Mathematics Review for the Field of Economics

Instructor Takuma Habu

This version: 23rd August 2024

Cornell University

Contents

Introduction

The note is a collection of concepts and results that I scrounged from many places that I think you will find useful in navigating and completing your first year in the field of economics at Cornell University. It is intended to be a self-contained reference for you as you progress in your first year. The course itself prioritises practicality, rather than rigour. We will only go over proofs that deepens your understanding of the relevant result and/or to help you familiarise yourselves with the proof techniques that would be useful in your core classes. We will not be spending time to practice the mathematics you cover—you will be doing that throughout the rest of the year. Finally, even if some materials seem esoteric at times (especially at the beginning of the class), I promise you that there is a good reason behind why we are going over them.

Tips for success Reading books and notes and listening to lectures (i.e. passing learning) won't work—you have to learn actively by working through the proofs and examples we go over in class (active reading), taking notes during class (active listening), and solving practice questions. Working in groups is strongly encouraged BUT always try to work through all problems on your own before meeting with others. And check your understanding by explaining it to others!

Textbooks

- Sundaram, "A First Course in Optimization Theory".
- Simon and Blume, "Mathematics for Economists".
- Sydsaester, Hammond, Seierstad and Strom, "Further Mathematics for Economic Analysis".
- Hansen, "Probability and Statistics for Economists"

If you find any mistakes and/or typos, please let me know by emailing me at takumahabu@cornell.edu.

Part I
Foundational concepts

1 Preliminaries

Throughout, I will use the symbol ":=" to define things.[1] For example, if I want to define a function f to be a constant function that always equals 1, I will write $f(x) \coloneqq 1$. Note that I can write $1 \coloneqq 2$ and it wouldn't technically be wrong—it just means that I will use 1 to mean 2 (of course, I will not write such nonsense!).

I denote $\mathbb{N} := \{1, 2, \ldots\}$ as the set of *natural* numbers,[2] $\mathbb{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ as the set of *integers*, $\mathbb{Q} := \{\frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{N}\}$ as the set of *rational* numbers, and $\mathbb{R}$ as the set of *real* numbers.

1.1 Proofs

1.1.1 Propositional logic

A *proposition* is a statement that has either value *true* (T) or *false* (F) but not both. Let p and q be two propositions. Two statements p and q are *logically equivalent*, written $p \equiv q$, if they always have the same truth/false value; i.e., they have the same *truth value*s. Write the *negation* of proposition p as $\neg p$ (read "not p"). The compound statement "p *and* q" is written $p \wedge q$. The compound statement "p *or* q" is written $p \vee q$. The table below sets is a *truth table* the truth values of negation and compounding of proposition(s).

Table 1: A truth table.

p	q	$\neg p$	$\neg q$	$p \wedge q$	$p \vee q$
T	T	F	F	T	T
T	F	F	T	F	T
F	T	T	F	F	T
F	F	T	T	F	F

Exercise 1 (De Morgan's laws)**.** When negating compound statements, prove that we can move the negation "inside" and flip the "sign" as below:

$$\neg (p \vee q) \equiv \neg p \wedge \neg q,$$
$$\neg (p \wedge q) \equiv \neg p \vee \neg q.$$

A statement that is always false is called a *contradiction* (e.g., $p \wedge (\neg p)$). A statement that is always true is called a *tautology* (e.g., $p \vee (\neg p)$).

We say p *implies* q, denoted $p \Rightarrow q$, if it is the case that "if p (is true), then q (is true);" importantly, no conclusion about q can be drawn if p is false. The *converse* of $p \Rightarrow q$ is $q \Rightarrow p$ (read either as "if q, then p" or "p only if q"). A *two-way implication*, written $p \Leftrightarrow q$ (read *if and only if (iff)*), means $p \Rightarrow q$ and $q \Rightarrow p$; i.e.,

$$p \Leftrightarrow q \equiv (p \Rightarrow q) \wedge (q \Rightarrow p).$$

The *contrapositive* of $p \Rightarrow q$ is $\neg q \Rightarrow \neg p$.

[1] Other notations include $\triangleq$ and $=_{def}$.
[2] Sometime people also include 0 as part of $\mathbb{N}$.

Exercise 2 (Truth table)**.** Extend the truth table, Table 1, to include $p \Rightarrow q$, $q \Rightarrow p$, $p \Leftrightarrow q$, and $\neg q \Rightarrow \neg p$.

Exercise 3 (Modus Ponens)**.** Suppose it is the case that "if it is raining, then there are clouds" and "it is raining." Can you conclude whether there are any clouds?

Exercise 4 (Modus Tollens)**.** Suppose it is the case that "if it is raining, then there are clouds" and "there are no clouds." Can you conclude whether it is raining?

A *variable proposition*, or *predicates*, are propositions whose truth value depends on some other parameter, say x that can take values from a collection of values X. Write $p(x)$ as the truth value of statement p given parameter x. Say that $p(x)$ is *sufficient* for $q(x)$ if $p(x) \Rightarrow q(x)$ is true for all possible values of x, and say that $p(x)$ is *necessary* for $q(x)$ if the converse, $q(x) \Rightarrow p(x)$, is true for all possible values of x. The statement "$\forall x \in X,\ p(x)$", or equivalently, "$p(x)\ \forall x \in X$", means that $p(x)$ is true for all possible parameter x in X. The symbol $\forall$ is called the *universal quantifier* and can also be read as "for every." The statement "$\exists x \in X,\ p(x)$" means that $p(x)$ is true for some parameter x in X. The symbol $\exists$ is called the *existential qualifier* and can also be read as "there exists" or "for at least one." Write $\exists!$ to mean "there exists unique".

Proposition 1. *When negating qualifiers, we can move the negation "inside" and swap the "sign" as below:*

$$\neg\left(\forall x \in X,\ p\left(x\right)\right) \equiv \left(\exists x \in X,\ \neg p\left(x\right)\right),$$
$$\neg\left(\exists x \in X,\ p\left(x\right)\right) \equiv \left(\forall x \in X,\ \neg p\left(x\right)\right).$$

Proof. Note that

$$\forall x \in X,\ p\left(x\right) \equiv p\left(x\right) \wedge p\left(x'\right) \wedge \cdots.$$

Applying De Morgan law repeatedly to gives

$$\neg\left(\forall x \in X,\ p\left(x\right)\right) \equiv \neg p\left(x\right) \vee \neg p\left(x\right) \wedge \cdots$$

and so there must be at least one x in X for which $p(x)$ is false. The proof for the negation of the existential qualifier is analogous. ■

Exercise 5. *Read* the following statement.

$$\forall \epsilon > 0,\ \exists \delta_{\epsilon,x} > 0,\ \left|y - f\left(s\right)\right| < \epsilon\ \forall s \in S \backslash \left\{x\right\} : \left|x - s\right| < \delta_{\epsilon,x}.$$

Convince yourself that the statement above can equivalently be written as

$$\forall \epsilon > 0,\ \exists \delta_{\epsilon,x} > 0,\ \left(\left|x - s\right| < \delta_{\epsilon,x},\ s \in S \backslash \left\{x\right\}\right) \Rightarrow \left|y - f\left(s\right)\right| < \epsilon.$$

Finally, write and read the negation of the statement(s) above?

