in S. That is: whenever there exists $\lambda \in (0,1)$ and $x_1, x_2 \in S$ such that

$$x = \lambda x_1 + (1-\lambda) x_2,$$

we must have $x_1 = x_2 = x$.

Hereafter, we use $\text{ext}(S)$ to denote the collection of extreme points of a convex set S.

The next three theorems establish the importance of extreme points.

Proposition 3.2.5. *Let X be a normed linear space and $S \subseteq X$ be a convex and compact subset of X. Then* $\text{ext}(S) \neq \emptyset$

[17] That X being a normed linear space is not necessary for the following results. These results hold whenever X is a *locally convex Hausdorff topological vector space.* For the sake of exposition, we use normed linear space to demostrate.

Theorem 3.2.1 (Carathédory's Theorem)**.** *For any $n \in \mathbb{N}$, let $S \subset \mathbb{R}^n$ be a subset of $\mathbb{R}^n$. Then for any $\mathbf{x} \in \text{co}(S)$, there exists $\{\mathbf{x}^i\}_{i=1}^{n+1} \subseteq S$ and $\{\lambda_i\}_{i=1}^{n+1} \subset [0,1]$ with $\sum_{i=1}^{n+1} \lambda_i = 1$ such that*

$$\mathbf{x} = \sum_{i=1}^{n+1} \lambda_i \mathbf{x}^i$$

Carathédory's theorem states that in an n-dimensional Euclidean space, any point in the convex hull of a set S can be represented by a convex combination of at most $n+1$ points in S, which means that finitely many elements in S is sufficient to describe the convex hull of a set S.

Theorem 3.2.2 (Krein-Milman Theorem)**.** *Let X be a normed linear space and let $S \subset X$ be a nonempty convex and compact subset of X. Then S is the closed convex hull of its extreme points.*

The Krein-Milman theorem implies that any convex and compact subset in a normed linear space can (almost) be described by it's extreme points. This sometimes allows us to reduce the complexity of the set we are dealing with by examining only the extreme points. For instance, combining the two theorems above, we can see that any point in a convex and compact subset in an n-dimensional Euclidean space can be written as a convex combination of at most $n+1$ of its extreme points.

Return to Example 3.2.1, we can see that the end "corners" of the triangle given by the budget constraint and nonnegativity constraint is precisely the extreme points of the budget set. In fact, in addition to being useful for describing a convex and compact set, extreme points are also useful for characterizing the solutions of some optimization problems.

Theorem 3.2.3 (Bauer Maximum Principle)**.** *Let X be a normed linear space and $S \subset X$ be a nonempty, convex and compact subset. Suppose that $f : S \to \mathbb{R}$ is upper-semicontinuous and convex. Then*

$$\underset{x \in S}{\text{argmax}}\, f(x) \bigcap \text{ext}(S) \neq \emptyset.$$

The Bauer maximum principle implies that whenever the objective function is convex and upper-semicontinuous and the feasible set is compact and convex, there is at least one extreme point of the feasible set that is the solution of the maximization problem. This is consistent with the observation in Example 3.2.1.

Example 3.2.2 (Monopolist Screening Revisited). Recall that in section 2.4, we formulated a problem for a monopolist who can design *any* selling mechanism to maximize her revenue. By Proposition 2.4.1 the problem can be reduced to

$$\max_{p:[0,1]\to[0,1]} \int_0^1 p(v)\left(v - \frac{1-F(v)}{f(v)}\right) f(v)dv \text{ s.t. } p \text{ is increasing.}$$

It is easy to verify that the set of increasing functions on $[0,1]$ that take values in $[0,1]$, denoted by $\mathcal{M}$, is convex and by Remark 2.2.3, $\mathcal{M}$ is a compact subset of $L^1([0,1])$. Furthermore, by linearity of Lebesgue integral, the objective

$$p \mapsto \int_0^1 p(v)\left(v - \frac{1-F(v)}{f(v)}\right) f(v)dv$$

is linear (and thus is convex) on $\mathcal{M}$. Together, by the Bauer maximum principle, there exists an extreme point of $\mathcal{M}$ that attains the maximum. As such, the problem is now reduced to finding extreme points of the set $\mathcal{M}$. We claim that $p \in \mathcal{M}$ is an extreme point of $\mathcal{M}$ if and only if $p(x) \in \{0,1\}$ for any $x \in [0,1]$. Indeed, suppose that there exists $x_0 \in [0,1]$ such that $p(x_0) \in (0,1)$. Define $\hat{p}$ as follows:

$$\hat{p}(x) := \begin{cases} p(x), & \text{if } p(x) \in [0, \frac{1}{2}] \\ 1 - p(x), & \text{if } p(x) \in (\frac{1}{2}, 1] \end{cases}.$$

Then,

$$q_1(x) := p(x) + \hat{p}(x) = \max\{2p(x), 1\}; \quad q_2(x) := p(x) - \hat{p}(x) = \max\{0, 2p(x) - 1\}$$

are both in $\mathcal{M}$. Furthermore, since $p(x_0) \in (0,1)$, $\hat{p}(x_0) \neq 0$ and therefore $q_1 \neq p$, $q_2 \neq p$. Finally, since $p = \frac{1}{2}q_1 + \frac{1}{2}q_2$, $p \notin \text{ext}(\mathcal{M})$.

Conversely, suppose that $p(x) \in \{0,1\}$ for any $x \in [0,1]$ and suppoes that $p = \lambda q_1 + \lambda q_2$ for some $\lambda \in (0,1)$ and $q_1, q_2 \in \mathcal{M}$. Then, if $p(x) = 0$, $q_1(x) = q_2(x) = 0$, whereas if $p(x) = 1$, $q_1(x) = q_2(x) = 1$ and hence $q_1 = q_2 = p$. Thus, $p \in \text{ext}(\mathcal{M})$.

As such, the problem is now reduced to choosing an extreme point to maximize the objective. Notice that since any $p \in \text{ext}(\mathcal{M})$ is increasing, p must take form of

$$p(x) = \begin{cases} 0, & \text{if } x \in [0, p^*) \\ 1, & \text{if } x \in (p^*, 1] \end{cases}$$

and $p(p^*) \in \{0,1\}$, for some $p^* \in [0,1]$. Notice that if we take such step function into the objective,

$$\int_0^1 p(x)\left(x - \frac{1-F(x)}{f(x)}\right) f(x)dx = \int_{p^*}^1 (xf(x) - (1-F(x)))dx,$$

using integration by parts, we have:

$$\int_0^1 p(x)\left(x - \frac{1-F(x)}{f(x)}\right) f(x)dx = \int_{p^*}^1 (xf(x) - (1-F(x)))dx = p^*(1-F(p^*)).$$

Therefore, the seemingly difficult infinite dimensional problem is now reduced to a one-dimensional problem

$$\max_{p^* \in [0,1]} p^*(1-F(p^*)).$$

As a final remark, notice that the mechanism corresponds to any extreme point, by Proposition 2.4.1, corresponds to a *posted price* mechanism: The seller can simply post a price and the buyer buys the object if his valuation is higher than the price and not buy if lower. The above analysis shows that, although the monopolist can choose *any* (possibly very complicated, by choosing a large M) selling mechanism, the simplest posted price mechanism can always achieve the maximized revenue!

3.2.2 Separation of Convex Sets and Supporting Hyperplane

The notion of *domination* is common in many economic problems, such as Pareto dominance or dominant strategy in games. Sometimes domination is a concept that is difficult to analyze as it involves comparing two alternatives under *all* circumstances. One way to characterize such concept is to find a *tangent plane* on the feasible set and regard the alternatives *as if* they are solutions to some constraint optimization problems that we are familiar with. This section is devoted to the mathematical backgrounds for this characterization. In fact, the separation and support hyperplane theorems introduced in this section have much wider applications. We will give a more abstract statement and then apply the results to Euclidean space and illustrate by an economic example.

Theorem 3.2.4 (Separating Hyperplane Theorem)**.** *Let X be a normed linear space and $A, B \subset X$ be two nonempty convex subsets of X with $\text{int}(A) \neq \emptyset$. Then $\text{int}(A) \cap B = \emptyset$ if and only if there exists $\alpha \in \mathbb{R}$ and a bounded linear functional $L : X \to \mathbb{R}$, $L \neq 0$, such that*

$$L(y) \geq \alpha \geq L(x), \forall y \in B, x \in A; \quad \alpha > L(x), \forall x \in \text{int}(A).$$

Theorem 3.2.5 (Supporting Hyperplane Theorem)**.** *Let X be a normed linear space and $S \subset X$ be a closed and convex subset. Suppose that* $\text{int}(S) \neq \emptyset$*. The for any $x \in S \backslash \text{int}(S)$, there exists a bounded linear functional $L : X \to \mathbb{R}$, $L \neq 0$, such that*

$$L(x) \geq L(y), \forall y \in S; \text{ and } L(x) > L(y), \forall y \in \text{int}(S).$$

Geometrically, the separating hyperplane theorem states that for any two convex sets in a normed linear space such that one of them has nonempty interior, one can always find a *hyperplane*, given by $\{x \in X | L(x) = \alpha\}$, for some bounded linear functional L and some number α, to "separate" these two sets so that each of them lies in exactly one of the half spaces generated by the hyperplane.

On the other hand, the supporting hyperplane theorem states that whenever a convex subset has nonempty interior, for any point on the boundary of that set, there exists a hyperplane that tangents to the set at that point. As an example, if we consider the upper-contour of $f : \mathbb{R} \to \mathbb{R}$, $\{(x, y) \in \mathbb{R}^2 | y \geq f(x)\}$ as a convex set in $\mathbb{R}^2$, Proposition 3.2.3 is exactly a version of the supporting hyperplane theorem, as each point on the boundary is precisely a pointwise supremum of a class of affine functions, which are exactly a class of hyperplanes.

In some normed linear spaces where inner product is well-defined (e.g. Hilbert space), for instance, the Euclidean space, it is well known that the set of bounded linear functionals on such space is isomorphic to the space itself and every bounded linear functional has a unique vector that corresponds to it by inner product operation. As a result, we have two immediate corollaries from the above theorems when $X = \mathbb{R}^n$.

Corollary 3.2.1 (Separating Hyperplane Theorem on $\mathbb{R}^n$)**.** *Let $A, B \subset X$ be two nonempty convex subsets of $\mathbb{R}^n$ with* $\text{int}(A) \neq \emptyset$*. Then* $\text{int}(A) \cap B = \emptyset$ *if and only if there exists $\alpha \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n \backslash \{0\}$ such that*

$$\mathbf{v}^\intercal \mathbf{y} \geq \alpha \geq \mathbf{v}^\intercal \mathbf{x}, \forall \mathbf{y} \in B, \mathbf{x} \in A; \quad \alpha > \mathbf{v}^\intercal \mathbf{x}, \forall \mathbf{x} \in \text{int}(A).$$

Corollary 3.2.2 (Supporting Hyperplane Theorem on $\mathbb{R}^n$)**.** *Let $S \subset \mathbb{R}^n$ be a closed and convex subset. Suppose that* $\text{int}(S) \neq \emptyset$*. The for any $\mathbf{x} \in S \backslash \text{int}(S)$, there exists $\mathbf{v} \in \mathbb{R}^n \backslash \{0\}$ such that*

$$\mathbf{v}^\intercal \mathbf{x} \geq \mathbf{v}^\intercal \mathbf{y}, \forall \mathbf{y} \in S; \text{ and } \mathbf{v}^\intercal \mathbf{x} > \mathbf{v}^\intercal \mathbf{y}, \forall y \in \text{int}(S).$$

In Euclidean spaces, it is easier to see the geometric interpretation of a *hyperplane*. Indeed, for any vector $\mathbf{v}$ and any $\alpha \in \mathbb{R}$, the set $\{\mathbf{x} \in \mathbb{R}^n | \mathbf{v}^\intercal \mathbf{x} = \alpha\}$ is exactly a "plane".

We now consider an example to illustrate the application of the separating and supporting hyperplane theorems.

Example 3.2.3 (Charazterizing Pareto Optimal Allocations). Consider an economy that is described by the set of commodities $\mathbb{R}^n$ and a vector of total endowments $\mathbf{e} = (e_1, \ldots, e_n)$ and $I \in \mathbb{N}$, $I \geq 2$ individuals with utility functions $\{u_i\}_{i=1}^I$. Assume that each u_i is continuous on $\mathbb{R}^n$ and is quasi-concave and strictly increasing. The set of feasible allocation in this economy, X, is defined by

$$X := \left\{ \mathbf{x} := (\mathbf{x}^i)_{i=1}^I \in \mathbb{R}^{n+I} \middle| \sum_{i=1}^I \mathbf{x}^i = \mathbf{e} \right\}.$$

We say that an allocation $\mathbf{x} \in X$ is *Pareto dominated* by an allocation $\mathbf{y} \in X$ if for any $i \in \{1, \ldots, I\}$,

$$u_i(\mathbf{x}^i) \leq u_i(\mathbf{y}^i)$$

and the strict inequality holds for some $i \in \{1, \ldots, I\}$. We say that an allocation is *efficient* if there is not other allocation in X that dominates it.

The set of efficient allocations are sometimes hard to identify. However, using the supporting hyperplane theorem, it reduces the problem to finding solutions to a family of optimization problems.

Proposition 3.2.6. *An allocation $\mathbf{x} \in X$ in the economy $(I, \{u_i\}_{i=1}^I, X)$ is efficient if and only if there exists $\{\lambda_i\} \subset [0,1]$ with $\sum_{i=1}^I \lambda_i = 1$ such that*

$$\mathbf{x} \in \operatorname*{argmax}_{\mathbf{x} \in X} \sum_{i=1}^I \lambda_i u_i(\mathbf{x}^i)$$

Proof. For necessity, suppose that $\mathbf{x}$ is dominated by some $\mathbf{y} \in X$. Then $u_i(\mathbf{x}) \geq u_i(\mathbf{y})$ for all $i \in \{1, \ldots, I\}$ and the strict inequality holds for some $i \in \{1, \ldots, I\}$. Therefore, for any $\{\lambda_i\} \subset [0,1]$ with $\sum_{i=1}^I \lambda_i = 1$,

$$\sum_{i=1}^I \lambda_i u_i(\mathbf{x}^i) \leq \sum_{i=1}^I \lambda_i u_i(\mathbf{y}^i).$$

If the inequality is strict, then the proof of this part is complete. If not, then it must be that $u_i(\mathbf{x}^i) = u_i(\mathbf{y}^i)$ whenever $\lambda_i > 1$ and thus $\lambda_i = 0$ for some $i \in \{1, \ldots, I\}$ such that $u_i(\mathbf{x}^i) < u_i(\mathbf{y}^i)$. Take any $i, j \in \{1, \ldots, I\}$ such that $u_i(\mathbf{x}^i) = u_i(\mathbf{y}^i)$ and $\lambda_i > 0$ and $u_j(\mathbf{x}^j) < u_j(\mathbf{y}^j)$ and $\lambda_j = 0$. Let $\tilde{\mathbf{y}}$ be defined as $\tilde{\mathbf{y}}^i = \mathbf{y}^i + 1$ and $\tilde{\mathbf{y}}^j = \mathbf{y}^j - 1$, and $\tilde{\mathbf{y}^l} = \mathbf{y}^l$ for all $l \notin \{i, j\}$. By construction, $\tilde{\mathbf{y}} \in X$. Then, since u_i is strictly increasing and $\lambda_j = 0$

$$\sum_{i=1}^{I} \lambda_i u_i(\mathbf{x}^i) \leq \sum_{i=1}^{I} \lambda_i u_i(\mathbf{y}^i) < \sum_{i=1}^{I} \lambda_i u_i(\tilde{\mathbf{y}}^i),$$

as desired.