1.1.2 Proof methods

A *proof* of a statement shows that the statement is *always* true given some assumptions. An example does not constitute a proof; however, a *counterexample* can be used to show that a statement is false (i.e., can be used to *disprove* a statement). Following are the main methods to approach a proof.

- Deduction/direct proof: Start with what you know, set up a chain of argument to end up with what you wanted to show.

- Contraposition: direct proof of the contrapositive statement.
- Contradiction: Assume that the required result is false then work to obtain a contradiction.
- Induction: Prove the base case (e.g., $n = 0$) then the induction step while assuming the induction hypothesis (assume it holds for n, then show that this implies the result for $n+1$). Then, invoke the Principle of Mathematical Induction.

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

Example 1. Suppose we wish to prove that the sum of the first n natural numbers is $\frac{n(n+1)}{2}$ for any $n \in \mathbb{N}$.

- A direct proof is to fix $n \in \mathbb{N}$, and let the sum be S_n:

  $$S_n = 1 + 2 + \cdots + (n-1) + n.$$

  Writing in reverse order:

  $$S_n = n + (n-1) + \cdots + 2 + 1.$$

  And os

  $$2S_n = (n+1) + (n+1) + \cdots + (n+1) + (n+1).$$

  Rearranging gives $S_n = \frac{n(n+1)}{2}$. Since we chose n arbitrarily, the result holds for all $n \in \mathbb{N}$.

- A proof by induction proceeds first by ensuring that the base case $n = 1$ holds—the sum of just the first natural number is just 1. From that formula, we also have $S_1 = \frac{1(1+1)}{2} = 1$. Thus, the formula holds for $n = 1$. To proceed, now suppose that the formula holds for some specific value of n, say $n = k$ for some $k \in \mathbb{N}$. We ask now if the formula holds for $n = k+1$. That is, suppose

  $$S_k = 1 + 2 + \cdots + k - 1 + k = \frac{k\,(k+1)}{2}.$$

  Then,

  $$S_{k+1} = S_k + (k+1) = \frac{k\,(k+1)}{2} + (k+1) = \frac{(k+1)\,(k+1+1)}{2}.$$

  Because the last expression is the same as the formula for S_k when we replace k by $k+1$, we can conclude that if the formula is correct for k, then it is also correct for $k+1$. Finally, we invoke the Principle of Mathematical Induction to conclude that formula is correct for all n.

Example 2 (Proof by induction)**.** We take the overlapping generation (OLG) model that you will cover in Macroeconomics. An agent, or *a generation*, that lives for two periods is born in each time period $t \in \mathbb{N}$. Let $x_t^i \in \mathbb{R}$ and $e_t^i \in \mathbb{R}_+$ denote generation i's consumption and endowment in period t respectively in a pure exchange economy. Assume that there is an "initial old" generation in $t = 1$ whose consumption is denoted x_1^0. By assumption, $x_t^i = 0$ and $e_t^i = 0$ for all $t \notin \mathbb{N}\backslash\{i, i+1\}$, and each generation have strictly positive endowments while they are alive; i.e., for all $i \in \mathbb{N}$, $e_t^i > 0$ for $t \in \{i, i+1\}$. Market clearing (or feasibility) condition is that

$$x_t^{t-1} + x_t^t = e_t^{t-1} + e_t^t \ \forall t \in \mathbb{N}.$$

Let $p_t \in \mathbb{R}_+$ be the price of goods in period t, then the budget constraint for generation $t \in \mathbb{N}$ is

$$p_t x_t^t + p_{t+1} x_{t+1}^t = p_t e_t^t + p_{t+1} e_{t+1}^t, \tag{1.1}$$

where we have implicitly assumed that the preferences are strictly increasing so that the, budget constraint binds. We will prove by induction that any competitive equilibrium is *autarky* (meaning no trade)*;* i.e., for any price vector,

$$x_t^{t-1} = e_t^{t-1},\ x_t^t = e_t^t\ \forall t \in \mathbb{N}. \tag{1.2}$$

First, we prove the base case: Since the initial old care only about consumption in $t = 1$, it must be that

$$x_1^0 = e_1^0.$$

Combining with the period-1 market-clearing condition, $x_1^0 + x_1^1 = e_1^0 + e_1^1$, implies

$$x_1^1 = e_1^1.$$

Thus, (1.2) holds for $t = 1$.

Now, we make the inductive hypothesis that (1.2) holds for some period $t > 1$. The goal is to show that (1.2) then holds for period $t + 1$. Since $x_t^t = e_t^t$ by the inductive hypothesis, budget constraint, (1.1), implies that $x_{t+1}^t = e_{t+1}^t$. Combining this with the period-$t + 1$ market clearing condition, $x_{t+1}^t + x_{t+1}^{t+1} = e_{t+1}^t + e_{t+1}^{t+1}$, yields that $x_{t+1}^{t+1} = e_{t+1}^{t+1}$. Thus, (1.2) indeed holds for period $t + 1$.

Therefore, by induction, (1.2) holds for all $t \in \mathbb{N}$.

Remark 1. I often write "TFU: [some statement]". The TFU stands "True, False, or Uncertain," what I mean here is that I'm looking for a proof of the statement (the "T" case), or a counterexample (the "F" case), or cases in which the statement is sometimes true and other times false (the "U" case).

1.2 Sets

A *set* X is a collection of objects or *elements*. If x is an element of X, we write $x \in X$ and if x is not an element of X, then we write $x \notin X$. The symbol $\varnothing \coloneqq \{x : x \neq x\}$ denotes the *empty set* and $\varnothing \subseteq X$ for any set X. A set X is *nonempty* if $X \neq \varnothing$.

Suppose X and Y are sets. Then, X is a *subset* of Y if every element of X is also an element of Y; i.e.,

$$X \subseteq Y \Leftrightarrow \forall x \in X,\ x \in Y.$$

The set X is a *proper subset* of Y if X is a subset of Y and, in addition, there exists $y \notin X$ such that $y \in Y$.[3]

$$X \subset Y \Leftrightarrow (\forall x \in X,\ x \in Y) \wedge (\exists y \in Y,\ y \notin X)\,.$$

Two sets X and Y are equal if every element of one is contained in the other; i.e.,

$$X = Y \Leftrightarrow (X \subseteq Y) \wedge (Y \subseteq X) \Leftrightarrow (z \in X \Leftrightarrow z \in Y)\,.$$

The *union* of two sets X and Y are defined as

$$X \cup Y \coloneqq \{z : (z \in X) \vee (z \in Y)\}\,.$$

The *intersection* of two sets X and Y are defined as

$$X \cap Y \coloneqq \{z : (z \in X) \wedge (z \in Y)\} \equiv \{z \in X : z \in Y\} \equiv \{z \in Y : z \in X\}\,.$$

[3]Rather confusingly, some write $\subset$ to mean $\subseteq$, in which case a proper subset is denoted $\subsetneq$ or $\subsetneqq$.

The intersection and union are both *commutative* (i.e., $X \cap Y = Y \cap X$ and $X \cup Y = Y \cup X$) and *associative* (i.e., $(X \cap Y) \cap Z = X \cap (Y \cap Z)$ and $(X \cup Y) \cup Z = X \cup (Y \cup Z)$). Moreover,

$$X \cap Y = X \Leftrightarrow X \subseteq Y,$$
$$X \cup Y = X \Leftrightarrow Y \subseteq X.$$

Let I be some set, and X_i be sets for each $i \in I$. We refer to $\{X_i\}_{i \in I}$ as a *collection of sets* (i.e., a set of sets) and I as an *index set*. Denote the union and intersection of all X_i's, respectively, as

$$\bigcup_{i \in I} X_i \coloneqq \{x : \exists i \in I,\ x \in X_i\},$$
$$\bigcap_{i \in I} X_i \coloneqq \{x : \forall i \in I,\ x \in X_i\}.$$

When $I = \{1, \ldots, n\}$ for some $n \in \mathbb{N}$, we may also write $\{X_i\}_{i=1}^n$, $\bigcup_{i=1}^n X_i$ and $\bigcap_{i=1}^n X_i$. The intersection is *distributive with respect to the union*; i.e.,

$$Y \cap \left(\bigcup_{i \in I} X_i \right) = \bigcup_{i \in I} (Y \cap X_i);$$

The union is *distributive with respect to the intersection*; i.e.,

$$Y \cup \left(\bigcap_{i \in I} X_i \right) = \bigcap_{i \in I} (Y \cup X_i).$$

The two sets X and Y are *disjoint* if they do not share elements; i.e., if $X \cap Y = \varnothing$. The *difference* of two sets X and Y, or *complement of Y relative to X*,[4] are defined as

$$X \backslash Y \coloneqq \{z : (z \in X) \wedge (z \notin Y)\} \equiv \{x \in X : x \notin Y\}.$$

If X and Y are disjoint, then $X \backslash Y = X$. Let Z bet the set of all "relevant" objects and suppose $X \subseteq X$. The *complement* of X is defined as

$$X^c \coloneqq Z \backslash X.$$

Note $x \notin X \Leftrightarrow x \in X^c$ and $(X^c)^c = X$.

Suppose $(X_i)_{i \in I} \subseteq Z$. *De Morgan's laws* tell us that

$$\left(\bigcup_{i \in I} A_i \right)^c \equiv \bigcap_{i \in I} A_i^c, \quad \left(\bigcap_{i \in I} A_i \right)^c \equiv \bigcup_{i \in I} A_i^c.$$

Given a set X, the *power set* of X, denoted 2^X or $\mathcal{P}(X)$, is the collection of all subsets of X (including the empty set); i.e.,

$$2^X \equiv \mathcal{P}(X) \coloneqq \{S : S \subseteq X\}.$$

Set of all sets is not well defined (google *Russell's paradox*)—this is a potential issue that can arise in, for example, information design when thinking about the set of all possible "signals" about an unknown state of the world.

[4] Sometimes, $X \backslash Y$ is written $X - Y$.

Example 3. Let $X \coloneqq \{a, b, c\}$. Then,

$$2^X = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}, X\}.$$

The *cardinality* of a set X, denoted $|X|$, is the number of elements the set X contains. The set X is *finite* if $|X| < \infty$ and *infinite* if it is not finite. If X is finite, then

$$X \subset Y \Rightarrow |X| < |Y|.$$

Given two sets X and Y, and $x \in X$ and $y \in Y$, (x, y) is an *ordered pair* (or a *tuple*) such that (x, y) is different from (y, x). The *Cartesian product* of two sets X and Y is the set of all ordered pairs; i.e.,

$$X \times Y \coloneqq \{(x, y) : x \in X,\ y \in Y\}.$$

More generally, given a finite index set $I = \{1, \ldots, n\}$, the Cartesian product of sets $X_1, \ldots, X_n$ is the set of all *n-tuples*; i.e.,[5]

$$\times_{i=1}^{n} X_i \coloneqq \{(x_1, \ldots, x_n) : x_i \in X_i\ \forall i \in \{1, \ldots, n\}\}.$$

If $X_i = X$ for all $i \in I$ (and $|I| = n$), then we write

$$X^n \equiv \times_{i \in I} X_i.$$

A *partition* of Y is a collection of disjoint sets whose union is Y; i.e., a collection $\{X_i\}_{i=1}^n$ is a partition of Y if

$$\left(\bigcup_{i=1}^{n} X_i = Y\right) \wedge (X_i \cap X_j = \varnothing\ \forall i \neq j).$$

("$\forall i \neq j$" means $\forall (i, j) \in \{(x, y) \in \{1, \ldots, n\}^2 : x \neq y\}$.)

1.3 Binary relations, partial and total orders

1.3.1 Binary relations

Given sets X and Y, $R \subseteq X \times Y$ is called a *binary relation from X to Y*. Write

$$xRy \Leftrightarrow (x, y) \in R,$$
$$\neg xRy \Leftrightarrow (x, y) \notin R.$$

The *inverse* of a binary relation R from X to Y is a binary relation from Y to X defined as

$$R^{-1} \coloneqq \{(y, x) \in Y \times X : (x, y) \in R\}.$$

Say that a binary relation on X is:

- *reflexive* if $\forall x \in X$, xRx;
- *symmetric* if $\forall x, y \in X$, $xRy \Leftrightarrow yRx$;
- *transitive* if $\forall x, y, z \in X$, $(xRy \wedge yRz) \Rightarrow xRz$;
- *antisymmetric if* $\forall x, y \in X$, $(xRy \wedge yRx) \Rightarrow x = y$ (i.e., rules out ties);
- *complete* if $\forall x, y \in X$, either xRy or yRx (i.e., every pair is ordered).

[5] Sometimes, Cartesian product is also written as $\prod_{i \in I} X_i$.

Note, in particular, that completeness implies reflexiveness.

1.3.2 Other relations

A binary relation $\geq$ on X (i.e., $\geq \subseteq X \times X$) is:

- a *preorder* if it is reflexive and transitive;
- an *equivalence* relation, denoted $\sim$, if it is reflexive, symmetric and transitive;
- a *partial order* if it is reflexive, transitive and antisymmetric
- a *total order* if it is complete, transitive and antisymmetric.[6]

Given a partial order on X, say that $(X, \geq)$ is a *partially ordered set* (*poset*). Given a total order on X, say that $(X, \geq)$ is a *totally order set* (*toset*).

For any $x \in X$, *equivalence class of x relative to* $\sim$ is defined as

$$[x]_{\sim} \coloneqq \{y \in X : y \sim x\}.$$

A *quotient set of* X relative to $\sim$, denoted $X/\sim$, is the class of all equivalence classes relative to $\sim$; i.e.,

$$X/\sim \coloneqq \{[x]_{\sim} : x \in X\}.$$

Totally ordered set is a special case of partially ordered sets. Based on a partially ordered set $(X, \geq)$, we may define

$$\begin{aligned} > &\coloneqq \left\{(x, y) \in X^2 : (x \geq y) \wedge \neg (y \geq x)\right\}, \\ \leq &\coloneqq \geq^{-1}, \\ < &\coloneqq >^{-1}. \end{aligned}$$

Observe that if $\geq$ is a binary relation on X, then $\geq$ is a binary relation on any $S \subseteq X$. A *rational preference relation* on X is a binary relation on X that is complete and transitive. You will see more of this in consumer theory.