For sufficiency, let

$$A := \{\mathbf{w} \in \mathbb{R}^I | w_i \leq u_i(\mathbf{x}^i), \forall i \in \{1, \ldots, I\}, \text{ for some } \mathbf{x} \in X\}.$$

Notice first that the set X is convex. It is then easy to verify, by using the definition of quasi-concavity, that A is convex. Furthermore, it can also be directly verified, by continuity of u_i, that if $\mathbf{x}$ is efficient, then $(u_i(\mathbf{x}^i))_{i=1}^I \in A \backslash \text{int}(A)$. Therefore, by the supporting hyperplane theorem, there exists $\mathbf{v} \in \mathbb{R}^I \backslash \{0\}$ such that

$$\sum_{i=1}^{I} v_i u_i(\mathbf{x}^i) \geq \sum_{i=1}^{I} v_i w_i,$$

for all $\mathbf{w} \in A$. Furthermore, since A is not bounded from below, it must be that $v_i \geq 0$ for all $i \in \{1, \ldots, I\}$ and thus $\sum_{i=1}^I v_i > 0$. Therefore, by defining $\lambda_i := v_i / \sum_{i=1}^I v_i$, we have

$$\sum_{i=1}^{I} \lambda_i u_i(\mathbf{x}^i) \geq \max_{\mathbf{w} \in A} \sum_{i=1}^{I} \lambda_i w_i \geq \max_{\mathbf{y} \in X} \sum_{i=1}^{I} \lambda_i u_i(\mathbf{y}^i).$$

The proof is then completed by the fact that $\mathbf{x}^i \in X$. ■

Proposition 3.2.6 implies that looking for an efficient allocation is equivalent to looking for the optimal allocation for a hypothetical social planner whose objective is some weighted sum of the individuals in the economy. This characterization has an important implication, called the *second welfare theorem*, which states that any efficient allocation can be supported by an equilibrium under some exchange economy.

3.2.3 Duality Theorem of Constraint Optimization

Recall that in section 3.1, we have seen the problem

$$\begin{aligned} &\sup_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \\ &\text{s.t. } g_j(\mathbf{x}) \geq 0,\ \forall j \in \{1,\dots,J\} \\ &\qquad h_l(\mathbf{x}) = 0,\ \forall l \in \{1,\dots,k\} \end{aligned} \tag{10}$$

and used the first order approach to solve the problem. In fact, such problem can be studied in another approach, called the *duality method*, which does not require as much differentiabiliy as the first-order approach. To formulate the result, we let the optimal value of (10) (could be infinity) be p^*. Notice that we have not imposed any assumptions on the functions f, $\{g_j\}_{j=1}^J$ and $\{h_l\}_{l=1}^k$.

Consider now the *dual* problem of (10):

$$\inf_{\boldsymbol{\lambda}\in\mathbb{R}_+^J,\,\boldsymbol{\mu}\in\mathbb{R}^k} q(\boldsymbol{\lambda},\boldsymbol{\mu}), \tag{11}$$

where

$$q(\boldsymbol{\lambda},\boldsymbol{\mu}) := \sup_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \boldsymbol{\lambda}^\intercal \mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^\intercal \mathbf{h}(\mathbf{x}).$$

Notice that $q : \mathbb{R}_+^J \times \mathbb{R}^k \to \mathbb{R}$ is convex as it is a pointwise supremum of affine functions. Let d^* be the optimal value of (11).

Theorem 3.2.6 (Weak Duality Theorem)**.** $d^* \geq p^*$

Proof. For any $\mathbf{x} \in \mathbb{R}^n$ such that $g_j(\mathbf{x}) \geq 0$ for all $j \in \{1,\dots,J\}$ and $h_l(\mathbf{x}) = 0$ for all $l \in \{1,\dots,k\}$ and for any $\boldsymbol{\lambda} \in \mathbb{R}_+^J$ and any $\boldsymbol{\mu} \in \mathbb{R}^k$,

$$f(\mathbf{x}) \leq f(\mathbf{x}) + \sum_{j=1}^J \lambda_j g_j(\mathbf{x}) + \sum_{l=1}^k \mu_l h_l(\mathbf{x}).$$

As such, for any $\mathbf{x} \in \mathbb{R}^n$ such that $g_j(\mathbf{x}) \geq 0$ for all $j \in \{1,\dots,J\}$ and $h_l(\mathbf{x}) = 0$ for all $l \in \{1,\dots,k\}$,

$$f(\mathbf{x}) \leq \sup_{\mathbf{x}\in\mathbb{R}^n} \left[f(\mathbf{x}) + \sum_{j=1}^J \lambda_j g_j(\mathbf{x}) + \sum_{l=1}^k \mu_l h_l(\mathbf{x}) \right] = q(\boldsymbol{\lambda},\mu).$$

Thus, $p^* \leq d^*$, as desired. ■

Weak duality theorem allows us to give an upper bound of the value of (10). However, the bound may not be tight. There might be cases in which $p^* < d^*$. We call the value $d^* - p^* \geq 0$ the *duality gap* of the problem (10). Under certain sufficient conditions, the duality gap will be zero, which we refer as the *strong duality theorem*. We give one sufficient that is commonly seen in economics here.

Proposition 3.2.7 (Slater's Condition)**.** *If f is quasi-concave, the feasible set is convex and there exists $\mathbf{x} \in \mathbb{R}^n$ such that $g_j(\mathbf{x}) > 0$ for all $j \in \{1, \ldots, J\}$, then the duality gap for problem* (10) *is zero.*

Whenever the duality gap is zero, solving for the constraint optimization problem (10) is equivalent to solving the dual problem (11), which is presumably easier since the objective is now convex.

Example 3.2.4 (Utility Maximization and Cost Minimization). Consider again the consumer's problem in Example 1.2.2. Assume that u is continuous, strictly increasing in each arguments, and strictly quasi-concave. Also, suppose that $m > 0$. Consider again the problem.

$$\max_{\mathbf{x} \in \mathbb{R}^n} u(\mathbf{x}) \text{ s.t. } \mathbf{p}^\intercal \mathbf{x} \leq m \tag{12}$$

Recall that we say the (unique) solution to this problem is called the *Marshallian demand*, we denote it by $\mathbf{x}^M(p, m)$. By Slater's condition, strong duality theorem holds and therefore

$$u(\mathbf{x}^M(\mathbf{p}, m)) = \min_{\lambda \geq 0} \max_{\mathbf{x} \in \mathbb{R}^n} [u(\mathbf{x}) + \lambda(\mathbf{p}^\intercal \mathbf{x} - m)]. \tag{13}$$

On the other hand, consider a related *expenditure minimization problem*,

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{p}^\intercal \mathbf{x} \text{ s.t. } u(\mathbf{x}) \geq \bar{u}. \tag{14}$$

We refer the solution to this problem as *Hicksian demand*, denoted by $\mathbf{x}^H(p, \bar{u})$. Again, by the strong duality theorem, we have

$$\mathbf{p}^\intercal \mathbf{x}^H(\mathbf{p}, \bar{u}) = \max_{\mu \leq 0} \min_{\mathbf{x} \in \mathbb{R}^n} [\mathbf{p}^\intercal \mathbf{x} + \mu(u(\mathbf{x}) - \bar{u})].$$

Fix any $\mathbf{p} \in \mathbb{R}^n_{++}$ and any $\bar{u} \in \mathbb{R}$, let $e(\mathbf{p}, \bar{u})$ denote the optimal expenditure given $\bar{u}$ and $\mathbf{p}$. We will now show a well-known identity, called the *Slutsky* identity by the observations above strong duality theorem.

Corollary 3.2.3 (Slutsky's Identity)**.** $\mathbf{x}^M(\mathbf{p}, e(\mathbf{p}, \bar{u})) = \mathbf{x}^H(\mathbf{p}, \bar{u})$.

Proof. For conciseness of notations, we let $\mathbf{x}^* := \mathbf{x}^M(\mathbf{p}, e(\mathbf{p}, \bar{u}))$ and $\mathbf{y}^* := \mathbf{x}^H(\mathbf{p}, \bar{u})$. Since u is strictly quasi-concave, the solution must be unique. Therefore, by the strong duality theorem, since

$$u(\mathbf{x}^*) = \min_{\lambda \geq 0} \max_{\mathbf{z} \in \mathbb{R}^n} [u(\mathbf{z}) + \lambda(e(\mathbf{p}, \bar{u}) - \mathbf{p}^\intercal \mathbf{z}))],$$

$$\mathbf{p}^\intercal \mathbf{y}^* = \max_{\mu \leq 0} \min_{\mathbf{z} \in \mathbb{R}^n} [\mathbf{p}^\intercal \mathbf{z} + \mu(u(\mathbf{z}) - \bar{u})].$$

it must be that $\mathbf{p}^\intercal \mathbf{x}^* = \mathbf{p}^\intercal \mathbf{y}^* = e(\mathbf{p}, \bar{u})$ and $u(\mathbf{y}^*) = \bar{u}$. To see this, let

$$g(\lambda) := \sup_{\mathbf{z} \in \mathbb{R}^n} [u(\mathbf{z}) + \lambda(e(\mathbf{p}, \bar{u}) - \mathbf{p}^\intercal \mathbf{z})]$$

so that $u(\mathbf{x}^*) = \min_{\lambda \geq 0} g(\lambda)$. Since g is convex, by Proposition 3.2.4, $\operatorname{argmin}_{\lambda \geq 0} g(\lambda) \neq \emptyset$. Furthermore, notice that for any $\lambda^* \in \operatorname{argmin}_{\lambda \geq 0} g(\lambda)$, it must be that $\lambda^* > 0$, since otherwise $g(\lambda^*) = \sup_{\mathbf{z} \in \mathbb{R}^n} [u(\mathbf{z}) + \lambda^*(e(\mathbf{p}, \bar{u}) - \mathbf{p}^\intercal \mathbf{z})] = \infty$, a contradiction. As such,

$$u(\mathbf{x}^*) = g(\lambda^*) = \sup_{\mathbf{z} \in \mathbb{R}^n} [u(\mathbf{z}) + \lambda^*(e(\mathbf{p}, \bar{u}) - \mathbf{p}^\intercal \mathbf{z})]$$

and therefore it must be the $\mathbf{p}^\intercal \mathbf{x}^* \geq e(\mathbf{p}, \mathbf{x}^*)$. Together with the definition of $\mathbf{x}^*$ from (12), we have $\mathbf{p}^\intercal \mathbf{x}^* = e(\mathbf{p}, \bar{u})$. By uniqueness of solution of (14), it must be that $\mathbf{x}^* = \mathbf{y}^*$.

■

Slutsky's identity is particularly useful for deriving the well-known *Slutsky's equation*, which draws a connection between the price elasticities of Marshallian and Hicksian demand and is the mathematical foundation of the substutition effect-income effect decomposition. We will return to this in the next section.

3.3 Application: Information Design—Bayesian Persuasion

Contrast to mechanism design introduced in the previous chapter, which focuses on how to induce certain behavior and outcome by designing rules of play, the study of information design examines how the designer can affect agents' behavior via the *information channel*. By providing different information to the agents, the designer can induce different actions since agents' behavior depends on their posterior beliefs. In this section, we will introduce

a well-known framework in the information design literature, called *Bayesian Persuasion*. The discussions below are mostly based on Kamenica & Gentzkow (2011).

Consider the following environment. There is a sender S and a receiver R, a finite state space Ω and a common prior $p \in \Delta(\Omega)$. Assume that $p(\omega) > 0$ for all $\omega \in \Omega$. The receiver can take action in a finite set A. Both the sender and the receiver have VNM utility function $u_S : \Omega \times A \to \mathbb{R}$ and $u_R : \Omega \times A \to \mathbb{R}$, respectively. The sender can commit to an *information structure* (S, π), where S is a finite set that contains all possible *signals* and $\pi : \Omega \to \Delta(S)$ is a conditional distribution conditioning on the state.

Given an information structure (S, π), when the receiver sees signal $s \in S$, according to Bayes' rule, his posterior belief about the state is given by:

$$q_s(\omega) = \frac{\pi(s|\omega)p(\omega)}{\sum_{\omega \in \Omega} \pi(s|\omega)p(\omega)}, \ \forall \omega \in \Omega.$$

The receiver then takes action $a^*(s) \in A$ to maximize payoff. That is, the receiver solves

$$\max_{a \in A} \sum_{\omega \in \Omega} u_R(a, \omega) q_s(\omega).$$

To ensure that the solution exists, when the receiver is indifferent, he will always select whichever is preferred by the sender (see Chapter 1 for the discussion of upper-semicontinuity). As such, a^* is a function that maps from signal s to actions. Anticipating this behavior, the sender then wishes to find an information structure (S, π) to maximize her expected payoff:

$$\sum_{\omega \in \Omega} \sum_{s \in S} u_S(a^*(s), \omega) \pi(s|\omega) p(\omega).$$

To solve the sender's problem, first observe that for any information structure (S, π) and for any signal realization $s \in S$, what really matters for the receiver's decision making (and hence the sender's payoff) is the induced posterior belief q_s. Since the singla s is drawn form the conditional distribution $\pi(\cdot|\omega)$ for any state ω, all the possible posteriors $\{q_s\}_{s \in S}$ follows the distribution induced by the marginal $\sum_{\omega \in \Omega} \pi(s|\omega)p(\omega)$. Furthermore, by Bayes' rule, the expected posterior under this distribution is exactly the prior:

$$\begin{aligned}\sum_{s \in S} q_s(\omega) \sum_{\omega' \in \Omega} \pi(s|\omega')p(\omega') &= \sum_{s \in S} \frac{\pi(s|\omega)p(\omega)}{\sum_{\omega' \in \Omega} \pi(s|\omega')p(\omega')} \sum_{\omega' \in \Omega} \pi(s|\omega')p(\omega') \\ &= \sum_{s \in S} \pi(s|\omega)p(\omega) = p(\omega), \ \forall \omega \in \Omega\end{aligned}$$

In fact, there is a one-to-one and onto mapping between information structures (S, π) and *distribution of posterior beliefs*. That is, all the possible information structures can be represented by the distribution of posterior beliefs that comply to Bayes' rule:[18]

$$\left\{\tau \in \Delta(\Delta(\Omega))\middle| \int_{\Delta(\Omega)} q\tau(dq) = p\right\}.$$

Given any posterior $q \in \text{supp}(\tau)$, the receiver will then take action $a^*(q)$. Let

$$v(q) := \sum_{\omega\in\Omega} u_S(a^*(q), \omega)q(\omega)$$

be the sender's expected payoff when the receiver's posterior is q. Notice that since the receiver always takes the sender-preferred action when indifferent, v is upper-semicontinous. The sender's problem can be simplified to:

$$\sup_{\tau\in\Delta(\Delta(\Omega))} \int_{\Delta(\Delta(\Omega))} v(q)\tau(dq) \text{ s.t. } \int_{\Delta(\Delta(\Omega))} q\tau(dq) = p. \tag{15}$$

Using the techniques introduced above, we can have a nice characterization of the solution to the sender's problem. To begin with, let

$$\mathcal{V} := \text{co}(\{(q, v) \in \Delta(\Omega) \times \mathbb{R} | v \leq v(p)\})$$

be the convex hull of the lower-contour set of v and let

$$V(q) := \sup\{v \in \mathbb{R} | (v, p) \in \mathcal{V}\}, \forall q \in \Delta(\Omega).$$

By construction, the function $V : \Delta(\Omega) \to \mathbb{R}$ is concave, and is in fact the smallest concave function that majorizes v. The function V is often referred as the *concave closure* of v and this process is often called *concavification*. With the definition of V, we can state the following result that serves as a cornerstone of the Bayesian persuasion literature.

Proposition 3.3.1.

$$V(p) = \sup_{\tau\in\Delta(\Delta(\Omega))} \int_{\Delta(\Omega)} v(q)\tau(dq) \text{ s.t. } \int_{\Delta(\Omega)} q\tau(dq) = p.$$

[18]Recall that $\Delta(\Omega)$ denotes the set of probability distributions on the finite set Ω, which is a compact and convex subset of $\mathbb{R}^|\Omega|$ with dimension $|\Omega| - 1$. $\Delta(\Delta(\Omega))$ denotes the set of probability distributions on the set $\Delta(\Omega)$. At this stage, it is not rigorously defined. Using the notions introduced in chapter 4, $\Delta(\Delta(\Omega))$ is the set of probability measures on the measurable set $\Delta(\Omega)$ endowed with the Borel σ-algebra. From the discussions in section 4.5, this is a compact set under the weak-* topology and a convex set.