Example 4. $(\mathbb{R}, \geq)$ is a total order defined as:

$$\geq \coloneqq \left\{(x, y) \in \mathbb{R}^2 : y - x \text{ is nonnegative}\right\}.$$

However, $(\mathbb{R}^n, \geq)$ with $n \in \mathbb{N}\backslash\{1\}$ is a partial order defined as

$$\geq \coloneqq \left\{\left((x_i)_{i=1}^n, (y_i)_{i=1}^n\right) \in \mathbb{R}^n \times \mathbb{R}^n : y_i - x_i \text{ is nonnegative } \forall i \in \{1, \ldots, n\}\right\}.$$

Observe that above is not complete (e.g., with $n = 2$, $(1, 2)$ and $(2, 1)$ are not ordered).

Example 5. Given a set X, $(2^X, \subseteq)$ is a partial order.

There are many other examples of partial orders: e.g., first-order and second-order stochastic dominance of distributions, ordering of utility functions using absolute risk aversion.

[6] Some people refer to total orders as linear orders.

1.3.3 Upper and lower bounds

Let $(X, \geq)$ be a partially ordered set and $S \subseteq X$. An element $u \in X$ is an *upper bound* of S if, $\forall s \in S$, $u \geq s$. If such an u exists, then we say that the set S is *bounded from above.* An element in S is a *maximum* of S if it is an upper bound of S and we denote it as $\max S$. Similarly, $\ell \in X$ is a *lower bound* of S if, $\forall s \in S$, $\ell \leq s$. If such an ℓ exists, then we say that the set S is *bounded from below.* A set S is bounded if its both bounded from above and bounded from below. An element is a *minimum* of S if it is a lower bound of S and we denote it as $\min S$.

An element is the *least upper bound* or the *supremum* of S, denoted $\sup S$, if (i) $\sup S$ is an upper bound of S; and (ii) $\sup S \leq u$ for any upper bound u of S. An element is the *greatest lower bound* or the *infimum* of S, denoted $\inf S \in X$, if (i) $\inf X$ is a lower bound of S; and (ii) $\inf X \geq \ell$ for any lower bound $\ell \in X$ of S.

Whenever a maximum or a minimum of S exists, they must be unique so that it make sense to refer to them as *the* maximum or minimum.

Proposition 2. *If a maximum or a minimum of S exists, then it must be unique.*

Proof. By way of contradiction (BWOC), suppose that there are two distinct maxima of S, denoted $\overline{s}_1$ and $\overline{s}_2$, where $\overline{s}_1 \neq \overline{s}_2$. By definition of maximum, $\overline{s}_1, \overline{s}_2 \geq s \ \forall s \in S$. In particular, it must be that

$$\overline{s}_1 \geq \overline{s}_2 \text{ and } \overline{s}_2 \geq \overline{s}_1.$$

Since S is a partially ordered set, antisymmetry implies that $\overline{s}_1 = \overline{s}_2$; a contradiction. ■

By definition, the supremum is the minimum of the set of upper bounds and the infimum is the maximum of the set of lower bounds.

Proposition 3. *If a maximum exists, then it is also the supremum. If a minimum exists, then it is also the infimum.*

The following tells us that taking supreme of a smaller set is lower and infimum of a smaller set is greater.

Proposition 4. *Suppose that $S \subseteq X$ where $(X, \geq)$ is a partially ordered set. If $\sup S$ and $\sup X$ exist, then $\sup S \leq \sup X$. If $\inf S$ and $\inf X$ exist, then $\inf S \geq \inf X$.*

Proof. Since $\sup X$ is an upper bound of X and $S \subseteq X$, $\sup X$ is also an upper bound of S. Since $\sup S$ is the least upper bound of S, we must have $\sup S \leq \sup X$. Since $\inf X$ is a lower bound of X and $S \subseteq X$, $\inf X$ is also a lower bound of S. Since $\inf S$ is the greatest lower bound of S, we must have $\inf X \leq \inf S$. ■

Supremum or infimum may not exist even when upper or lower bounds exist. For example, suppose $X := \mathbb{R}\backslash\{0\}$ and $S := \{x \in \mathbb{R} : x < 0\}$. Then, $S \subseteq X$, $(X, \leq)$ is a partially ordered set, and, for example, $1 \in X$ is an upper bound of S. However, there is no least upper bound. Thus, we say that a partially ordered set $(X, \leq)$ has the *least upper bound (resp. greatest lower bound) property* if any nonempty subset of X bounded from above (resp. below) has a least upper (resp. greatest lower) bound. We will study such sets in details in ECON 6701 in the context of *lattices.* We also briefly mention a lemma that we will use a little later.

Lemma 1 (Zorn's lemma)**.** *Suppose a partially order set $(X, \geq)$ has the property that every totally ordered subset of X has an upper bound in X. Then, the set X contains at least one maximal element.*

1.4 Functions

A *function* (or a *mapping*) f is a binary relation f from X to Y such that: $\forall x \in X$, (i) $\exists y \in Y$, $(x, y) \in f$; and (ii) $(x, y_1), (x, y_2) \in f \Rightarrow y_1 = y_2$. That is, a function assigns a single element in Y to every element in X. A function from X to Y is denoted

$$f : X \to Y \text{ or } f \in Y^X$$

and the assigned element is denoted $f(x)$. We call X as the *domain* and Y as the *codomain* of f. If $X = Y$, then the function is a *self-map* on X.

The *image* of a set $S \subseteq X$ under a function $f : X \to Y$ is

$$f(S) \coloneqq \{y \in Y : \exists s \in S,\ f(s) = y\}.$$

The *range* of $f : X \to Y$ is the set of all values that f can generate in its codomain and is denoted $f(X)$. It is possible that $f(X) \subset Y$.

The *graph* of a function $f : X \to Y$ is defined

$$\operatorname{gr}(f) \coloneqq \{(x, y) \in X \times Y : y = f(x)\},$$

The *preimage* (or the *inverse image*) of $f : X \to Y$ is defined as subset of the set of all elements in the domain that maps to members of the codomain; i.e.,

$$f^{-1}(Y) \coloneqq \{x \in X : f(x) \in Y\}.$$

Proposition 5. *Suppose $f : X \to Y$, $X_1, X_2 \subseteq X$, and $Y_1, Y_2 \subseteq Y$.*

(i) $X_1 \subseteq X_2 \Rightarrow f(X_1) \subseteq f(X_2)$ *and* $Y_1 \subseteq Y_2 \Rightarrow f^{-1}(Y_1) \subseteq f^{-1}(Y_2)$*;*

(ii) $f(X_1 \cup X_2) = f(X_1) \cup f(X_2)$*;*

(iii) $f(X_1 \cap X_2) \subseteq f(X_1) \cap f(X_2)$ *(why not equal?);*

(iv) $f^{-1}(Y_1 \cup Y_2) = f^{-1}(Y_1) \cup f^{-1}(Y_2)$*;*

(v) $f^{-1}(Y_1 \cap Y_2) = f^{-1}(Y_1) \cap f^{-1}(Y_2)$ *(why is this equal?);*

(vi) $f^{-1}(Y \backslash Y_1) = X \backslash f^{-1}(Y)$.

Exercise 6. In the previous proposition, given an example in which (iii) holds "strictly." Prove (v).

A function $f : X \to Y$ is:

- *one-one* or *injective* if every $x \in X$ is assigned to a unique element in Y; i.e.,

$$\forall x_1, x_2 \in X,\ f(x_1) = f(x_2) \Rightarrow x_1 = x_2.$$

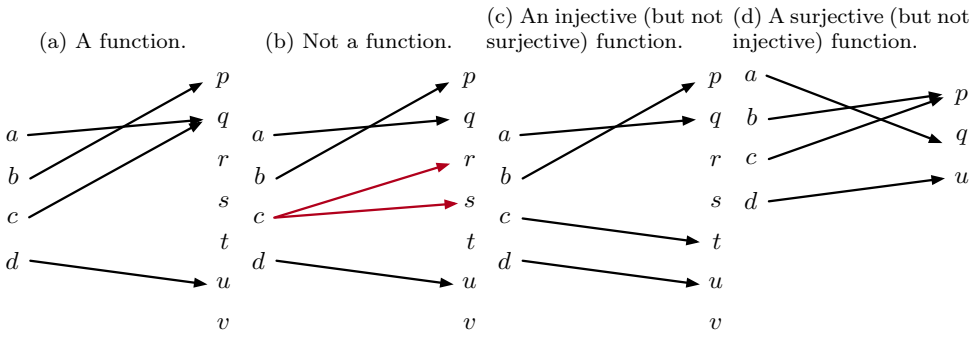
- *onto* or *surjective* if, for all $y \in Y$, there exists $x \in X$ such that $f(x) = y$; i.e.,

$$f(X) = Y;$$

- *bijective* if it is injective and surjective so that every element in the domain is mapped to a unique element in the codomain;

- *invertible* if the inverse image mapping is a function from Y to X.

Figure 1.1: Functions.



Observe that preimage of f, f^{-1}, may not be a function if multiple x's are assigned to the same element $y \in Y$. We say that a function f is invertible if its preimage is itself a function.

Proposition 6. *A function is invertible if and only if it is bijective.*

Exercise 7. Prove the proposition above.

Suppose $S \subseteq X \subseteq Z$ and $f : X \to Y$. The *restriction of f to S* is a function $f|_S : S \to Y$ defined as

$$f|_S(s) \coloneqq f(s).$$

An *extension* of f to Z is a function $f^* : Z \to Y$ such that $f^*|_X = f$.

A *projection from $X \times Y$ onto X* is a function $f : X \times Y \to X$ defined by $f(x, y) \coloneqq x$. Similarly, a projection from $X \times Y$ onto Y is a function $f : X \times Y \to Y$ defined by $f(x, y) \coloneqq y$.

Proposition 7. *Suppose $f : X \to Y$, $A \subseteq X$, and $B \subseteq Y$.*

(i) $f(f^{-1}(Y)) = f(X)$ *and* $f^{-1}(f(X)) = X$;

(ii) $f(f^{-1}(B)) \subseteq B$ *with equality if f is surjective;*

(iii) $A \subseteq f^{-1}(f(A))$ *with equality if f is injective.*

Exercise 8. Prove the previous proposition.

Remark 2 (Cardinality of sets). For example, $X = \{1, 2, 3, 4\}$ is numerically equivalent to $Y = \{4, 7, 10, 13\}$ under the (bijective) function $f(x) = 3x + 1$. A set X is *countable* if it is numerically equivalent to $\mathbb{N}$ and *uncountable* if it is not countable. Note $\mathbb{Q}$ is countable while $\mathbb{R}$ is uncountable.

1.4.1 Composition and arithmetic of functions

Given two functions $f : X \to Y$ and $g : Y \to Z$, the *composite function* is a mapping from X to Z defined as

$$g \circ f \coloneqq g(f(x)) \ \ \forall x \in X.$$

The composite operation is associative (i.e., $(h \circ g) \circ f = h \circ (g \circ f)$) but not commutative (i.e., $g \circ f \neq f \circ g$).

If $f, g : X \to Y$, then:

- the *sum*, denoted $f + g$, is defined by $(f + g)(x) \coloneqq f(x) + g(x)$;
- *multiplication by a scalar constant*, denoted λf with $\lambda \in \mathbb{R}$, is defined by $(\lambda f)(x) \coloneqq \lambda f(x)$;
- the *product*, denoted fg, is defined by $(fg)(x) \coloneqq f(x)g(x)$;
- the *quotient*, denoted $\frac{f}{g}$, is defined by $(\frac{f}{g})(x) \coloneqq \frac{f(x)}{g(x)}$.

1.4.2 Monotonicity

When both the domain and codomain of functions are ordered sets, we can talk about monotonicity of a function.

Suppose $(X, \geq_X)$ and $(Y, \geq_Y)$ are posets, then a function $f : X \to Y$ is

- *weakly increasing* (or *nondecreasing*) if $\forall x, x' \in X,\ x \geq_X x' \Rightarrow f(x) \geq_Y f(x')$;
- *weakly decreasing* (or *nonincreasing*) if $\forall x, x' \in X,\ x \geq_X x' \Rightarrow f(x) \leq_Y f(x')$;
- *strictly increasing* if $\forall x, x' \in X,\ x >_X x' \Rightarrow f(x) >_Y f(x')$;
- *strictly decreasing* if $\forall x, x' \in X,\ x >_X x' \Rightarrow f(x) <_Y f(x')$.

A *monotone* function is a function that is either weakly increasing or weakly decreasing. A *strictly monotone* function is either strictly increasing or strictly decreasing.

1.4.3 Correspondence

A *correspondence* F is a binary relation F from X to Y such that: $\forall x \in X,\ \exists y \in Y,\ (x, y) \in F$. That is, a correspondence assigns not-necessarily single elements Y to every element in X; it is a generalisation of a function by removing the uniqueness requirement. A correspondence F from X to Y is denoted

$$F : X \rightrightarrows Y.$$

We will study correspondences in detail in ECON 6170.

1.5 Fields

The goal now is to formalise the idea of "numbers". We want to allow for number system that allows basic operations like addition and multiplication. We may define additions and multiplications as *binary operations*, which are functions from $X \times X$ to X.

A *field* is a 3-tuple $(X, +, \cdot)$, where X is a set and $+, \cdot : X \times X \to X$ that satisfy the following properties:

- (*Associativity*) $(x + y) + z = x + (y + z)$ and $x \cdot (y \cdot z) = (x \cdot y) \cdot z\ \forall x, y, z \in X$;
- (*Commutativity*) $x + y = y + x$ and $x \cdot y = y \cdot x\ \forall x, y \in X$;
- (*Distributivity*) $x \cdot (y + z) = x \cdot y + x \cdot z\ \forall x, y, z \in X$;
- (*Existence of identity elements*) $\exists! 0, 1 \in X,\ x + 0 = 0 + x = x$ and $x \cdot 1 = 1 \cdot x = x\ \forall x \in X$;
- (*Existence of inverse elements*) $\forall x \in X,\ \exists!(-x) \in X,\ x + (-x) = (-x) + x = 0$ and $\forall x \in X \backslash \{0\},\ \exists! x^{-1}.\ x \cdot x^{-1} = x^{-1} \cdot x = 1$.