Proof. To begin with, let $W(p)$ be the value of the sender's problem. Consider first the dual of this constraint maximization problem:

$$D(\lambda) = \sup_{\tau \in \Delta(\Delta(\Omega))} \int_{\Delta(\Omega)} (v(q) - \lambda^\intercal q)\tau(dq).$$

By Theorem 3.2.6,

$$D^* := \inf_{\lambda \in \mathbb{R}^{|\Omega|}} D(\lambda) \geq W(p). \tag{16}$$

By construction, $\mathcal{V}$ is convex, with $\text{int}(\mathcal{V}) \neq \emptyset$. Moreover, since v is upper-semicontinous, $(p, V(p)) \in \mathcal{V} \backslash \text{int}(\mathcal{V})$. Thus, by Corollary 3.2.1, there exists a vector $(u, w) \in \mathbb{R} \times \mathbb{R}^{|\Omega|}$ with $u > 0$ (why?) such that

$$uV(p) + w^\intercal p \geq uv + w^\intercal q, \, \forall (v, q) \in \mathcal{V}. \tag{17}$$

If $V(p) = v(p)$, then for any $q \in \Delta(\Omega)$, (17) implies that

$$v(p) + \frac{1}{u} \cdot w^\intercal p \geq v(q) + \frac{1}{u} \cdot w^\intercal q, \, \forall q \in \Delta(\Omega). \tag{18}$$

We will now verify that $\tau = \delta_{\{p\}}$ solves the sender's problem and hence $W(p) = v(p) = V(p)$. Using (16), it suffices to find some $\lambda \in \mathbb{R}^{|\Omega|}$ such that

$$\delta_{\{p\}} \in \underset{\tau \in \Delta(\Delta(\Omega))}{\text{argmax}} \int_{\Delta(\Omega)} (v(q) - \lambda^\intercal q)\tau(dq),$$

since this would then imply that then exists a feasible τ in the primal problem that yields expected payoff $D(\lambda)$. $\delta_{\{p\}}$ being feasible then implies that

$$W(p) \geq D(\lambda) \geq D^* \geq W(p)$$

and hence $D(\lambda) = W(p)$, which means that $\delta_{\{p\}}$ indeed solves the primal problem. To find this λ, simply take

$$\lambda := \frac{-1}{u} \cdot w.$$

(18) can then be written as

$$v(p) - \lambda^\intercal p \geq v(q) - \lambda^\intercal q, \, \forall q \in \Delta(\Omega).$$

As such, if $\tau = \delta_{\{p\}}$,

$$\int_{\Delta(\Omega)} (v(q) - \lambda^\intercal q)\tau(dq) \geq \int_{\Delta(\Omega)} (v(q) - \lambda^\intercal q)\tau'(dq), \, \forall \tau' \in \Delta(\Delta(\Omega)).$$

Therefore, when $\lambda = -1/u \cdot w$, $\delta_{\{p\}}$ yields expected payoff $D(\lambda)$, as desired.

On the other hand, if $V(p) > v(p)$, by Theorem 3.2.1, there exists $\{(v_i, q_i)\}_{i=1}^{|\Omega|+2} \subseteq \{(v, q) \in \mathbb{R} \times \Delta(\Omega) | q \leq v(q)\}$ and $\alpha \in \Delta(\{1, \ldots, |\Omega| + 2\})$ such that

$$\sum_{i=1}^{|\Omega|+2} \alpha_i (q_i, v_i) = (p, V(p)). \tag{19}$$

Notice that (17) and the fact that $u > 0$ implies that $v_i = v(q_i)$ for all $i \in \{1, \ldots, |\Omega| + 2\}$. Furthermore, together with (19), we must have

$$uv(q_i) + w^\intercal q_i = uV(p) + w^\intercal p \geq uv(q) + w^\intercal q, \, \forall q \in \Delta(\Omega), \, \forall i \in \{1, \ldots, |\Omega| + 2\}.$$

Using the same argument, let τ be defined as $\tau(q_i) = \alpha_i$ for all $i \in \{1, \ldots, |\Omega| + 2\}$ and let $\lambda := -1/u \cdot w$, we then have

$$\int_{\Delta(\Omega)} (v(q) - \lambda^\intercal q) \tau(dq) \geq \int_{\Delta(\Omega)} (v(q) - \lambda^\intercal q) \tau'(dq), \, \forall \tau' \in \Delta(\Delta(\Omega)).$$

Thus, τ indeed solves the sender's problem and $W(p) = V(p)$ as (19) dictates. This completes the proof. ■

Proposition 3.3.1 implies that, to solve the sender's optimal persuasion problem, one only needs to concavify the function v, as the value of the sender's problem is exactly the concave closure of v. This result extremely useful when the the number of states is small. For instance, consider an example in which $\Omega = \{0, 1\}$, $A = \{0, 1\}$, $u_S(a, \omega) = a$ and $u_R(a, \omega) = \mathbf{1}\{a = \omega\}$. Clearly, the receiver follows a cutoff strategy so that $a^*(q) = \mathbf{1}\{q \geq \frac{1}{2}\}$. The sender's induced payoff function v is then

$$v(q) = \begin{cases} 1, & \text{if } q \in [\frac{1}{2}, 1] \\ 0, & \text{if } q \in [0, \frac{1}{2}). \end{cases}.$$

The concave closure of v is then

$$V(q) = \min\{2q, 1\}.$$

As such, for any prior $p \geq \frac{1}{2}$, the best for the sender is to provide no further information to the receiver and for any $p < \frac{1}{2}$, the best information structure for the sender is the ones such that with probability $\frac{1}{2}$, the receiver's posterior is $q = 0$ and with probability $\frac{1}{2}$, his posterior is $q = \frac{1}{2}$, this can be done by an information structure that take form of $S = \{0, 1\}$, $\pi(0|1) = 0$, $\pi(0|0) = \frac{1}{2(1-p)}$.

3.4 Comparative Statics

In addition to solving optimization problems and equilibrium, one of the main tasks in economic analysis is *comparative statics*: examining how the solutions will change as the environment changes. Comparative statics is a useful tool for exploring *causality*. Once we have a clear prediction of how the *endogenous variables* (optimal solutions, equilibria etc) vary as the *exogenous variables* (variables that are taken as given when solving the model), we then have a causal interpretation of how the economy reacts to change of the environment. For instance, how will the equilibrium price in a market change when government raise taxes; how will the monopolist's revenue change as number of buyers increase. In this section, we will introduce some widely used methods for comparative statics.

3.4.1 Envelope Theorem

Let us motivate by a simple example. Let $f : [a, b] \times [0, 1] \to \mathbb{R}$ be a strictly concave function such that $f_{11}(x, \theta)$ exists for all $x \in [a, b]$ and is continuous on $[a, b]$, for all $\theta \in [0, 1]$. Consider the maximization problems parameterized by θ:

$$V(\theta) = \max_{x \in [a,b]} f(x, \theta).$$

Since f is concave, and continuous, there exists a unique solution, which we denote by $x^*(\theta)$. For the sake of exposition, we assume that $x^*(\theta) \in (a, b)$. Therefore, by the discussions in section 3.1, $x^*(\theta)$ solves the first order condition. That is:

$$f_1(x^*(\theta), \theta) = 0.$$

If we further assume that x^* is differentiable on $(0, 1)$,[19] then by chain rule, we have:

$$V'(\theta) = f_1(x^*(\theta), \theta)x'(\theta) + f_2(x^*(\theta), \theta) = f_2(x^*(\theta), \theta),$$

where the second equality follows from the first order condition. As a result, we know that if we wish to examine the change of optimal value as the parameter changes, which is taken as given in the optimization problem, we only need to consider the *direct effect* of the changing

[19] By the assumptions here, this is directly implied by the implicit function theorem as we will introduce later

the parameter on the objective and ignore the *indirect effect* induced by optimal choices. This turns out to be an extremely useful observation, since the indirect effects can sometimes be very complicated.

We now begin examining a more general version of the observation above. Let X be a nonempty set and let $f : X \times [0,1] \to \mathbb{R}$ be a real-valued function. Let

$$V(\theta) := \sup_{x \in X} f(x, \theta), \, \forall \theta \in [0,1]$$

be the optimal value as a function of θ and let

$$X^*(\theta) := \{x \in X | f(x, \theta) = V(\theta)\}$$

be the set of maximizers (could be empty). Notice that so far we have not imposed any assumptions on the choice set X nor the objective function f, and yet we have the following result:

Lemma 3.4.1 (Milgrom & Segal (2002))**.** *Consider any $\theta \in [0,1]$ and any $x^* \in X^*(\theta)$ and suppose that $f_2(x^*, \theta)$ exists. If $\theta > 0$ and V is left-differentiable at θ,* [20] *then $V'(\theta^-) \leq f_2(x^*, \theta)$. If $\theta < 1$ and V is right-differentiable at θ,*[21] *then $V'(\theta^+) \geq f_2(x^*, \theta)$. If $\theta \in (0,1)$ and V is differentiable at θ, then $V'(\theta) = f_2(x^*, \theta)$.*

Proof. By definition of V and x^*, for any $\theta' \in [0,1]$

$$f(x^*, \theta') - f(x^*, \theta) \leq V(\theta') - V(\theta).$$

Suppose that $\theta < 1$ and V is right-differentiable at θ. For any $\theta' \in (\theta, 1)$, we have

$$\frac{f(x^*, \theta') - f(x^*, \theta)}{\theta' - \theta} \leq \frac{V(\theta') - V(\theta)}{\theta' - \theta}$$

and thus, since $f_2(x^*, \theta)$ exists and V is right-differentiable at θ, we have

$$f_2(x^*, \theta) = \lim_{\theta' \downarrow \theta} \frac{f(x^*, \theta') - f(x^*, \theta)}{\theta' - \theta} \leq \lim_{\theta' \downarrow \theta} \frac{V(\theta') - V(\theta)}{\theta' - \theta} =: V'(\theta^+),$$

as desired. The proof for $\theta > 0$ and V begin left-differentiable is analogous and the result for $\theta \in (0,1)$ and V being differentiable follows from combining both cases. ■

[20] That is, $V'(\theta^-) := \lim_{\delta \downarrow 0} \frac{V(\theta) - V(\theta - \delta)}{\delta}$ exists.

[21] That is, $V'(\theta^+) := \lim_{\delta \downarrow} \frac{V(\theta + \delta) - V(\theta)}{\delta}$ exists.

Lemma 3.4.1 does not rely on any assumptions on the choice set nor the nature of the solution of the maximization problem. However, it does rely on the assumption that the optimal value V is differentiable. As V is an endogenous object, we wish to find sufficient conditions under which differentiability of V can be ensured. This is given by the next result.

Theorem 3.4.1 (Envelope Theorem (Milgrom & Segal, 2002))**.** *Suppose that $f(x,\cdot)$ is absolutely continuous on $[0,1]$ for all $x \in X$ and that there exists a (Lebesgue) integrable function $g : [0,1] \to \mathbb{R}$ such that $|f_2(x,\theta)| \leq g(\theta)$ for almost all $\theta \in [0,1]$, for all $x \in X$. Then V is absolutely continuous on $[0,1]$. Furthermore, if $f(x,\cdot)$ is differentiable for all $x \in X$ and $X^*(\theta) \neq \emptyset$ for all $\theta \in [0,1]$, then for any selection x^* of X^*,*[22]

$$V(\theta) = V(0) + \int_0^\theta f_2(x^*(z), z)dz. \tag{20}$$

In particular,

$$V'(\theta) = f_2(x^*(\theta), \theta) \tag{21}$$

almost everywhere.

Proof. Observe that for any $\theta, \theta' \in [0,1]$ with $\theta < \theta'$, by definition of V, we observe that

$$-\sup_{x\in X} |f(x,\theta') - f(x,\theta)| \leq \sup_{x\in X} f(x,\theta') - \sup_{x\in X} f(x,\theta) = V(\theta') - V(\theta) \leq \sup_{x\in X} |f(x,\theta') - f(x,\theta)|.$$

Thus,

$$\begin{aligned} |V(\theta') - V(\theta)| \leq & \sup_{x\in X} |f(x,\theta') - f(x,\theta)| \\ & = \sup_{x\in X} \left| \int_\theta^{\theta'} f_2(x,z)dz \right| \qquad \text{(Fundamental Theorem of Calculus)} \\ & = \leq \int_\theta^{\theta'} \sup_{x\in X} |f_2(x,z)|dz \leq \int_\theta^{\theta'} |g(z)|dz. \end{aligned}$$

Therefore, for any $\varepsilon > 0$, by countable additivity and characterization of integrability of Lebesgue integral (Proposition 2.2.3), there exists $\delta > 0$ such that $\lambda(A) < \delta$ implies

$$\int_A |g(z)|dz < \varepsilon$$

[22]For nonempty sets X, Y. Let $\Gamma \Rightarrow Y$ be a correspondence. $\gamma : X \to Y$ is a selection of Γ if $\gamma(x) \in \Gamma(x)$ for all $x \in X$.

and therefore for any finite disjoint intervals $\{[a_k, b_k]\}_{k=1}^n$ with $\lambda(\cup_{k=1}^n [a_k, b_k]) = \sum_{k=1}^n (b_k - a_k) < \delta$, we must have

$$\sum_{k=1}^n |V(b_k) - V(a_k)| \leq \sum_{k=1}^n \int_{a_k}^{b_k} |g(z)|dz = \int_{\cup_{k=1}^n [a_k, b_k]} |g(z)|dz < \varepsilon$$

and hence V is absolutely continuous. By the Fundamental Theorem of Calculus,

$$V(\theta) = V(0) + \int_0^\theta V'(z)dz.$$

If, furthermore, $f(x, \cdot)$ is differentiable for all $x \in X$ and X^* is nonempty-valued, by Lemma 3.4.1, since V is absolutely continuous and thus is differentiable almost everywhere on $(0, 1)$, for any selection x^* of X^*, for almost all $\theta \in (0, 1)$,

$$V'(\theta) = f_2(x^*(\theta), \theta).$$

Together, we have

$$V(\theta) = V(0) + \int_0^\theta f_2(x^*(z), z)dz.$$

for all $\theta \in [0, 1]$. ■

To sum up, the envelope theorem states that the derivative of the optimal value as a function of the parameter is exactly the partial derivative of the objective with respect to the parameter, evaluated at the optimal choice under the parameter. This is summarized by (21). Whenever the objective is differentiable with respect to the parameter, this is true for almost all the parameters.

Using Theorem 3.4.1 and the duality theorem in the previous section, we then have the envelope theorem for constraint optimization problems as well.

Corollary 3.4.1 (Envelope Theorem for Constraint Optimization)**.** *Let* $f : \mathbb{R}^n \times [0, 1] \to \mathbb{R}$, $\{g^j : \mathbb{R}^n \times [0, 1] \to \mathbb{R}\}_{j=1}^n$, $\{h^l : \mathbb{R}^n \times [0, 1] \to \mathbb{R}\}_{l=1}^k$ *be real-valued functions that* $f(\mathbf{x}, \cdot)$, $g^j(\mathbf{x}, \cdot)$ *and* $h_l(\mathbf{x}, \cdot)$ *are differentiable for all* $j \in \{1, \ldots, J\}$, $l \in \{1, \ldots, k\}$, $\mathbf{x} \in \mathbb{R}^n$. *Let*

$$X(\theta) := \{\mathbf{x} \in \mathbb{R}^n | g^j(\mathbf{x}, \theta) \geq 0, \forall j \in \{1, \ldots, J\}; h^l(\mathbf{x}, \theta) = 0, \forall l \in \{1, \ldots, k\}\}.$$

and let

$$V(\theta) := \sup_{\mathbf{x} \in X(\theta)} f(\mathbf{x}) \tag{22}$$

$$X^*(\theta) := \{\mathbf{x} \in X(\theta) | f(\mathbf{x}, \theta) = V(\theta)\}$$

be the optimal value and the solutions of the constraint maximization problem, respectively. Suppose that the duality gap for (22) *is zero and that* X^* *is nonempty-valued. Then for almost all* $\theta \in (0,1)$*, any selection* x^* *of* X^*

$$V'(\theta) = f_\theta(\mathbf{x}^*(\theta), \theta) + \sum_{j=1}^{J} \lambda_j(\theta) g_\theta^j(\mathbf{x}, \theta) + \sum_{l=1}^{k} \mu_l(\theta) h_\theta^l(\mathbf{x}, \theta),$$

where $\boldsymbol{\lambda_j} = (\lambda_j(\theta))_{j=1}^J$ *and* $\boldsymbol{\mu}(\theta) = (\mu_l(\theta))_{l=1}^k$ *are the solutions for*

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^J; \boldsymbol{\mu} \in \mathbb{R}^k} \max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + \boldsymbol{\lambda}^\intercal \mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^\intercal \mathbf{h}(\mathbf{x}). \tag{23}$$

Proof. Since the duality gap is zero,

$$V(\theta) = \min_{\boldsymbol{\lambda} \in \mathbb{R}_+^J; \boldsymbol{\mu} \in \mathbb{R}^k} \max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}, \theta) + \boldsymbol{\lambda}^\intercal \mathbf{g}(\mathbf{x}, \theta) + \boldsymbol{\mu}^\intercal \mathbf{h}(\mathbf{x}, \theta).$$

Applying Theorem 3.4.1 twice, and observe that the solutions of (22) are the same as optimal solutions of $\mathbf{x}$ in (23). Together, we have the result. ■

Example 3.4.1 (Incentive Compatibility Revisited). Recall that in Proposition 2.4.1, we characterized the incentive compatible direct mechanisms. In fact, Theorem 3.4.1 simplifies the proof for necessity. Recall that in the proof of Proposition 2.4.1, we defined

$$\Pi(v) := p(v)v - t(v),$$

which, by incentive compatibility, implies that

$$\Pi(v) = \max_{v' \in [0,1]} \pi(v', v),$$

where $\pi(v', v) := p(v')v - t(v')$. Clearly $\pi(v', \cdot)$ is differentiable for all $v' \in [0,1]$ and the derivative is simply $\pi_2(v', v) = p(v')$, which is itself Lebesgue integrable since $p(v') \in [0,1]$ for all $v' \in [0,1]$. Therefore, by Theorem 3.4.1,

$$\Pi(v) = \Pi(0) + \int_0^v p(z)dz.$$

Furthermore, since $\pi(v', \cdot)$ is an affine function for all $v' \in [0,1]$, Π is convex by Proposition 3.2.3. It then follows from Lemma 3.2.1 that $\Pi' = p$ is increasing.

Example 3.4.2 (Slutsky's Equation). Recall that in Proposition 4.2.2, we derived the Slutsky's identity:

$$\mathbf{x}^M(\mathbf{p}, e(\mathbf{p}, \bar{u})) = \mathbf{x}^H(\mathbf{p}, \bar{u}).$$

Notice that $e(\mathbf{p}, \bar{u})$ is defined to be the optimal value of the expenditure minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n_+} \mathbf{p}^\intercal\mathbf{x} \text{ s.t. } u(\mathbf{x}) = \bar{u}.$$

By Corollary 3.4.1, we have:

$$\nabla e(\cdot, \bar{u}) = \mathbf{x}^H(\mathbf{p}, \bar{u}),$$

for all $\mathbf{p} \in \mathbb{R}^n_{++}$ and any $\bar{u} \in \mathbb{R}$. Now differentiate Slutsky's identity with respect to p_i and apply the chain rule, let $\mathbf{x}^* := \mathbf{x}^H(\mathbf{p}, \bar{u})$, we have

$$\frac{\partial x_i^M}{\partial p_i} = \frac{\partial x_i^H}{\partial p_i} - \frac{\partial x_i^M}{\partial m} x_i^* \tag{24}$$

(24) is the well-known *Slutsky's equation*, which draws the connection between change in Marshallian demand and change in Hicksian demand as a commodity's price varies. The left hand side of (24) is the change of Marshallian demand, the first term on the right hand side of (24) is called the *substutition effect*, which, as will be shown later, is always negative; and the second term on the right hand side of (24) is called the *income effect*, which might be positive or negative. This is then the substitution effect-income effect decomposition that you learned in undergraduate microeconomics!

3.4.2 Implicit Function Theorem

While the envelope theorem establishes a characterization of derivatives of the *optimal value*, it cannot provide any further understandings about the derivatives of the *optimal choice*. To address this problem, we need the *implicit function theorem*. Again, let us motivate by the simple example

$$\max_{x\in[a,b]} f(x, \theta),$$

for $\theta \in [0, 1]$. Assume that the Hessian matrix of f exists and each component is continuous. Also assume that f is strictly concave and the (unique) optimal solution $x^*(\theta)$ is in (a, b)

and that x^* is differentiable. Thus, by first order condition,

$$f_1(x^*(\theta), \theta) = 0.$$

Since f_1 is differentiable in θ, by chain rule

$$f_{11}(x^*(\theta), \theta)x^{*'}(\theta) + f_{12}(x^*(\theta), \theta) = 0.$$

Since f is strictly concave $f_{11} < 0$ and therefore,

$$x^{*'}(\theta) = -\frac{f_{12}(x^*(\theta), \theta)}{f_{11}(x^*(\theta, \theta))}.$$

As such, we can examine the sign of $x^{*'}$ by simply seeing the sign of f_{12}.

We will now introduce a more generalized result, which and be applied when a problem involves solving a system with n unknowns and n equations, together with m parameters.

Theorem 3.4.2 (Implicit Function Theorem)**.** *For each $i \in \{1, \dots n\}$, $k \in \mathbb{N} \cup \{\infty\}$, let $f_i : \mathbb{R}^{n+m} \to \mathbb{R}$ be a function in $C^k(\mathbb{R}^{n+m})$ and let $\mathbf{f} := (f_i)_{i=1}^n$. Suppose that $f(\mathbf{x}^0, \mathbf{y}^0) = \mathbf{z}^0$ for some $\mathbf{x}^0, \mathbf{z}^0 \in \mathbb{R}^n$ and some $\mathbf{y}^0 \in \mathbb{R}^m$. Suppose also that the Jacobian of $\mathbf{f}$ with respect to $\mathbf{x}$, at $(\mathbf{x}^0, \mathbf{y}^0)$, is invertible. Then there exists an open set $U \subset \mathbb{R}^m$ such that $\mathbf{y}^0 \in U$ and a unique function $\mathbf{g} : U \to \mathbb{R}^n$ such that*

$$\mathbf{g}(\mathbf{y}^0) = \mathbf{x}^0$$

and

$$\mathbf{f}(\mathbf{g}(\mathbf{y}), \mathbf{y}) = \mathbf{z}^0,$$

for all $\mathbf{y} \in U$. Moreover, the partial derivatives of $\mathbf{g}$ on U is given by

$$J_{\mathbf{g}}(\mathbf{y}) = -(J_{\mathbf{f},\mathbf{x}}(\mathbf{g}(\mathbf{y}), \mathbf{y}))^{-1} J_{\mathbf{f},\mathbf{y}}(\mathbf{g}(\mathbf{y}), \mathbf{y}),$$

for all $\mathbf{y} \in U$, where

$$\begin{aligned} J_{\mathbf{g}}(\mathbf{y}) &:= \left(\frac{\partial g_i}{\partial y_j}(\mathbf{y})\right)_{i\in\{1,\dots,n\}, j\in\{1,\dots,m\}} \\ J_{\mathbf{f},\mathbf{x}}(\mathbf{x}, \mathbf{y})) &:= \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}, \mathbf{y})\right)_{i\in\{1,\dots,n\}, j\in\{1,\dots,n\}} \\ J_{\mathbf{f},\mathbf{y}}(\mathbf{x}, \mathbf{y}) &:= \left(\frac{\partial f_i}{\partial y_j}(\mathbf{x}, \mathbf{y})\right)_{i\in\{1,\dots,n\}, j\in\{1,\dots,m\}} \end{aligned}$$

Theorem 3.4.2 is in fact intuitive to understand. The theorem states that if a system of implicit equation $f(\mathbf{x}, \mathbf{y}) = \mathbf{z}$, with $\mathbf{x}$ being the endogenous variables and $\mathbf{y}$ being exogenous variables, has an explicit solution—$\mathbf{x}$ can be expressed as a function of $\mathbf{y}$ around a neighborhood, provided that there exists a particular solution given some particular exogenous variables (i.e. $f(\mathbf{x}^0, \mathbf{y}^0) = \mathbf{z}^0$) and the derivative is not zero at this particular solution. Moreover, the solution $\mathbf{g}(\mathbf{y})$ has the same degree of smoothness as the system itself. As such, the chain rule can be applied and the derivatives of the solutions $\mathbf{g}$ with respect to the exogenous variables $\mathbf{y}$ can be derived.

We conclude this section by a simple example

Example 3.4.3 (Hicksian Demand Revisited). Recall that we have derived Slutsky's equation

$$\frac{\partial x_i^M}{\partial p_i} = \frac{\partial x_i^H}{\partial p_i} - \frac{\partial x_i^M}{\partial m} x_i^*.$$

To complete the discussion of substitution effect-income effect decomposition, we need to examine the sign of $\partial x_i^H / \partial p_i$ for all $i \in \{1, \ldots, n\}$. Assume that u is in C^2 and is strictly quasi-concave. Then the first order condition is sufficient for the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbf{p}^\intercal \mathbf{x} \text{ s.t. } u(\mathbf{x}) = \bar{u}.$$

That is,

$$\begin{aligned} \mu(\mathbf{p}) \nabla u(\mathbf{x}^H(\mathbf{p})) &= \mathbf{p} \\ u(\mathbf{x}^H(\mathbf{p})) &= \bar{u} \end{aligned}$$

Rewrite the system as:

$$\begin{aligned} \mu \nabla u(\mathbf{x}) - \mathbf{p} &= \mathbf{0}_{n \times 1} \\ u(\mathbf{x}) - \bar{u} &= 0 \end{aligned} \tag{25}$$

Clearly there exists $\mathbf{x} = \mathbf{x}^H(\mathbf{p}) \in \mathbb{R}^n_+$ and $\mu \in \mathbb{R}$ such that the system (25) holds. Moreover, notice that the Jacobian matrix of the system with respect to $\mathbf{x}, \mu$ is given by

$$J(\mathbf{x}, \mu) := \begin{pmatrix} 0 & \nabla u(\mathbf{x})^\intercal \\ \nabla u(\mathbf{x}) & \mu H_u(\mathbf{x}) \end{pmatrix},$$

which is exactly the negative of the Borden Hessian of the minimization problem. Since u is quasi-concave,

$$\det \begin{pmatrix} 0 & \nabla u(\mathbf{x})^\intercal \\ \nabla u(\mathbf{x}) & \mu H_u(\mathbf{x}) \end{pmatrix} > 0$$

and therefore $J(\mathbf{x}, \mu)$ is invertible around $\mathbf{x}^H(\mathbf{p}), \mu(\mathbf{p})$. By the implicit function theorem, since u is twice differentiable, $\mathbf{x}^H$ is differentiable and thus for any $i \in \{1, \ldots, n\}$, after differentiating with respect to p_i in (25),

$$\begin{pmatrix} 0 & \nabla u(\mathbf{x})^\intercal \\ \nabla u(\mathbf{x}) & \mu H_u(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \frac{\partial \mu^H}{\partial p_i} \\ \frac{\partial x_1^H}{\partial p_i} \\ \vdots \\ \frac{\partial x_n^H}{\partial p_i} \end{pmatrix} = \mathbf{e}_i, \tag{26}$$

where $\mathbf{e}_{i+1} \in \mathbb{R}^{n+1}$ is the $i+1$-th standard basis that has all the components but the $i+1$-th zero and the $i+1$-th component is 1. Therefore, solving the system (26) gives

$$\frac{\partial x_i^H}{\partial p_i} = \frac{\det(J_i)}{\det(J)},$$

where J_i a matrix the is obtained by replacing the $i+1$-th column of J with $\mathbf{e}_{i+1}$. It is straightforward to show that $\det(J_i) \leq 0$. Together, we have

$$\frac{\partial x^H}{\partial p_i} < 0.$$

Therefore, the substitution effect must be negative.

3.4.3 Monotone Comparative Statics

Although the implicit function is useful, it requires differentiability and even further smoothness conditions on the primitives, which are far from necessary for economic analyses. However, even without differentiability, we can still obtain some unambiguous comparative statics if the problem has some special structures. This section will introduce the simplest form of this structure. Before stating the main result, we need to introduce some mathematical concepts in *lattice theory*.

To start off, let X be a set and $\geq$ be a transitive, reflexive and antisymmetric binary relation on X[23]. As such $(X, \geq)$ is a partially ordered set. For any $x, y \in X$, define $x \vee y$ (join) as the least upper bound of $\{x, y\}$, namely, $x \vee y \geq x$, $x \vee y \geq y$ and for any $z \geq x$ and $z \geq y$, $x \vee y \leq z$. Also, define $x \wedge y$ (meet) as the greast lower bound of x an y. i.e. $x \wedge y \geq x$, $x \wedge y \geq y$ and for any z such that $z \leq x$, $z \leq y$, $z \leq x \wedge y$. We can then define lattice as in the following definition.

Definition 3.4.1.

1. A partial ordered set $(X, \geq)$ is a lattice if for any $x, y \in X$, $x \wedge y$ and $x \vee y$ exists.

2. A subset $S \subset X$ is a sub-lattice of X if X is a lattice and if for any $x, y \in S$, $x \vee y \in S$ and $x \wedge y \in S$.

3. A sub-lattice S of X is complete if for any $S' \subset S$, $S' \neq \emptyset$, $\inf(S')$ and $\sup(S')$ exist and are in S.

Hereafter, unless it would be potentially confusing, we will suppress the notations for orderings and only say that X is a lattice and use $\geq$ to denote orders.

Since the main goal of this section is to examine the monotonicity of solutions as functions of parameters, and since the parameters sometimes involve sets (e.g. feasible sets) and the solutions are sometimes correspondences rather that functions, it is also necessary to have a proper order between sets. Let $(X, \geq)$ be a lattice, define the *strong set order*, $\geq_S$ as follows: For any $Y, Z \subset X$, $Z \leq_S Y$ if and only if for any $z \in Z$, $y \in Y$, $z \vee y \in Y$ and $z \wedge y \in Z$.

In many of the applications, the lattice being considered would be $\mathbb{R}^n$, $n > 1$. In such cases, the natural partial order would be component-wise comparisons. That is, $\mathbf{x} \geq \mathbf{y}$ if and only if $x_i \geq y_i$ for all $i \in \{1, \ldots, n\}$; $\mathbf{x} \wedge \mathbf{y} = (\min\{x_i, y_i\})_{i=1}^n$ and $\mathbf{x} \vee \mathbf{y} = (\max\{x_i, y_i\})_{i=1}^n$.

With the definitions above, as a side note, there is in fact another powerful fixed point theorem in lattice theory, which we record below:

Theorem 3.4.3 (Tarski's Fixed Point Theorem)**.** *Let $(X, \geq)$ be a lattice and $f : X \to X$ be an increasing function. That is, for any $x, y \in X$ such that $x \geq y$, $f(x) \geq f(y)$. Then f has a fixed point. Furthermore, the set of fixed points of f is a complete sublattice of X.*

[23] $\geq$ is transitive if for any $x, y, z \in X$, $x \geq y$ and $y \geq z$ implies $x \geq z$. $\geq$ is reflexive if for all $x \in X$, $x \leq x$. $\geq$ is antisymmetric if for all $x, y \in X$, $x \geq y$ and $y \geq x$ implies $x = y$.

The next definition is crucial to the results in this section.

Definition 3.4.2. Let X be a lattice and Θ be a partially ordered set.

1. A function $f : X \times \Theta \to \mathbb{R}$ is *quasi-supermodular* in x if for any $x, x' \in X$,

$$f(x, \theta) \geq (>) f(x \wedge x', \theta) \Rightarrow f(x \vee x', \theta) \geq (>) f(x', \theta), \forall \theta \in \Theta$$

2. A function $f : X \times \Theta \to \mathbb{R}$ is said to have *single-crossing property* in (x, θ) if for any $x' > x$ and $\theta' > \theta$,

$$f(x', \theta) \geq (>) f(x, \theta) \Rightarrow f(x', \theta') \geq (>) f(x, \theta').$$

As a remark, notice that quasi-supermodularity is trivially satisfied if X is totally ordered and thus only the single-crossing property will be relevant in one-dimensional problems. In addition, as it can be seen from the definition of single-crossing, such condition is conveying some ideas of *increasing marginal gains*. In fact, this is a strict generalization of the *increasing difference* property. If f is differentiable, there is an intuitive sufficient condition for quasi-supermodularity and single-crossing.

Proposition 3.4.1. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be twice continuously differentiable on an open set. Then f has single-crossing property in (x, θ) if $\frac{\partial^2 f}{\partial x_i \partial \theta_j} \geq 0$ for all $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$; and f is supermodular in x if $\frac{\partial^2 f}{\partial x_i x_j} \geq 0$ for all $i, j \in \{1, \ldots, n\}$, $i \neq j$.*

That is, single-crossing property and supermodularity is equivalent to the familiar *complementarity* conditions. We are now ready for the main result.

Theorem 3.4.4 (Monotone Comparative Statics (Milgrom & Shannon, 1994))**.** *Let X be a lattice, $S \subset X$ be any subset of X, and Θ be a partially ordered set. Let $f : X \times \Theta \to \mathbb{R}$ be a function on $X \times \Theta$. Then* $\operatorname{argmax}_{x \in S} f(x, \theta)$ *is monotone non-decreasing in (θ, S) if and only if f is quasi-supermodular in x and has single-crossing property in (x, θ), where order between sets is the strong set order.*

By Theorem 3.4.4, we can then conclude that the solution of

$$\max_{x \in S} f(x, \theta)$$

must be increasing the parameter θ whenever f is quasi-supermodular in x and single crossing in (x, θ) even if we do not assume any differentiability and continuity of the function f.