Given a field $(X, +, \cdot)$, $+$ is the *addition* operation and $\cdot$ is the *multiplication* operation. Given a field $(X, +, \cdot)$, we can define the *subtraction* operation as $- : X \times X \to X$ such that $x - y \equiv -(x, y) \coloneqq x + (-y)$ and the *division* operation as $\div : X \times X \to X$ such that $\frac{x}{y} \equiv \div(x, y) \coloneqq x \cdot y^{-1}$.[7] We often simply say that X is a field without specifying $+$ and $\cdot$.

Note that $\mathbb{N}$ is *closed* under the operations of addition and multiplication; i.e., the sum and the product of any two natural numbers is a natural number. However, $\mathbb{N}$ is not closed under subtraction and division (examples?). $\mathbb{Z}$ (unlike the natural numbers) is closed under subtraction, but not division (check). Finally, the set of rational numbers, $\mathbb{Q}$ is closed under all four operations.

[7] Strictly speaking, $\div$ is not a binary operation since $1/0$ is not defined in X.

Proposition 8. *$\mathbb{R}$ and $\mathbb{Q}$ are fields while $\mathbb{N}$ and $\mathbb{Z}$ are not.*

An *ordered field* is a 4-tuple $(X, +, \cdot, \geq)$ such that $(X, +, \cdot)$ is a field and $\geq$ is a partial order on X that is compatible with the operations $+$ and $\cdot$ meaning

$$x \geq y \Rightarrow x + z \geq y + z \ \forall x, y, z \in X,$$
$$x \cdot z \geq y \cdot z \ \forall x, y, z \in X : z \geq 0.$$

Given an ordered field, we define

$$X_+ \coloneqq \{x \in X : x \geq 0\}, \quad X_- \coloneqq \{x \in X : x \leq 0\},$$
$$X_{++} \coloneqq \{x \in X : x > 0\}, \quad X_{--} \coloneqq \{x \in X : x < 0\}.$$

In particular, we write $\mathbb{R}^n_+$ to denote nonnegative real vectors with $n \in \mathbb{N}$ components, and $\mathbb{R}^n_{++}$ to denote the set of vectors with $n \in \mathbb{N}$ components whose individual components are all strictly positive.

Let $(X, +, \cdot, \geq)$ be an ordered field. Define the *absolute value* function, $|\cdot| : X \to X$, as

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

The following proposition tells us that the Triangle Inequality is derived from the properties of an ordered field.

Proposition 9 (Triangle Inequality)**.** *Let $(X, +, \cdot, \geq)$ be an ordered field. Then,*

$$|x + y| \leq |x| + |y| \ \ \forall x, y \in X. \tag{1.3}$$

Lemma 2. *For any $x, y, z \in X$, (i) $x + y = x + z \Leftrightarrow y = z$; (ii) $-(-x) = x$; and (iii) $-(x + y) = -x + -y$.*

Proof. (i) For any $w \in X$, we have $w = 0 + w = (-x + x) + w = -x + (x + w)$. Letting $w = y$, if $x + y = x + z$, we obtain

$$y = -x + (x + y) = -x + (x + z) = (-x + x) + z = 0 + z = z.$$

For the converse, follow the steps in reverse.

(ii)

$$x = x + 0 = x + (-x - (-x)) = (x + -x) + - (-x) = 0 + - (-x).$$

(iii) Since the additive inverse of $x + y$ is unique, the claim follows since

$$\begin{aligned} (x + y) + (-x + -y) &= (x + y) + (-y + -x) \\ &= x + (y + (-y + -x)) \\ &= x + ((y + -y) + -x) \\ &= x + (0 + -x) \\ &= x + -x = 0. \end{aligned}$$

■

Proof of Proposition 9. We first show that (a) $|x| \geq x$. If $x \geq 0$, then $|x| = x$ by definition. If $x < 0$, then

$$|x| = -x \equiv 0 + -x \geq x + -x \equiv 0 \geq x.$$

Next, we show that (b) $x \geq 0 \Rightarrow -x \leq 0$ and $x \leq 0 \Rightarrow -x \geq 0$. Suppose $x \geq 0$, then

$$0 \equiv x + (-x) \geq 0 + (-x) \equiv -x.$$

Suppose instead $x \leq 0$, then

$$0 \equiv x + (-x) \leq 0 + (-x) \equiv -x \Leftrightarrow -x \geq 0.$$

Next, we show that (c) $x \geq -|x|$. If $x \geq 0$ so that $x = |x|$, then $x \geq 0 \geq -x = -|x|$, where the second inequality follows from (b). If $x < 0$, then $|x| = -x$ so that $x = -(-x) = -|x|$, where we used (ii) from Lemma 2.

We now show that (d) $x \geq y \Rightarrow -y \geq -x$.

$$\begin{aligned} x = x + 0 = x + (-y + y) &\geq y + (x + -x) = y + 0 = y \\ \Leftrightarrow (x + y) - y &\geq (x + y) + -x = (y + x) + -x \\ \Leftrightarrow -y &\geq -x, \end{aligned}$$

where, in the last line, we used the (i) from Lemma 2.

We are now ready to prove the triangle inequality. It suffices to show that

$$|x| + |y| \geq x + y \text{ and } |x| + |y| \geq -(x + y).$$

By applying (a) twice,

$$|x| + |y| \geq x + |y| \geq x + y.$$

Apply (c) twice gives

$$x + y \geq -|x| + y \geq -|x| + -|y| = -(|x| + |y|),$$

where we used (iii) from Lemma 2 in the last equality. By (d) and (ii) from Lemma 2, we have

$$x + y \geq -(|x| + |y|) \Rightarrow |x| + |y| = -(-(|x| + |y|)) \geq -(x + y). \qquad \blacksquare$$

Corollary 1. *Let $(X, +, \cdot, \geq)$ be an ordered field. Then, for any $n \in \mathbb{N}$,*

$$|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n| \ \ \forall x_1, \ldots, x_n \in X.$$

Proof. The inequality trivially holds if $n = 1$. For $n \geq 2$, we prove this by induction on n. For $n = 2$, we have already shown that the inequality holds. We now make the inductive hypothesis that the inequality holds some arbitrary $n > 2$. Our goal is to to show that the expression holds for $n + 1$. Define $z_n \coloneqq x_1 + \cdots + x_n$. By Proposition 9, since $z_n, x_{n+1} \in X$,

$$|z_n + x_{n+1}| \leq |z_n| + |x_{n+1}|.$$

By the induction hypothesis, $|z_n| \leq |x_1| + \cdots + |x_n|$ so that

$$|z_n| + |x_{n+1}| \leq |x_1| + \cdots + |x_n| + |x_{n+1}|.$$

Combining, we have that

$$|x_1 + \cdots + x_n + x_{n+1}| = |z_n + x_{n+1}| \leq |x_1| + \cdots + |x_n| + |x_{n+1}|.$$

The desired result then follows by the Principle of Mathematical Induction.