Example 3.4.4 (Bellman Equation Revisited). Recall that in Chapter 1, we saw that the Bellman equation

$$V(x) = \max_{y \in \Gamma(x)} F(x, y) + \beta V(y)$$

can be used to characterize the value of the sequence problem

$$\max_{\{x_t\}_{t=0}^{\infty} \subset X} \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}), \text{ s.t. } x_t \in \Gamma(x_{t-1}), \forall t \in \mathbb{N}.$$

Theorem 3.4.4 can then be used to examine the *policy function* $g : X \to X$. That is:

$$g(x) \in \operatorname*{argmax}_{y \in \Gamma(x)} F(x, y) + \beta V(y).$$

Since V is a fixed point of a functional, it is not obvious that V will be twice differentiable everywhere and thus the implicit function theorem is not very powerful here. However, by Theorem 3.4.4, if we further assume that F is single-crossing in (x, y) and that Γ is increasing in set inclusion order, the objective function of the problem

$$\max_{y \in \Gamma(x)} F(x, y) + \beta V(y)$$

is then single-crossing in (x, y). Thus, we can conclude that $g(x)$ must be increasing in x, for any selection g. You will see more about this in the second half of the course.

3.5 Exercises

1. Prove Proposition 3.1.7.

2. Consider a consumer's problem in an closed economy. Let $\mathbf{e} \in \mathbb{R}^n_+$ be the consumer's endowment. Then given any price vector $\mathbf{p} \in \mathbf{R}^n_+$, the consumer's problem is

$$\max_{\mathbf{x} \in \mathbb{R}^n_+} u(\mathbf{x}) \text{ s.t. } \mathbf{p}^\intercal \mathbf{x} \leq \mathbf{p}^\intercal \mathbf{e}.$$

   Let $\mathbf{x}^*(\mathbf{p})$ denote the solution given $\mathbf{p}$. Show that $\mathbf{x}$ is homogeneous of degree zero.

3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function for some $n \in \mathbb{N}$. Show that for each $\mathbf{x} \in \mathbb{R}^n$, there exists some $m \in \mathbb{R}^n$ such that

$$f(\mathbf{x}) \geq f(\mathbf{y}) + m^\intercal(\mathbf{x} - \mathbf{y}), \forall \mathbf{y} \in \mathbb{R}^n.$$

That is, show that for any $\mathbf{x} \in \mathbb{R}^n$, subgradient of f at $\mathbf{x}$ exists. (Hint: Use the supporting hyperplane theorem.)

4. Show that a function $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-concave if and only if it's upper contours are convex. That is.

$$\{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) \geq r\}$$

is convex for all $r \in \mathbb{R}$.

5. Show that when $u : \mathbb{R}^n_+ \to \mathbb{R}$ is strictly increasing, continuous and strictly quasi-concave. That is: for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$, $\mathbf{x} \neq \mathbf{y}$, $\lambda \in (0, 1)$.

$$u(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) > \min\{u(\mathbf{x}, \mathbf{y})\}.$$

The the problem

$$\max_{\mathbf{x} \in \mathbb{R}^n_+} u(\mathbf{x}) \text{ s.t. } \mathbf{p}^\intercal\mathbf{x} \leq m$$

has a unique solution for any $\mathbf{p} \in \mathbb{R}^n_{++}$, any $m > 0$.

6. Let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function. Define

$$\text{epi}(f) := \{(x, y) \in \mathbb{R}^2 | y \geq f(x)\}$$

as the *epigraph* of f.

   (a) Show that f is convex if and only if epi(f) is a convex set.

   (b) Let $E := \text{co}(\text{epi}(f))$ and let $g(x) := \inf\{y|(x, y) \in E\}$. g is called the *convexification* of f. Show that

$$g(x) = \sup\{\alpha + \beta x | (\alpha, \beta) \in \Gamma\},$$

   where $\Gamma := \{(\alpha, \beta) \in \mathbb{R}^2 | \alpha + \beta y \leq f(y), \forall y \in \mathbb{R}\}$. That is, the convexification of f is precisely the pointwise supremum over the family of affine functions that are majorized by f.

7. Consider the firm's profit maximization problem that is decomposed as follows: Fix any $y \geq 0$. The firm first solves a *cost minimization* problem

$$\min_{L,K \geq 0} wL + rL \text{ s.t. } F(L,K) = y$$

to minimize total cost of production to produce y units. Use first order approach, assume whatever differentiability and concavity you need, to give a characterization of the firm's cost-minimizing solutions. What are the relevant first order conditions? Can you give an interpretation of the optimal Lagrange multiplier? (Hint: Use the envelope theorem)

Now let $c(y)$ denote the value of the solution. What is the firm's *profit maximizing* problem? What can you say about the solution? Use this to conclude that the amount of labor and capital at optimum is the same as in Example 3.1.1.

4 Introduction to Probability Theory

Uncertainty and incomplete information are two of the most popular topics in economics. Probability theory is a powerful tool to model uncertainty and incomplete information. Furthermore, many economic analyses are often combined with econometrics, which is based on statistics and probability theory. In this chapter, we will introduce some of the most important concepts in probability theory, which is based on *measure theory*. We will first introduce abstractly some measure-theoretical concepts that generalizes some of the results in Chapter 2. Then we will properly introduce the notion of *random variables*, *expectations*, *density functions* using measure-theoretic language. Finally, we will briefly introduce a topological space for probability measures and characterize the convergence in distribution.

4.1 General Measure Spaces

4.1.1 Measurable Space and Measure

Recall that in Chapter 2, we introduced the notion of σ-algebra. Let Ω be a nonempty set. We say that $\mathcal{F}$ is a σ-algebra on Ω if:

1. $\Omega \in \mathcal{F}$.

2. For any $A \in \mathcal{F}$, $A^C \in \mathcal{F}$.

3. For any $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$, $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Given a nonempty set Ω and a σ-algebra $\mathcal{F}$ on Ω, we call $(\Omega, \mathcal{F})$ a *measurable space*. For instance, as shown in chapter 2, $(\mathbb{R}, \mathcal{M})$ is a measurable space, where $\mathcal{M}$ is the collection of Lebesgue measurable sets. Also, $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measurable space, where $\mathcal{B}(\mathbb{R})$ is the Borel algebra on $\mathbb{R}$. Recall also that we defined the Borel algebra on $\mathbb{R}$ to be the smallest σ-algebra that contains all the open sets in $\mathbb{R}$, we can generalize this definition as well. Specifically, let $(X, \mathcal{T})$ be a topological space, the *Borel algebra* on X is the smallest (why does this make sense?) σ-algebra that contains $\mathcal{T}$, which we will denote by $\mathcal{B}(X, \mathcal{T})$, or simply $\mathcal{B}(X)$ when there are no confusions.

Given a measurable space, we can then abstractly define a *measure* on this space.

Definition 4.1.1. Let $(\Omega, \mathcal{F})$ be a measurable space, $\mu : \mathcal{F} \to \mathbb{R}$ is called a *measure* if:

1. $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$.

2. For any countable disjoint collection $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$,

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

If, furthermore, $\mu(\Omega) = 1$, we then say μ is a *probability measure*. Finally, given a (probability) measure μ on the measurable space $(\Omega, \mathcal{F})$, we say that $(\Omega, \mathcal{F}, \mu)$ is a (probability) *measure space*. Also, we denote $\Delta(\Omega, \mathcal{F})$ by the collection of probability distributions on the measurable space $(\Omega, \mathcal{F})$.

Recall that we defined the collection of Lebesgue measurable sets $\mathcal{M}$ in a way such that the outer measure λ^* is countably additive. Now we work reversely—starting with a σ-algebra $\mathcal{F}$, we define a measure so that it is nonnegative-valued, assigns measure zero to the empty set and being countably additive. With these axioms, many of the properties of the Lebesgue measure still holds. Specifically,

Proposition 4.1.1. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then*

1. *μ is monotone. That is, for any $A, B \in \mathcal{F}$ with $A \subseteq B$, $\mu(A) \leq \mu(B)$.*

2. *μ is continuous. That is, for any $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ and $\{B_n\}_{n=1}^{\infty}$ with $A_n \subseteq A_{n+1}$ and $B_{n+1} \subseteq B_n$, $\mu(B_1) < \infty$,*

$$\lim_{n\to\infty} \mu(A_n) = \mu\left(\bigcup_{n=1}^{\infty} A_n\right); \quad \lim_{n\to\infty} \mu(B_n) = \mu\left(\bigcap_{n=1}^{\infty} B_n\right).$$

3. *μ is countably subadditive. That is, for any $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ (need not to be disjoint),*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n).$$

4.1.2 Cumulative Distribution Function

In many economic problems that involves probability theory, we are particularly interested in the probability measures on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Under this measurable space, there is a very useful representation of probability measures, also known as *cumulative distributions functions*, (or CDF for short). To formally introduce this representation, we first need the following well-known theorem.

Definition 4.1.2. Let Ω be a nonempty set, and let $\mathcal{A}$ be a collection of subsets of Ω. We say that $\mathcal{A}$ is a π-system if for any $A, B \in \mathcal{A}$, $A \cap B \in \mathcal{A}$. On the other hand, we say that $\mathcal{A}$ is a λ-system if

1. $\Omega \in \mathcal{A}$.

2. For any $A, B \in \mathcal{A}$ with $A \subseteq B$, $B \backslash A \in \mathcal{A}$.

3. For any $\{A_n\}_{n=1}^{\infty} \subset \mathcal{A}$ with $A_n \subseteq A_{n+1}$, $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Let $\mathcal{A}$ be a collection of subsets of a nonempty set Ω, we say that $\sigma(\mathcal{A})$ is the σ-algebra *generated* by $\mathcal{A}$ if it is the smallest sigma algebra that contains $\mathcal{A}$. For instance, consider the collection of intervals of form $[a, b)$, $(a, b]$, $[a, b]$ or (a, b), with $b \geq a$, $a, b \in \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$, denoted by $\mathcal{I}$. First notice that $\mathcal{I}$ is a π-system. Moreover, since all the open sets in $\mathbb{R}$ can be written as disjoint unions of open intervals, the σ-algebra generated by $\mathcal{I}$ contains all the open sets, and therefore, since $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra that contains all the open sets, $\mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{I})$. We can now introduce the theorem and the representation.

Theorem 4.1.1 ($\pi - \lambda$ Theorem)**.** *Let Ω be a nonempty set, $\mathcal{P}$ be a π-system and $\mathcal{L}$ be a λ-system such that $\mathcal{P} \subset \mathcal{L}$. Then $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

Definition 4.1.3. Let $F : \mathbb{R} \to [0,1]$ be a function on $\mathbb{R}$, F is said to be a *cumulative distribution function* (CDF) if:

1. F is increasing.
2. F is right-continuous. That is, $\lim_{\delta \downarrow 0} F(x+\delta) = F(x)$ for all $x \in \mathbb{R}$.
3. $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to\infty} F(x) = 0$.

We will let $\mathcal{G}$ denote the collection of all CDF henceforth.

We can now show that any probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ can be represented by a CDF.

Proposition 4.1.2. *There exists a bijection between $\Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\mathcal{G}$.*

Proof. For any $\mu \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, let $F_\mu(x) := \mu((-\infty, x]))$. We now show that the map

$$\mu \mapsto F_\mu$$

is a bijection. First notice that for any $\mu \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, F_μ is a CDF. Indeed, by monotonicity of μ, F is increasing. Furthermore, $\lim_{x\to\infty} F_\mu(x) = \mu(\mathbb{R}) = 1$, $\lim_{x\to-\infty} F_\mu(x) = \mu(\emptyset) = 0$ since μ is a probability measure. Finally, for any $x \in \mathbb{R}$ and for any sequence $\{x_n\}$ such that $x_{n+1} \leq x_n$ and $\{x_n\} \downarrow x$,

$$\lim_{n\to\infty} F_\mu(x_n) = \lim_{n\to\infty} \mu((-\infty, x_n]) = \mu\left(\bigcap_{n=1}^{\infty} (-\infty, x_n]\right) = \mu((-\infty, x]) = F_\mu(x)$$

by continuity of μ. As such, F_μ is indeed a CDF.

We now show that the map is onto. Indeed, for any CDF $F \in \mathcal{G}$, define

$$F^{-1}(y) := \sup\{x \in [0,1] | F(x) < y\}$$

and let $\mu(A) := \underline{\lambda}(\{y \in [0,1] | F^{-1}(y) \in A\})$ for all $A \in \mathcal{B}(\mathbb{R})$, where $\underline{\lambda}(B) = \lambda(B \cap [0,1])$ for all $B \in \mathcal{B}(\mathbb{R})$ is the Lebesgue measure on $[0,1]$. Then since $\underline{\lambda}$ is a measure and F only takes

value on $[0,1]$, μ is indeed a probability measure. Furthermore, notice that for any $x \in \mathbb{R}$ and any $y \in [0,1]$, if $y \leq F(x)$, then $x \notin \{x \in \mathbb{R} | F(x) < y\}$ and therefore $x \geq F^{-1}(y)$ since F is increasing. On the other hand, if $y > F(x)$, since F is right-continuous, there exists $\delta > 0$ such that $y > F(x+\delta)$ and thus $F^{-1}(y) \geq x + \delta > x$. Together, $F^{-1}(y) \leq x$ if and only if $y \leq F(x)$. Therefore, for any $x \in \mathbb{R}$,

$$\mu((-\infty, x]) = \underline{\lambda}(\{y \in [0,1] | F^{-1}(y) \leq x\}) = \underline{\lambda}(\{y \in [0,1] | y \leq F(x)\}) = \underline{\lambda}([0, F(x)]) = F(x).$$

Therefore, $F \equiv F_\mu$.

Finally, we show that the mapping in one-to-one. Suppose that there exists $\mu_1, \mu_2 \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\mu_1((-\infty, x]) = \mu_2(-\infty, x])$ for all $x \in \mathbb{R}$. Since μ_1 and μ_2 are countably additive μ_1 and μ_2 agree on the π-system $\mathcal{I}$. Let $\mathcal{L} := \{A \in \mathcal{B}(\mathbb{R}) | \mu_1(A) = \mu_2(A)\}$. Clearly $\Omega \in \mathcal{L}$ since μ_1 and μ_2 are probability measures. Also, by countable additivity of μ_1 and μ_2, for any $A, B \in \mathcal{L}$ such that $A \subseteq B$,

$$\mu_1(B \backslash A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \backslash A)$$

and hence $B \backslash A \in \mathcal{L}$. Finally, for any $\{A_n\}_{n=1}^\infty \subset \mathcal{L}$ with $A_n \subset A_{n+1}$ for all $n \in \mathbb{N}$, by continuity of μ_1 and μ_2,

$$\mu_1\left(\bigcup_{n=1}^\infty A_n\right) = \lim_{n\to\infty} \mu_1(A_n) = \lim_{n\to\infty} \mu_2(A_n) = \mu_2\left(\bigcup_{n=1}^\infty A_n\right)$$

and thus $\cup_{n=1}^\infty A_n \in \mathcal{L}$. Together, $\mathcal{L}$ is a λ-system and thus, by Theorem 4.1.1, since $\mathcal{I}$ is a π-system and $\mathcal{I} \subset \mathcal{L}$, $\mathcal{B}(\mathbb{R}) \subseteq \sigma(\mathcal{I}) \subseteq \mathcal{L} \subseteq \mathcal{B}(\mathbb{R})$. Thus, $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{B}(\mathbb{R}) = \mathcal{L}$. This completes the proof. ■

With Proposition 4.1.2, every probability distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ can be represented by a unique CDF $F \in \mathcal{G}$ and vice versa.

4.1.3 Measurable Functions and Integration

Recall that in chapter 2, we used simple approximation theorem to construct the Lebesgue integral. In fact, all the steps we used in chapter depend only on the properties that we abstracted aways to define general measures. As a result, integration can be defined on general measure spaces with any measure μ.

Definition 4.1.4. Let $(\Omega, \mathcal{F})$ and $(X, \mathcal{M})$ be two measurable spaces and $f : \Omega \to X$ be a function, f is $\mathcal{F}$-measurable if for any $B \in \mathcal{M}$,

$$f^{-1}(B) := \{\omega \in \Omega | f(\omega) \in B\} \in \mathcal{F}.$$

Henceforth, if there is no further confusions, we sometimes supress the σ-algebra when saying a function is measurable.

It can be easily verified that this is a generalized definition of Definition 2.1.3, and therefore, for a measure space $(\Omega, \mathcal{F}, \mu)$ and a function $f : \Omega \to \mathbb{R}$, simple approximation theorem is still valid, with the measure begin replaced by μ. As such, integral of a function can be constructed by the same procedure as in chapter 2. Formally, let $(X, \mathcal{M}, \mu)$ be a measure space, $f : X \to \mathbb{R}$ be a measurable function, the integral of f with respect to μ is denoted by

$$\int_X f d\mu, \text{ or } \int_X f(x)\mu(dx)$$

whenever $\int_X |f| d\mu < \infty$, which we will call μ-integrable.

By Proposition 4.1.2, for any probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, there exists a unique CDF $F_\mu \in \mathcal{G}$ such that $\mu((-\infty, x]) = F_\mu(x)$ for all $x \in \mathbb{R}$. As such, we sometimes use the associated CDF F_μ to denote the integral. That is, if $X \subseteq \mathbb{R}$ and $\mathcal{M}$ is a sigma algebra on X and μ is a probability measure on $(X, \mathcal{M})$,

$$\int_X f d\mu = \int_X f(x)\mu(dx) = \int_X f(x) F_\mu(dx).$$

This is a Lebesgue version of Riemann-Stieltjes integral.

For completeness, we summarize all the properties in chapter 2 here again, with a general measure μ. Fix a measure space $(X, \mathcal{M}, \mu)$.

Proposition 4.1.3. *Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be two real-value functions that are μ-integrable. Then:*

1. *(Monotonicity)*
$$f \geq g \Rightarrow \int_X f d\mu \geq \int_X g d\mu.$$
2. *(Linearity) For any $\alpha, \beta \in \mathbb{R}$, $\alpha f + \beta g$ is μ-integrable and*
$$\int_X (\alpha f + \beta g) d\lambda = \alpha \int_X f d\mu + \beta \int_X g d\mu.$$

3. *(Characterization of integrability) f is integrable if and only if for any $\varepsilon > 0$, there exists $\delta > 0$ such that*
$$\int_A f d\mu < \varepsilon$$
for any $A \in \mathcal{M}$ with $\mu < \delta$.

4. *(Countable additivity) For any disjoint collection $\{E_n\} \subset \mathcal{M}$,*
$$\int_{\cup_{n=1}^{\infty} E_n} f d\mu = \sum_{n=1}^{\infty} \int_{E_n} f d\mu.$$

5. *(Almost everywhere equivalence) For any $A \in \mathcal{M}$ with $\mu(A) = 0$,*
$$\int_{X \setminus A} f d\mu = \int_X f d\mu.$$

6. *(Converse of almost everywhere equivalence) If $f \geq 0$, then*
$$\int_X f d\mu = 0 \Rightarrow f \equiv 0 \text{ almost everywhere .}$$

Also, the L^p spaces can be defined with general measure μ. That is, we define:

$$L^p(X, \mathcal{M}, \mu) := \left\{ f : X \to \mathbb{R} \middle| \int_X |f|^p d\mu < \infty \right\}$$

and

$$||f||_p := \left(\int_X |f|^p d\mu \right)^{\frac{1}{p}}$$

is a norm on $L^p(X, \mathcal{M}, \mu)$, for any $p \in [1, \infty)$. It can be shown that $L^p(X, \mathcal{M}, \mu)$ is complete for any $p \in [1, \infty)$ and that

$$\int_X |fg| d\mu \leq ||f||_p \cdot ||g||_q$$

for any $f \in L^p(X, \mathcal{M}, \mu)$, $g \in L^q(X, \mathcal{M}, \mu)$, $p \in (1, \infty)$ and $\frac{1}{p} + \frac{1}{q} = 1$. That is, the Hödel inequality still holds.

Furthermore, all the convergence theorems in chapter 2 are still valid.

Lemma 4.1.1 (Fatou's Lemma)**.** *For any sequence of nonnegavie measurable functions $\{f_n\}$ on X such that $\{f_n\}$ converges pointwisely μ-almost everywhere for some $f : X \to \mathbb{R}$, f is μ-integrable and*
$$\int_X f d\mu \leq \liminf_{n \to \infty} \int_X f_n d\mu$$

Theorem 4.1.2 (Monotone Convergence Theorem)**.** *For any sequence of nonnegative measurable functions $\{f_n\}$ such that $0 \leq f_n \leq f_{n+1}$ for all $n \in \mathbb{N}$ and that $\{f_n\}$ converges poitnwisely μ-almost everywhere to some $f : X \to \mathbb{R}$, f is μ-integrable and*

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X f d\mu$$

Theorem 4.1.3 (Dominance Convergence Theorem)**.** *For any sequence of measurable functions $\{f_n\}$ that converges pointwisely μ-almost everywhere to some $f : X \to \mathbb{R}$, if there exists an μ-integrable function $g : X \to \mathbb{R}$ such that $|f_n| \leq |g|$ for all $n \in \mathbb{N}$, then f is μ-integrable and*

$$\lim_{n\to\infty} \int_X f_n d\mu = \int_X f d\mu.$$

Finally, the formula for integration by parts and Fubini's theorem are also valid.

Proposition 4.1.4. *Let $[a, b] \subset \mathbb{R}$ be an interval in $\mathbb{R}$ and let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ be a measure space with $\mu([a, b]) < \infty$. Suppose that $f : [a, b] \to \mathbb{R}$ is differentiable on X Lebesgue-almost everywhere. Then*

$$\int_{[a,b]} f(x)\mu(dx) = f(b)F_\mu(b) - f(a)F_\mu(a^-) - \int_a^b f'(x)F_\mu(x)dx,$$

where F_μ is defined as in Proposition 4.1.2 and $F_\mu(a^-) := \lim_{\delta\downarrow 0} F(a - \delta)$.

Proposition 4.1.5. *Let $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{F}, \nu)$ be two measure spaces. Let $\mu \times \nu$ denote the unique measure on the measurable space $(X \times Y, \mathcal{S})$ such that $\mu \times \nu(A \times B) = \mu(A)\nu(B)$ for all $A \in \mathcal{M}$, $B \in \mathcal{F}$, where $\mathcal{S}$ is the σ-algebra generated by $\{A \times B | A \in \mathcal{M}, B \in \mathcal{F}\}$. Let $f : X \times Y \to \mathbb{R}$ be a measurable function. Then if $f \geq 0$ or $\int_{X\times Y} |f| d\mu \times \nu < \infty$,*

$$\int_X \left(\int_Y f(x, y)\nu(dy) \right) \mu(dx) = \int_{X\times Y} f d\mu \times \nu = \int_Y \left(\int_X f(x, y)\mu(dx) \right) \nu(dy).$$

4.2 Random Variable and Expectation

4.2.1 Random Variable

In this section, we will use the notions of measurable space and measurable functions to define random variables and expectations formally. Fix a probability measure space $(\Omega, \mathcal{F}, \mathbb{P})$, we can think of this as an underlying probabilistic description of the world that summarizes all

the uncertainties. That is, each $\omega \in \Omega$ is a *state of world*, $E \in \mathcal{F}$ is an *event* and $\mathbb{P}(E)$ is the *probability* that event E occurs. For instance, for a (fair) coin toss. $\Omega = \{H, T\}$, $\mathcal{F} = \{\{\emptyset\}, \{H\}, \{T\}, \{H, T\}\}$ and $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2$. Since the state space Ω is abstract and we are often interested in uncertainties in actual *numbers* (e.g. stock price, productivity, valuations etc.), translating uncertainties from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to real numbers is essential. *Random variables* can achieve this purpose.

Definition 4.2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space. $X : \Omega \to \mathbb{R}$ is a *random variable* if for any $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) := \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{F}.$$

That is, a random variable is a measurable function that maps from Ω to $\mathbb{R}$. Furthermore, for any random variable X, for any (Borel) measurable function $f : \mathbb{R} \to \mathbb{R}$, $f(X) : \Omega \to \mathbb{R}$ is still a measurable function on Ω and thus is also a random variable.

For any random variable $X : \Omega \to \mathbb{R}$ and any σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, we say that X is *measurable with respect to* $\mathcal{G}$ if

$$X^{-1}(B) \in \mathcal{G}, \forall B \in \mathcal{B}(\mathbb{R}).$$

Conversely, given a random variable X, we can identify a σ-algebra on Ω that describes the informational content of this random variable. That is, we can let

$$\sigma(X) := \sigma(\{X^{-1}(B) | B \in \mathcal{B}(\mathbb{R})\})$$

Conceptually, the larger this σ-algebra is, the more information this random variable is conveying. For instance, if a random variable X is degenerate. That is, if $X(\omega) = x_0 \in \mathbb{R}$ for all $\omega \in \Omega$, then $\sigma(X) = \{\Omega, \emptyset\}$, which means that X contains no information.

Notice that since a random variable X is measurable, the measure $\mathbb{P} \circ X^{-1}$ is a probability measure on $\mathbb{R}$. We call this probability measure the *distribution* or *law* of the random variable X, denoted by $\mu_X \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Therefore, by Proposition 4.1.2, for any random variable X, there exists a unique CDF $F_X \in \mathcal{G}$ such that $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\})$ for all $x \in \mathbb{R}$. We say that F_X is a CDF of X.

Conversely, for any CDF $F \in \mathcal{G}$, if we let U be the random variable of *uniform distribution*. That is, let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and $\mathbb{P}$ be defined as $\mathbb{P}(A) := \lambda(A \cup [0,1])$. $U : \Omega \to [0,1]$ defined as $U(\omega) := \omega$ for all $\omega \in [0,1]$ is then a random variable with uniform distribution. Let $X := F^{-1}(U)$, where

$$F^{-1}(y) := \sup\{x \in \mathbb{R} | F(x) < y\}$$

is a measurable function. By the proof of Proposition 4.1.2, $X := F^{-1}(U)$ is then a random variable with a CDF F.

4.2.2 Expectation

With the tools above, we can now formally define *expectations* of a random variable X.

Definition 4.2.2. Let $(\Omega, \mathcal{F}, \mathbb{R})$ be a probability measure space and $X : \Omega \to \mathbb{R}$ be a random variable, the *expectation* of X, $\mathbb{E}[X]$, is defined by

$$\mathbb{E}[X] := \int_\Omega X d\mathbb{P},$$

provided that $\int_\Omega |X| d\mathbb{P} < \infty$.

An immediate observation is that for any random variable X,

$$\mathbb{E}[X] = \int_\Omega X d\mathbb{P} = \int_{\mathbb{R}} x \mu_X(dx) = \int_{\mathbb{R}} x F_X(dx),$$

notice that the second integral is simply an integration on $\mathbb{R}$, which is often easier to compute.

Moreover, expectations of functions of random variables can also be properly defines as:

$$\mathbb{E}[f(X)] := \int_\Omega f \circ X d\mathbb{P} = \int_{\mathbb{R}} f(x) \mu_X(dx).$$

Since expectations are simply integrals, all the properties in Proposition 4.1.3, the convergence theorems, and the Hölder inequality are true for expectation operators. For instance, it is linear. That is, for any random variables X, Y and any $\alpha, \beta \in \mathbb{R}$,

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y].$$

Also, it is monotone, that is, for any random variables X, Y such that $X \geq Y$, $\mathbb{P}$-almost every where

$$\mathbb{E}[X] \geq \mathbb{E}[Y].$$

In addition, if we define *variance* of a random variable X as

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2],$$

provided that $\mathbb{E}[X^2] < \infty$, and define *corvariance* between to random variables as

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])],$$

provided that $\mathbb{E}[X^2] < \infty$, $\mathbb{E}[Y^2] < \infty$. Notice that the variance is simply the L^2-norm and therefore, the by Hölder inequality with $p = q = 2$,

$$\mathrm{Cov}(X, Y) \leq \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)},$$

which is exactly the *Cauchy-Schwartz* inequality that you are familiar with.

Another useful inequality for expectation operator is the *Jensen's inequality.*

Proposition 4.2.1 (Jensen's Inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space, $X : \Omega \to \mathbb{R}$ be a random variable and $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Suppose that $\mathbb{E}[|f(X)|] < \infty$ and $\mathbb{E}[|X|] < \infty$. Then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Proof. Since f is a convex function, by Proposition 3.2.3, f is a pointwise supremum of a family of affine functions. That is, $f(x) = \sup_{L \in \mathcal{L}} L(x)$ for some collection of affine functions $\mathcal{L}$. Therefore,

$$\mathbb{E}[f(X)] = \mathbb{E}[\sup_{L \in \mathcal{L}} L(X)] \geq \sup_{L \in \mathcal{L}} \mathbb{E}[L(X)] = \sup_{L \in \mathcal{L}} L(\mathbb{E}[X]) = f(\mathbb{E}[X]),$$

where the second equality follows from linearity of the expectation operator. ■

4.2.3 Independence

In probability and statistics, one of the most important notions is called *independence*, this subsection introduces basic definition and implications of independence.

Definition 4.2.3. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a collection of sub σ-algebras $\{\mathcal{F}_i\}_{i=1}^n \subseteq \mathcal{F}$ is said to be independent if for any $\{A_i\}_{i=1}^n$ with $A_i \in \mathcal{F}_i$, for all $i \in \{1, \ldots, n\}$,

$$\mathbb{P}\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \mathbb{P}(A_i).$$

A collection of random variables $\{X_i\}_{i=1}^n$ is said to be independent if the collection of σ-algebras $\{\sigma(X_i)\}_{i=1}^n$ is independent.

Using the $\pi - \lambda$ theorem, it is easy to check independence of random variables.

Proposition 4.2.2. *Suppose that $\{\mathcal{A}_i\}_{i=1}^n$ is a collection of π systems such that for any $\{A_i\}_{i=1}^n$ with $A_i \in \mathcal{A}_i$, for all $i \in \{1, \ldots, n\}$,*

$$\mathbb{P}\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \mathbb{P}(A_i).$$

The the collection of σ-algebras $\{\sigma(\mathcal{A}_i)\}_{i=1}^n$ is independent.

Corollary 4.2.1. *A collection of random variables $\{X_i\}_{i=1}^n$ is independent if for any $\{x_i\}_{i=1}^n \subseteq \mathbb{R}$,*

$$\mathbb{P}(\{\omega \in \Omega | X_i(\omega) \leq x_i, \, \forall i \in \{1, \ldots, n\}\}) = \prod_{i=1}^n \mathbb{P}(\{\omega \in \Omega | X_i(\omega) \leq x_i\}).$$

Proof. For each $i \in \{1, \ldots, n\}$, let $\mathcal{A}_i$ be the collection of sets that take form of $\{\omega \in \Omega | X_i(\omega) \leq x_i\}$ for some $x_i \in \mathbb{R}$. Then $\mathcal{A}_i$ is a π system and $\sigma(\mathcal{A}_i) = \sigma(X_i)$. the result then follows from Proposition 4.2.2. ■

Corollary 4.2.1 is then the familiar definition of independent random variables learned in undergraduate probability. It says that it suffices to check if the joint CDF of random variables are multiplicatively separable in order to ensure independence.

Independent random variables have an important implication.

Lemma 4.2.1. *Let $\{X_i\}_{i=1}^n$ be a collection of random variables with $\mathbb{E}[|X_i|] < \infty$ for all $i \in \{1, \ldots, n\}$. Then*

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i].$$

Lemma 4.2.1 means that independence implies zero correlation, which is defined as $\mathrm{Cov}(X, Y) = 0$ for two random variables X and Y, as long as expectations are finite. An immediate consequence is that for a collection of independent random variable $\{X_i\}_{i=1}^n$,

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}\left(X_i\right).$$

4.3 Absolute Continuity and Conditional Expectation

The goal in this section is to introduce the *conditional expectation*. Conceptually, conditional expectation is the expectation of a random variable after given certain information about the underlying state of the world. For instance, if there are two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ and if one is able to observe the realization of Y. Then one's expectation about X should also incorporate the information from observing Y. We will provide a formal definition of such conditional expectation, denoted by $\mathbb{E}[X|Y]$. Before introducing this, we need some preliminaries.

4.3.1 Absolute Continuity and Density Function

Definition 4.3.1. Let $(X, \mathcal{M})$ be a measurable space and let μ and ν be two measures on $(X, \mathcal{M})$. We say that μ is *absolutely continuous* with respect to ν, denoted as $\mu \ll \nu$, if for any $E \in \mathcal{M}$, $\mu(E) = 0$ whenever $\nu(E) = 0$.