We also have the following (the second is called the *reverse triangle inequality*). ■

Proposition 10. *Let $(X, +, \cdot, \geq)$ be an ordered field. Then, $\forall x, y \in X$, (i) $|x \cdot y| = |x| \cdot |y|$; and (ii) $|x - y| \geq ||x| - |y||$.*

We can define operations on correspondences too. For example, the *sumset* of two sets X and Y is defined as

$$X + Y \coloneqq \{x + y : (x, y) \in X \times Y\}.$$

1.5.1 Real numbers, $\mathbb{R}$

Recall the various sets of numbers:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C},$$

where $\mathbb{C}$ is the set of complex numbers, $\mathbb{R}$ is the set of real numbers, $\mathbb{Q}$ is the set of rational numbers (i.e., numbers that can be expressed as fractions), $\mathbb{Z}$ is the set of integers, and $\mathbb{N}$ is the set of natural numbers.[8] You may know that there is a sense in which $\mathbb{Q}$ has "holes" and the set of real number, $\mathbb{R}$, is obtained by filling in these holes. We will not attempt to construct numbers formally, and only note that the crucial difference between $\mathbb{Q}$ and $\mathbb{R}$ is the following.

Axiom 1 (Completeness Axiom)**.** *Every nonempty subset S of $\mathbb{R}$ that is bounded from above (resp. below) has a supremum (resp. infimum) in $\mathbb{R}$.*

Given $X \subseteq \mathbb{R}$, if X is not bounded above, we write $\sup X = \infty$. If X is not bounded below, we write $\inf X = -\infty$.

We have already noted that, $(\mathbb{R}, \geq)$ is a total order and that $(\mathbb{R}, +, \cdot, \geq)$ is an ordered field. Thus, above tells us that $\mathbb{R}$ is a complete ordered field.

Proposition 11 (The Archimedean Property)**.** *For any $(x, y) \in \mathbb{R} \times \mathbb{R}_{++}$, there exists an $n \in \mathbb{N}$ such that $x < n \cdot y$.*

Proof. By way of contradiction, suppose that the claim is false; i.e., $\exists (x, y) \in \mathbb{R} \times \mathbb{R}_{++}$ such that $x \geq n \cdot y$ for all $n \in \mathbb{N}$. In other words, there must exist $y \in \mathbb{R}_{++}$ such that $\{n \cdot y : n \in \mathbb{N}\}$ is bounded from above (by x). But then $s \coloneqq \sup\{n \cdot y : n \in \mathbb{N}\}$ would be a real number by the Completeness Axiom. Since $y > 0$ and s is the least upper bound, it must be that $s - y < s$ is not an upper bound of $\{n \cdot y : n \in \mathbb{N}\}$; i.e., there exists an $n^* \in \mathbb{N}$ such that $s - y < n^* \cdot y \Leftrightarrow s < (n^* + 1) \cdot y$, which contradicts the choice of s as an upper bound. ■

We often use the Archimedean Property reformulated (check) as follows:

$$\forall \epsilon > 0,\ \exists n \in \mathbb{N},\ \frac{1}{n} < \epsilon.$$

We say that $E \subseteq \mathbb{R}$ is *dense* in $\mathbb{R}$, if an element of E lies between any two real numbers. The Archimedean Property is used to establish that rational numbers are dense in real numbers.

Corollary 2. *For any $x, y \in \mathbb{R}$ such that $x < y$, there exists $q \in \mathbb{Q}$ such that $x < q < y$.*

Proof. Fix $x, y \in \mathbb{R}$ such that $y > x$. By the Archimedean Property (Proposition 11)—setting $y = y - x > 0$ and $x = 1$—there exists an $n \in \mathbb{N}$ such that $n(y - x) > 1 \Leftrightarrow ny > nx + 1$. Let $m \coloneqq \min\{k \in \mathbb{Z} : k > na\}$.[9] By definition, $na < m$ and $na \geq m - 1$ (why?) and so $na < m \leq 1 + na < nb$. Letting $q \coloneqq \frac{m}{n}$ and noting that q is rational completes the proof. ■

[8] I will use the convention that $0 \notin \mathbb{N} \equiv \{1, 2, \ldots\}$.

[9] One way to formally prove the existence of m is to prove that every nonempty subset of $\mathbb{N}$ that is bounded from below has a minimum.

Exercise 9. Let $x, y \in \mathbb{R}$. Prove that, for all $\epsilon > 0$, $x \leq y + \epsilon \Rightarrow x \leq y$.

Exercise 10. Let $x, y \in \mathbb{R}$. Prove that, $|x - y| \leq \epsilon$ for all $\epsilon > 0$ implies $x = y$.

1.5.2 Extended real numbers

The *extended real numbers* is the set that contains $\mathbb{R}$, $-\infty$ and ∞, denoted

$$\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, \infty\}.$$

We extend the total order $\geq$ of $\mathbb{R}$ to $\overline{\mathbb{R}}$ by letting $\infty > -\infty$ and $\infty > x > -\infty$ for all $x \in \mathbb{R}$. Thus, $(\overline{\mathbb{R}}, \geq)$ is a totally ordered set. The benefit from doing so is that every subset S in $\overline{\mathbb{R}}$ has an infimum and a supremum. We can extend the standard operations of addition and multiplication to $\overline{\mathbb{R}}$: for any $x \in \mathbb{R}$,

$$\begin{aligned} x + \infty &\coloneqq \infty =: \infty + x, & \infty + \infty &\coloneqq \infty, \\ x - \infty &\coloneqq -\infty =: -\infty + x, & -\infty + -\infty &\coloneqq -\infty, \\ x \cdot \infty &\coloneqq \begin{cases} \infty & \text{if } 0 < x \leq \infty \\ -\infty & \text{if } -\infty \leq x < 0 \end{cases} =: \infty \cdot x, \\ x \cdot (-\infty) &\coloneqq \begin{cases} -\infty & \text{if } 0 < x \leq \infty \\ \infty & \text{if } -\infty \leq x < 0 \end{cases} =: (-\infty) \cdot x. \end{aligned}$$

Note that $+\infty$ and $-\infty$ are *not* real numbers. Thus, statements on real numbers do not (automatically) extend to them. Plausible facts like $x + \infty = \infty$, $(-\infty) + (-\infty) = -\infty$, etc. are true in $\overline{\mathbb{R}}$. However, expressions like $+\infty + (-\infty)$, $\infty \cdot 0$, etc. are left undefined (just like $1/0$ is undefined in $\mathbb{R}$). Consequently, $\overline{\mathbb{R}}$ is not a field.

1.5.3 Intervals

For any $x, y \in \mathbb{R}$, with $x < y$, we can define the following intervals:

$$\begin{aligned} (x, y) \{z \in \mathbb{R} : x < z < y\}, \qquad (x, y] &\coloneqq (x, y) \cup \{y\}, \\ [x, y] \coloneqq \{z \in \mathbb{R} : x \leq z \leq y\}, \; [x, y) &\coloneqq \{x\} \cup (x, y). \end{aligned}$$

We refer to (x, y) as an *open interval*, $(x, y]$ and $[x, y)$ as *semi-open* (or *half-open*) intervals, and $[x, y]$ as a *closed* interval. A set $S \subseteq \mathbb{R}$ is *open* if, for each $x \in S$, there exists $r > 0$ such that $(x - r, x + r) \subseteq S$. We will generalise these concepts when we go over topology.