It is not difficult to verify that a probability measure $\mu \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is absolutely continuous with respect to the Lebesgue measure if and only if the associated CDF F_μ is absolutely continuous.

Absolute continuity has a crucial role in measure theory and probability theory, mostly due to the following theorem:

Theorem 4.3.1 (Radon-Nikodym Theorem)**.** *Let $(X, \mathcal{M})$ be a measurable space and let μ,ν be measures on $(X, \mathcal{M})$ with $\mu(X) < \infty$, $\nu(X) < \infty$ such that μ is absolutely continuous with respect to ν. Then there exists $f : X \to \mathbb{R}$ that is $\mathcal{M}$-measurable such that*

$$\mu(E) = \int_E f d\nu, \forall E \in \mathcal{M}. \tag{27}$$

Furthermore, if there exists $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ such that (27) holds, then $f \equiv g$ ν-almost everywhere.

From Theorem 4.3.1, if $\mu \ll \nu$, then there exists an essentially unique function f such that the measure μ can be represented by the integral of f with respect to ν. We call this function f the *Radon-Nikodym* derivative of μ with respect to ν, often denoted by $\frac{d\mu}{d\nu}$. As

such, (27) can be written in a more intuitive way:

$$\mu(E) = \int_E 1 d\mu = \int_E \frac{d\mu}{d\nu} d\nu, \forall E \in \mathcal{M}.$$

As a corollary, we our familiar definition of *density function*.

Corollary 4.3.1. *Let $\mu \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a probability measure that is absolutely continuous with respect to the Lebesgue measure. Then there exists an essentially unique $f : \mathbb{R} \to \mathbb{R}$ such that*

$$\mu(E) := \int_E f d\lambda, \forall E \in \mathcal{M}$$

Such function f is exactly the density function of the probability distribution μ, as the Lebesgue integral of the function on any measurable set E is exactly the probability of E. Notice that for any such probability measure μ,

$$F_\mu(x) = \int_{-\infty}^{x} f(z) dz, \forall x \in \mathbb{R}.$$

Together with the fundamental theorem of calculus in chapter 2, we then have $F' \equiv f$ Lebesgue-almost everywhere.

Notice that if μ is not absolute continuous, then the statement above is not true. As we saw in chapter 2, the cantor function is in $\mathcal{G}$ and therefore there exists a probability measure $\mu_C \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that (27) fails.

Although not all the probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are absolutely continuous with respect to the Lebesgue measure, all of the measures can be *decomposed* so that part of it is absolutely continuous.

Definition 4.3.2. Let $(X, \mathcal{M})$ be a measurable space and μ, ν be two measures on $(X, \mathcal{M})$. We say that μ and ν are *mutually singular*, denoted as $\mu \perp \nu$, if there exists $A, B \in \mathcal{M}$ with $A \cap B = \emptyset$ and $A \cup B = X$ such that

$$\mu(A) = \nu(B) = 0.$$

For instance, consider again the Cantor function, let μ_C denote the associated probability measure. Since the Cantor set $\mathbb{C}$ has Lebesgue measure zero and $\mu_C(\mathbb{C}) = 1$, μ_C and the Lebesgue measure on $[0, 1]$ are mutually singular.

Theorem 4.3.2 (Lebesgue Decomposition)**.** *Let $(X, \mathcal{M})$ be a measurable space and let μ, ν be measures on $(X, \mathcal{M})$ such that $\mu(X) < \infty$, $\nu(X) < \infty$. Then there exists two measures μ_a, μ_s on $(X, \mathcal{M})$ such that μ_a is absolutely continuous with respect to ν and μ_s, ν are mutually singular and that*

$$\mu = \mu_a + \mu_s.$$

Notice that from Theorem 4.3.2, since μ_s and ν are mutually singular, there exists $A \in \mathcal{M}$ such that $\mu_s(A) = 0$ and $\nu(X \backslash A) = 0$. Since $\mu_a \ll \nu$, we then have $\mu_a(X \backslash A) = 0$ as well. We say that the absolutely continuous part μ_a *charges* on the measurable set A and the singular part *charges* on $X \backslash A$. That is, Theorem 4.3.2 allows us to decompose a measure into two parts, one that is absolutely continuous with respect to the underlying measure ν and the other is singular with ν, and each of them charges on disjoint subsets of X. For instance, any probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ can be decomposed into two parts, one of which charges on a measurable subset $A \in \mathcal{B}(\mathbb{R})$ and the other charges on $X \backslash A$. The part that charges on A has a density function with respect to the Lebesgue measure. That is, for any $E \subseteq A$, $E \in \mathcal{B}(\mathbb{R})$,

$$\mu_a(E) = \int_E f_a d\lambda$$

for an essentially unique $f : A \to \mathbb{R}$. For the probability measure induced by the Cantor function, the absolute continuous part is a zero measure and charges on the set D that is removed from the interval $[0, 1]$ when constructing $\mathbb{C}$.

4.3.2 Conditional Expectation

With the results from the previous section, we are now able to formally define conditional expectations.

Definition 4.3.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}[|X|] < \infty$. Given any σ-algebra $\mathcal{H} \subseteq \mathcal{F}$, we say that a *version* of the *conditional expectation* given $\mathcal{H}$ is a random variable $Y : \Omega \to \mathbb{R}$ such that Y is $\mathcal{H}$-measurable and that

$$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P}, \, \forall A \in \mathcal{H}.$$

Proposition 4.3.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and let X be a random variable with $\mathbb{E}[|X|] < \infty$. Then for any σ-algebra $\mathcal{H} \subseteq \mathcal{F}$, conditional expectation of X given $\mathcal{H}$ exists and is essentially unique.*

Proof. Consider any random variable X on the probability measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}[|X|] < \infty$, if $X \geq 0$, let

$$\nu(A) := \int_A X d\mathbb{P}, \forall A \in \mathcal{H}.$$

It can be shown that, by the dominant convergence theorem and Proposition 4.1.3, $\nu \ll \mathbb{P}|_{\mathcal{H}}$. Therefore, by Theorem 4.3.1, there exists an essentially unique $Y : \Omega \to \mathbb{R}$ that is $\mathcal{H}$-measurable such that

$$\int_A X d\mathbb{P} = \nu(A) = \int_A Y d\mathbb{P}, \forall A \in \mathcal{H}.$$

As such, the condition expectation of X given $\mathcal{H}$ exists and is essentially unique (in the sense that it is unique except for a $\mathbb{P}$-measure zero set.) Now for any random variable X, write $X := X^+ - X^-$ and let Y^+, Y^- be versions of conditional expectations of X^+ and X^-, respectively and let $Y := Y_1 - Y_2$. Then clearly Y is $\mathcal{H}$-measurable and

$$\int_A X d\mathbb{P} = \int_A X^+ d\mathbb{P} - \int_A X^- d\mathbb{P} = \int_A Y^+ d\mathbb{P} - \int_A Y^- d\mathbb{P} = \int_A (Y^+ - Y^-) d\mathbb{P}, \forall A \in \mathcal{H}.$$

Together, conditional expectation for any random variable X with $\mathbb{E}[|X|] < \infty$ exists and is essentially unique. ■

Since conditional expectation of a random variable X given $\mathcal{H}$ is essentially unique, we often use $\mathbb{E}[X|\mathcal{H}]$ to denote this random variable. As such, for any Y, a version of conditional expectation of X given $\mathcal{H}$, $\mathbb{E}[X|\mathcal{H}] \equiv Y$ $\mathbb{P}$-almost everywhere. A more intuitive way to characterize conditional expectation is then:

$$\int_A (X - \mathbb{E}[X|\mathcal{H}]) d\mathbb{P} = 0, \forall A \in \mathcal{H}.$$

It is obvious from definition that $\mathbb{E}[X|\mathcal{F}] = X$ $\mathbb{P}$-almost everywhere. On the other hand, for any measurable subsets $A, B \in \mathcal{F}$,[24]

$$\mathbb{E}[\mathbf{1}_A|\sigma(B)] = \mathbf{1}_B \frac{\int_B \mathbf{1}_A d\mathbb{P}}{\mathbb{P}(B)}.$$

[24] With a convention that $0/0 = 0$.

Therefore, if we define $\mathbb{P}(A|B) := \mathbb{E}[\mathbf{1}_A|\sigma(B)](\omega)$ for some $\omega \in B$, then we have our familiar Bayes' rule

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Also, for a special case when $\mathcal{H} = \sigma(P)$, where $P = \{\Lambda\}_{j=1}^{\infty}$ is a countable partition of Ω, for any random variable X with $\mathbb{E}[|X|] < \infty$,

$$\mathbb{E}[X|\mathcal{H}] = \sum_{j=1}^{\infty} \mathbf{1}_{\Lambda_j} \frac{\int_{\Lambda_j} X d\mathbb{P}}{\mathbb{P}(\Lambda_j)},$$

which brings us back to our familiar Bayes' formula. Indeed, for any $\Lambda_j \in P$,

$$\int_{\Lambda_j} \sum_{j=1}^{\infty} \mathbf{1}_{\Lambda_j} \frac{\int_{\Lambda_j} X d\mathbb{P}}{\mathbb{P}(\Lambda_j)} d\mathbb{P} = \int_{\Lambda_j} X d\mathbb{P}.$$

Therefore, for any countable subset of P, denoted by Q,

$$\begin{aligned}
\int_{\cup\{\Lambda \in Q\}} \sum_{j=1}^{\infty} \mathbf{1}_{\Lambda_j} \frac{\int_{\Lambda_j} X d\mathbb{P}}{\mathbb{P}(\Lambda_j)} d\mathbb{P} &= \sum_{\Lambda \in Q} \int_{\Lambda} \sum_{j=1}^{\infty} \mathbf{1}_{\Lambda_j} \frac{\int_{\Lambda_j} X d\mathbb{P}}{\mathbb{P}(\Lambda_j)} d\mathbb{P} \\
&= \sum_{\Lambda \in Q} \int_{\Lambda} X d\mathbb{P} \\
&= \int_{\cup\{\Lambda \in Q\}} X d\mathbb{P}
\end{aligned}$$

by countable additivity of integration. The result then follows from the observation that the collection of sets that take form of countable union of subsets in P, together with Ω and $\emptyset$ is a σ-algebra.

For any random variables X, Y on a probability measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}[|X|] < \infty$, we can then define the conditional expectation of X given Y as:

$$\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)].$$

Conditional expectation operator, as the expectation operator, has some nice properties, we list some of them here.

Proposition 4.3.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space, X, Y be random variables with $\mathbb{E}[|X|] < \infty$, $\mathbb{E}[|Y|] < \infty$, and $\{X_n\}$ be a sequence of random variables such that $\{X_n\} \to X$ pointwisely $\mathbb{P}$-almost everywhere. Then for any σ-algebra $\mathcal{H} \subseteq \mathcal{F}$,*

1. *(Linearity) For any* $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}[\alpha X + \beta Y|\mathcal{H}] = \alpha\mathbb{E}[X|\mathcal{H}] + \beta\mathbb{E}[Y|\mathcal{H}]$.

2. *(Monotonicity)* $\mathbb{E}[X|\mathcal{H}] \geq \mathbb{E}[Y|\mathcal{H}]$ $\mathbb{P}$*-almost everywhere if* $X \geq Y$ $\mathbb{P}$*-almost everywhere.*

3. *(Jensen's Inequality) For any convex function* $f : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}[f(X)|\mathcal{H}] \geq f(\mathbb{E}[X|\mathcal{H}])$ $\mathbb{P}$*-almost everywhere.*

4. *(Fatou's Lemma) If* $X_n \geq 0$ *for all* $n \in \mathbb{N}$*, then*

$$\liminf_{n\to\infty} \mathbb{E}[X_n|\mathcal{H}] \geq \mathbb{E}[X|\mathcal{H}]$$

   $\mathbb{P}$*-almost everywhere.*

5. *(Monotone Convergence Theorem) If* $0 \leq X_n \leq X_{n+1}$ *for all* $n \in \mathbb{N}$*, then*

$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$$

   $\mathbb{P}$*-almost everywhere.*

6. *(Dominant Convergence Theorem) If* $|X_n| \leq |Y|$ $\mathbb{P}$*-almost everywhere, then*

$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$$

   $\mathbb{P}$*-almost everywhere.*

The following two properties are crucial for computing conditional expectations, one of them leads to the well-known formula of *law of iterative expectation.*

Proposition 4.3.3. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability measure space and* X *be a random variable with* $\mathbb{E}[|X|] < \infty$*. Then for any* σ*-algebra* $\mathcal{H} \subseteq \mathcal{F}$ *and any random variable* Y *that is* $\mathcal{H}$*-measurable,*

$$\mathbb{E}[XY|\mathcal{H}] = Y\mathbb{E}[X|\mathcal{H}]$$

$\mathbb{P}$*-almost everywhere.*

Proof. First consider any $B \in \mathcal{H}$, since

$$\int_A \mathbf{1}_B X d\mathbb{P} = \int_{A\cap B} X d\mathbb{P} = \int_{A\cap B} \mathbb{E}[X|\mathcal{H}] d\mathbb{P} = \int_A \mathbf{1}_B \mathbb{E}[X|\mathcal{H}] d\mathbb{P}$$

for all $A \in \mathcal{H}$. By essential uniqueness of conditional expectation, $\mathbb{E}[\mathbf{1}_B X|\mathcal{H}] = \mathbf{1}_B \mathbb{E}[X|\mathcal{H}]$ $\mathbb{P}$-almost everywhere. Thus, for any simple function $Y : \Omega \to \mathbb{R}$ that is $\mathcal{H}$-measurable. Since $Y = \sum_{k=1}^n y_k \mathbf{1}_{B_k}$ for some $\{B_k\}_{k=1}^n \subset \mathcal{H}$ and some $\{y_k\}_{k=1}^n \subset \mathbb{R}$, by linearity of conditional expectation,

$$\mathbb{E}[YX|\mathcal{H}] = Y\mathbb{E}[X|\mathcal{H}]$$

$\mathbb{P}$-almost everywhere. Now if $Y : \Omega \to \mathbb{R}$ is $\mathcal{H}$-measurable and $Y \geq 0$, by the simple approximation theorem, there exists a sequence of simple functions $\{Y_n\}$ such that $0 \leq Y_n \leq Y_{n+1}$ and $\{Y_n\} \to Y$ pointwisely $\mathbb{P}$-almost everywhere. By the monotone convergence theorem, for $\mathbb{P}$-almost all $\omega \in \Omega$,

$$\mathbb{E}[YX|\mathcal{H}](\omega) = \lim_{n\to\infty} \mathbb{E}[Y_n X|\mathcal{H}](\omega) = \lim_{n\to\infty} Y_n(\omega)\mathbb{E}[X|\mathcal{H}](\omega) = Y(\omega)\mathbb{E}[X|\mathcal{H}](\omega).$$

$\mathbb{P}$-almost everywhere. Finally, if Y is $\mathcal{H}$-measurable, write $Y = Y^+ - Y^-$. Then by linearity again,

$$\mathbb{E}[YX|\mathcal{H}] = \mathbb{E}[(Y^+ - Y^-)X|\mathcal{H}] = (Y^+ - Y^-)\mathbb{E}[X|\mathcal{H}] = Y\mathbb{E}[X|\mathcal{H}]$$

$\mathbb{P}$-almost everywhere. ■

Proposition 4.3.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and X be a random variable with $\mathbb{E}[|X|] < \infty$. Then for any σ-algebras $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{F}$,*

$$\mathbb{E}[\mathbb{E}[X|\mathcal{H}_1]|\mathcal{H}_2] = \mathbb{E}[\mathbb{E}[X|\mathcal{H}_2]|\mathcal{H}_1] = \mathbb{E}[X|\mathcal{H}_1].$$

$\mathbb{P}$-almost everywhere.

Proof. We first show that $\mathbb{E}[\mathbb{E}[X|\mathcal{H}_2]|\mathcal{H}_1] = \mathbb{E}[X|\mathcal{H}_1]$ $\mathbb{P}$-almost everywhere. Indeed, let $Y := \mathbb{E}[\mathbb{E}[X|\mathcal{H}_2]|\mathcal{H}_1]$. By definition, Y is $\mathcal{H}_1$-measurable. Furthermore, by definition, for any $A \in \mathcal{H}_2$,

$$\int_A \mathbb{E}[X|\mathcal{H}_2]d\mathbb{P} = \int_A X d\mathbb{P}.$$

On the other hand, for any $A \in \mathcal{H}_1$,

$$\int_A \mathbb{E}[X|\mathcal{H}_2]d\mathbb{P} = \int_A \mathbb{E}[\mathbb{E}[X|\mathcal{H}_2]|\mathcal{H}_1]d\mathbb{P}.$$

Together with the condition that $\mathcal{H}_1 \subseteq \mathcal{H}_2$, for any $A \in \mathcal{H}_1$,

$$\int_A X d\mathbb{P} = \int_A \mathbb{E}[X|\mathcal{H}_2]d\mathbb{P} = \int_A Y d\mathbb{P}.$$

and hence

$$\int_A Y d\mathbb{P} = \int_A X d\mathbb{P}, \forall A \in \mathcal{H}_1.$$

By essential uniqueness of conditional expectation,

$$\mathbb{E}[X|\mathcal{H}_1] = \mathbb{E}[\mathbb{E}[X|\mathcal{H}_2]|\mathcal{H}_1]$$

$\mathbb{P}$-almost everywhere.

On the other hand, let $Z := \mathbb{E}[X|\mathcal{H}_1]$. Then Z is $\mathcal{H}_1$-measurable and hence is $\mathcal{H}_2$-measurable. By Proposition 4.3.2, $\mathbb{E}[Z|\mathcal{H}_2] = Z\mathbb{E}[1|\mathcal{H}_2] = Z = \mathbb{E}[X|\mathcal{H}_1]$ $\mathbb{P}$-almost everywhere. ■

Corollary 4.3.2 (Law of Iterative Expectation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and X, Y be random variables with $\mathbb{E}[|X|] < \infty$. Then*

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Proof. Apply Proposition 4.3.2, take $\mathcal{H}_1 := \sigma(Y)$ and $\mathcal{H}_2 = \{\Omega, \emptyset\}$. The result then follows immediately. ■

Conditional expectation has an even more intuitive interpretation when we treated in L^2 space.

Proposition 4.3.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability measure space and let X be a random variable with $\mathbb{E}[X^2] < \infty$. Then for any σ-algebra $\mathcal{H} \subseteq \mathcal{F}$, any $\mathcal{H}$-measurable random variable Z,*

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}])^2] \leq \mathbb{E}[(X - Z)^2].$$

Proof. Notice that for any $\mathcal{H}$-measurable random variable Z,

$$\begin{aligned}
\mathbb{E}[(X - Z)^2] &= \mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}] + \mathbb{E}[X|\mathcal{H}] - Z)^2] \\
&= \mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}])^2] + \mathbb{E}[(\mathbb{E}[X|\mathcal{H}] - Z)^2] + 2\mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}])(\mathbb{E}[X|\mathcal{H}] - Z)] \\
&= \mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}])^2] + \mathbb{E}[(\mathbb{E}[X|\mathcal{H}] - Z)^2] + 2\mathbb{E}[X(\mathbb{E}[X|\mathcal{H}] - Z) - \mathbb{E}[X(\mathbb{E}[X|\mathcal{H}] - Z)|\mathcal{H}]] \\
&= \mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}])^2] + \mathbb{E}[(\mathbb{E}[X|\mathcal{H}] - Z)^2] \\
&\geq \mathbb{E}[(X - \mathbb{E}[X|\mathcal{H}])^2]
\end{aligned}$$

where the third equality follows from Proposition 4.3.2 and the fact that Z and $\mathbb{E}[X|\mathcal{H}]$ are $\mathcal{H}$-measurable and the last equality follows from Proposition 4.3.2 with $\mathcal{H}_1 = \mathcal{H}$ and $\mathcal{H}_2 = \{\Omega, \emptyset\}$. ■

Therefore, if we regard the random variables as vectors in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, the collection of $\mathcal{H}$-measurable random variables is then a subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Proposition 4.3.5 means that the conditional expectation of a random variable X given $\mathcal{H}$ is the vector in the subspace of $\mathcal{H}$-measurable random variables that minimizes the distance given by the L^2 norm. Since $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is in fact a Hilbert space, this means that $\mathbb{E}[X|\mathcal{H}]$ is exactly the projection of X on the subspace of $\mathcal{H}$-measurable random variables.

4.4 Notions of Convergence

We now introduce some commonly used notion of convergence in probability theory. Throughout the section, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$

Definition 4.4.1. Let $\{X_n\}$ be a sequence of random variables on $(\Omega, \mathcal{F})$ and let X be a random variable on $(\Omega, \mathcal{F})$. We say that $\{X_n\}$ *converges to X in measure* (or *in probability*) if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega | |X_n(\omega) - X(\omega)| > \varepsilon) = 0.$$

Convergence in measure is closely related to other different notions of convergences, one of the most useful ways to establish these connections is the following inequality.

Lemma 4.4.1 (Chebychev's Inequality)**.** *Suppose that $X \geq 0$ is a random variable. Then for any $\lambda > 0$,*

$$\mathbb{P}(\{\omega \in \Omega | X(\omega) \geq \lambda\}) \leq \frac{1}{\lambda} \int_\Omega X d\mathbb{P}.$$

Using Chebychev's inequality, we can easily prove a simple version of the familiar *law of large numbers*

Proposition 4.4.1 (Weak Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables that are independent and have the same induced law with $\mathbb{E}[X_1^2] < \infty$. Then the sequence of random variables $\{\bar{X}_n\}$, where*

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$$

converges in measure to $\mathbb{E}[X]$.

Proof. Take any $\varepsilon > 0$, notice that for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}(\{\omega \in \Omega || \bar{X}_n(\omega) - \mathbb{E}[X]| \geq \varepsilon\}) =& \mathbb{P}(\{\omega \in \Omega || \bar{X}_n(\omega) - \mathbb{E}[X]|^2 \geq \varepsilon\}) \\ \leq& \frac{1}{\varepsilon^2} \int_\Omega (\bar{X}_n - \mathbb{E}[X])^2 d\mathbb{P} \\ =& \frac{1}{\varepsilon^2} \mathrm{Var}(\bar{X}_n) \\ =& \frac{1}{n\varepsilon^2} \mathrm{Var}(X_1).\end{aligned}$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega || \bar{X}_n(\omega) - \mathbb{E}[X]| \geq \varepsilon\}) \leq \lim_{n\to\infty} \frac{1}{n\varepsilon^2} \mathrm{Var}(X_1) = 0$$

for any $\varepsilon > 0$ since $\mathrm{Var}(X_1) \leq \mathbb{E}[X_1^2] < \infty$. ■

Another notion of convergence is in fact a concept that was introduced before.

Definition 4.4.2 (Convergence in L^p)**.** Given $p > 0$. Let $\{X_n\}$ be a sequence of random variables and let X be a random variable with $\mathbb{E}[|X_n|^p] < \infty$ for all $n \in \mathbb{N}$ and $\mathbb{E}[|X|^p] < \infty$. We say that $\{X_n\}$ converges in L^p to X if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

That is, given random variables $\{X_n\}$, X, if we regard them as vectors in the normed linear space $L^p(\Omega, \mathcal{F}, \mathbb{P})$, the convergence is L^p is simply convergence under this norm.

Proposition 4.4.2. *Let $\{X_n\}$ ba a sequence of random variables and let X be a random variable. Suppose that $\{X_n\}$ converges to X in L^p for some $p > 0$. Then $\{X_n\}$ converges to X in measure.*

Proof. Suppose that $\{X_n\}$ converges to X in L^p. Take any $\varepsilon, p > 0$, notice that for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}(\{\omega \in \Omega || X_n(\omega) - X(\omega)| \geq \varepsilon\}) =& \mathbb{P}(\{\omega \in \Omega || X_n(\omega) - X(\omega)|^p \geq \varepsilon^p\}) \\ \leq& \frac{1}{\varepsilon^p} \mathbb{E}[|X_n - X|^p].\end{aligned}$$

Thus,

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega || X_n(\omega) - X(\omega)| \geq \varepsilon\}) \leq \lim_{n\to\infty} \frac{1}{\varepsilon^p} \mathbb{E}[|X_n - X|^p] = 0.$$

■

Definition 4.4.3 (Almost Sure Convergence)**.** Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that $\{X_n\}$ converges to X almost surely if there exists $E \in \mathcal{F}$ with $\mathbb{P}(E) = 1$ such that $\{X_n(\omega)\} \to X(\omega)$ for all $\omega \in E$.

In other words, almost sure convergence is simply convergence pointwise $\mathbb{P}$-almost everywhere. A stronger version of Proposition 4.4.1 establishes almost sure convergence of sample avarege.

Theorem 4.4.1 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables that are independent and have the same induced law with $\mathbb{E}[|X_1|] < \infty$. Then the sequence of random variables $\{\bar{X}_n\}$, where*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

converges almost surely to $\mathbb{E}[X]$.

Clearly, almost sure convergence is stronger then convergence in measure and the converse may not be true. However, there is a further connection between the two.

Proposition 4.4.3. *Suppose that $\{X_n\}$ converges in measure to X. Then there exists a subsequence $\{X_{n_k}\} \subset \{X_n\}$ such that $\{X_{n_k}\}$ converges to X almost surely.*

One last notion of convergence involves the entire distribution associates to the random variable.

Definition 4.4.4. Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_n be the CDF associated with X_n for all $n \in \mathbb{N}$ and let F be the CDF associated with X. We say that $\{X_n\}$ converges in distribution to X if for all $x \in \mathbb{R}$ at which F is continuous,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

An important result associated with convergence in distribution is the *central limit theorem.*

Theorem 4.4.2 (Central Limit Theorem)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables that are independent and have the same induced law with $\mathbb{E}[X_1^2] < \infty$. Then the random*

variable $\sqrt{n}(\bar{X}_n - \mathbb{E}[X_1])$ *converges in distribution to* X*, where* X *has a law of normal distribution with variance* $Var(X_1)$*.*

Remark 4.4.1. Although convergence of sequences are often closely related to the notion of topology, not all of the convergence notions introduced above induces a topology. In fact, convergence in measure and almost sure convergence do not induce any topology. Convergence in L^p is exactly the (strong) topology induced by the $L^p(\Omega, \mathcal{F}, \mathbb{P})$ norm. Convergence in distribution is associated with a certain topology, called the weak-* topology. In the next section, we will introduce some concepts that allow us to further understand this topology.

4.5 Space of Probability Measures

Let X be a metric space, endowed with the Borel σ-algebra. Let $\Delta(X)$ be the collection of probability measures defined on X. In many economic applications, the set $\Delta(X)$ would either be a choice set (see section 3.3) or objects of comparative statics. It is then of great importance to understand the space $\Delta(X)$ and explore the structure of it. This section will give a brief introduction. The materials in this section is relatively more abstract and advance. These will not be necessary for understanding and solving most of the economic models, but knowing some fundamental knowledge about commonly used terminology would be helpful. To begin with, we first introduce some basic concepts about normed linear space.

In Definition 1.1.12, we defined a norm on a linear space X. Given any linear space X, we simply say that X is a normed linear space if it is endowed with a norm $||\cdot||$. Given a linear space X, let X^* be the collection of bounded linear functionals on X. That is:

$$X^* := \{\psi : X \to \mathbb{R} | \psi \text{ is linear, } |\psi(x)| \leq M||x||, \forall x \in X, \text{ for some } M > 0\}.$$

It is easy to verify that for any $\psi \in X^*$, ψ is a continuous function on X under the topology induced by the norm $||\cdot||$, which called the *strong topology*. We often call X^* the *dual space* of X. In fact, X^* is itself a normed linear space

Given any two topologies $\mathcal{T}_1$ and $\mathcal{T}_2$ on X, we say that $\mathcal{T}_1$ is *weaker* than $\mathcal{T}_2$ if $\mathcal{T}_1 \subseteq \mathcal{T}_2$.

Definition 4.5.1. Let X be a normed linear space and X^* be its dual space. The *weak topology* on X is the weakest topology under which the functions X^* is continuous on X.

In other words, weak topology is the coarsest topology under which all the bounded linear functionals on X remain continuous. The convergence notion associated with the weak topology is as follows: $\{x_n\} \subseteq X$ converges under the weak topology to $x \in X$ if and only if for any $\psi \in X^*$,

$$\lim_{n\to\infty} \psi(x_n) = \psi(x).$$

On the other hand, by definition, for any $\psi \in X^*$, there exists $M > 0$ such that

$$|\psi(x)| \leq M||x||, \forall x \in X.$$

Let $||\psi||$ be the infimum of these M. It can be verify that $||\cdot||$ is also a norm on X^* and hence X^* is a normed linear space. Now consider a family of functionals on X^* that takes form of J_x, for some $x \in X$, defined by:

$$J_x(\psi) := \psi(x), \forall \psi \in X^*$$

By definition, J_x is linear and bounded on the normed linear space X^*.

Definition 4.5.2. Let X be a normed linear space and X^* be its dual space. The *weak-* topology* on X^* is the weakest topology under which the family of functions $\{J_x\}_{x\in X}$ are continuous.

Analogously, the notion of convergence associated with the weak-* topology is that $\{\psi_n\} \subseteq X^*$ converges to $\psi \in X^*$ under the weak-* topology if and only if

$$\lim_{n\to\infty} J_x(\psi_n) = J_x(\psi),$$

which is equivalent to

$$\lim_{n\to\infty} \psi_n(x) \to \psi(x), \forall x \in X.$$

That is, weak-* convergence is simply pointwise convergence.

One of the most important features of X^* under the weak-* topology is that closed unit ballas are compact again.

Theorem 4.5.1 (Alaoglu)**.** *Let X be a normed linear space and X^* be its dual. Define*

$$B^* := \{\psi \in X^* | ||\psi|| \leq 1\}$$

as the closed unit ball in X^*. Then B^* is compact with respect to the weak- topology.*

With the notions introduced above, we are now ready to examine the topological structure of $\Delta(X)$. More specifically, we will consider the case when X is a compact metric space. This is not the most general case but will be sufficient for most of the applications. To begin with, let X be a compact metric space and let $C(X)$ be the set of continuous functions on X endowed with the sup norm. That is, for any $f \in C(X)$, $||f|| := \max_{x \in X} |f(x)|$. Under this norm, $C(X)$ is a normed linear space and has a dual $C(X)^*$. The most important result that connects probability measures and the concepts introduced above is the following:

Theorem 4.5.2 (Riesz Representation Theorem)**.** *Let X be a compact metric space and $C(X)$ be the collection of continuous functions on X. Then for any $\psi \in C(X)^*$ with $||\psi|| = 1$, there exists a unique Borel probability measure $\mu \in \Delta(X)$ such that*

$$\psi(f) = \int_X f d\mu, \forall f \in C(X).$$

That is, through the set of continuous functions $C(X)$, the space $\Delta(X)$ is in fact isomorphic to the closed unit ball in the dual space of $C(X)$. As a result, $\Delta(X)$ can be studied as the closed unit ball of the dual space of a normed linear space. Therefore, we can define a natural topology on $\Delta(X)$ by simply using the weak-* topology on $C(X)^*$. That is, we say that $\{\mu_n\} \subseteq \Delta(X)$ converges in weak-* to $\mu \in \Delta(X)$ if and only if the associated bounded linear functionals on $C(X)^*$ converges in weak-*, i.e.,

$$\lim_{n \to \infty} \int_X f d\mu_n = \int_X f d\mu, \forall f \in C(X).$$

Combining Theorem 4.5.1, Theorem 4.5.2 and linearity, the following properties can be established.

Proposition 4.5.1. *Let X be a compact metric space, the set of Borel probability measures on X, $\Delta(X)$, is convex and compact under the weak-* topology.*

Being a compact and convex set, the solutions of many economic applications can then be guaranteed to exist. For instance, the sender's problem in section 3.3, (15), must have a solution because $\Delta(\Omega)$ is a compact metric space and therefore $\Delta(\Delta(\Omega))$ is a compact set under the weak-* topology and the constraint set is a closed set under the weak-* topology and the objective is linear.

One final remark, the weak-* topology introduced here turns out to be very easy to check if X is a compact interval on $\mathbb{R}$. The following result states that, when X is an interval on $\mathbb{R}$, weak-* convergence is the same as convergence of distribution defined in the last section.

Theorem 4.5.3 (Portmanteau Lemma)**.** *Let X be a compact interval on $\mathbb{R}$, $\{\mu_n\} \subseteq \Delta(X)$ be a sequence of Borel probability measures, $\mu \in \Delta(X)$ be a Borel probability measure. Let F_n be the CDF associated with the measure μ_n for all $n \in \mathbb{N}$ and let F be the CDF associated with μ. Then $\{\mu_n\}$ converges to μ under the weak-* topology if and only if for any $x \in X$ at which F is continuous,*

$$\lim_{n\to\infty} F_n(x) = F(x).$$