Real intervals are *bounded* and of *length* given by the difference. An interval is *nondegenerate* if $y - x > 0$. We assume that intervals are nondegenerate whenever we write (x, y), $(x, y]$ or $[x, y)$. Unbounded intervals are defined as

$$\begin{aligned} (x, \infty) \coloneqq \{z \in \mathbb{R} : z > x\}, \;\; [x, \infty) &\coloneqq \{x\} \cup (x, \infty), \\ (-\infty, y) \coloneqq \{z \in \mathbb{R} : z < y\}, (-\infty, y] &\coloneqq (-\infty, y) \cup \{y\}. \end{aligned}$$

Note that

$$\begin{aligned} \sup(-\infty, y) = \sup(x, y) = \sup(x, y] = y, \\ \inf(x, \infty) = \inf(x, y) = \inf[x, y) = x. \end{aligned}$$

The Completeness Axiom says that every nonempty subset S of $\mathbb{R}$ that fits in an interval of finite

length has both an infimum and a supremum. Conversely, if S does not fit in any interval of the form $(-\infty, b)$, then $\sup S$ does not exist and we write $\sup S = \infty$. Similarly, if S does not fit in the interval of the form (x, ∞), then $\inf S$ does not exist and we write $\inf S = -\infty$.

As we will study more in ECON 6701, an interval that is both closed and bounded are *compact*.

1.5.4 $\mathbb{R}^n$

Consider $\mathbb{R}^2$ and define the additive and multiplicative operator as

$$\begin{aligned} x + y \equiv + (x, y) &\coloneqq (x_1 + y_1, x_2 + y_2), \\ x \cdot y \equiv \cdot (x, y) &\coloneqq (x_1 y_1 - x_2 y_2, x_1 y_2 + x_2 y_1), \end{aligned}$$

we can show that $(\mathbb{R}^2, +, \cdot)$ is a field, with the additive identity $(0, 0)$, the multiplicative identity $(1, 0)$, the additive inverse (of x) as $(-x_1, -x_2)$, and the multiplicative inverse (of x) $((x_1 + x_2 \cdot x_2 \cdot x_1^{-1})^{-1}, -x_2 x_1^{-1}(x_1 + x_2 \cdot x_2 \cdot x_1^{-1})^{-1})$.[10] Note that this allows complex numbers, $\mathbb{C}$, to be a field.

However, it turns out that generalisation of the example above is not possible. We will simply state the following result.

Proposition 12. *$\mathbb{R}^n$ cannot be a field for any $n > 2$ with $n \in \mathbb{N}$.*

This fact motivates us to generalise fields to linear spaces so that we can deal with $\mathbb{R}^n$.

1.6 Complex numbers

While it is somewhat rare, complex numbers do pop up in economics such as when studying the stability of dynamic systems in macroeconomics and in econometrics. Let us go over complex numbers briefly.

Consider the equation $x^2 + 1 = 0$ which has no solutions in real numbers. We know that a quadratic equations generally have 0, 1 or 2 solutions. What complex numbers allows us to do (among other things!) is to think that quadratic equations *always* have two solutions. We do so by extending the of real numbers. Let i denote the solution to $i^2 = -1$. With this notation, it is clear that the equation $x^2 + 1 = 0$ has exactly two solutions, $x = i$ and $x = -i$. More generally, a *complex number* is a number of the form $a + ib$, for any $a, b \in \mathbb{R}$. The set of all complex numbers if denoted $\mathbb{C}$.

Two complex numbers are added or multiplied term by term using the usual rules of algebra, remembering that $i^2 = -1$.

Example 6. Let $z \coloneqq a + bi$ and $w \coloneqq c + di$. Then,

$$\begin{aligned} z + w &= (a + c) + (b + d)\, i, \\ -w &= -c - di \\ z - w &= (a - c) + (b - d)\, i \\ zw &= (a + bi)\,(c + di) = ac + adi + bci + bdi^2 \\ &= (ac - bd) + (adi + bc), \end{aligned}$$

[10] The multiplicative operator cannot be defined as a dot product operator (i.e., $x \cdot y \neq x_1 y_1 + x_2 y_2$) since the result would be in $\mathbb{R}$ and not $\mathbb{R}^2$ (recall multiplicative operator is a function from $\mathbb{R}^2 \times \mathbb{R}^2$ to $\mathbb{R}^2$). Similarly, $x \cdot y = (x_1 \cdot y_1, x_2 \cdot y_2)$ is not a valid operator since $(0, 1)$ would not have a multiplicative inverse. To see this, first observe that the multiplicative identity must be $(1, 1)$. Then, for (y_1, y_2) to be the multiplicative inverse of $(0, 1)$, it implies a contradiction since

$$(1, 1) = (0, 1) \cdot (y_1, y_2) = (0, y_2) \Rightarrow 1 = 0.$$

and finally

$$z = w \Leftrightarrow (a = c) \wedge (b = d)\,.$$

Every complex number $z = a + bi$ can be represented by a point (a, b) in the (x, y) plane called *complex plane*, where the x-axis corresponds to the real part (a) and y-axis corresponds to the imaginary part (b). The point (a, b) can be expressed in *polar coordinates*; i.e., by a pair $\langle r, \theta\rangle$, where r is the distance from the origin to the point (a, b) and θ is the angle, measured in radians, relative to the x-axis. Recalling some trigonometry gives that

$$a = r\cos\theta \text{ and } b = r\sin\theta.$$

Note that

$$a^2 + b^2 = r^2\cos^2\theta + r^2\sin^2\theta = r^2\underbrace{\left(\cos^2\theta + \sin^2\theta\right)}_{=1} = r^2$$

and so $r = \sqrt{a^2 + b^2}$. Moreover,

$$\frac{b}{a} = \frac{r\sin\theta}{r\cos\theta} = \tan\theta \Leftrightarrow \theta = \arctan\left(\frac{b}{a}\right).$$

Since $\theta \equiv \theta + 2k\pi$ for any $k \in \mathbb{Z}$, we usually define the *principal value* of the argument to lie in the range $0 \leq \theta < 2\pi$. Using polar coordinates, one can show that, given $z_1 = \langle r_1, \theta_1\rangle$ and $z_2 = \langle r_2, \theta_2\rangle$

$$\begin{aligned} z_1 z_2 &= \langle r_1 r_2, \theta_1 + \theta_2\rangle \\ z_1^{-m} &= \left\langle r_1^{-m}, -m\theta_1\right\rangle \ \forall m \in \mathbb{N}, \\ \frac{z_1}{z_2} &= \left\langle \frac{r_1}{r_2}, \theta_1 - \theta_2\right\rangle. \end{aligned}$$

Note that (these are called Maclauren series—infinite Taylor expansions)

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \cdots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots, \\ \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots. \end{aligned}$$

For $z \in \mathbb{C}$, define

$$e^z \coloneqq 1 + z + \frac{z^2}{2!} + \cdots.$$

If we now write $z = i\theta$ (where $\theta \in \mathbb{R}$), then

$$e^{i\theta} = 1 + i\theta + \frac{(i\theta)^2}{2!} + \cdots = \cos\theta + i\sin\theta,$$

which is known as the *Euler's formula*. In this way, any complex number $z = r(\cos\theta + \sin\theta)$ can be written as $z = re^{i\theta}$ which allows us to use the usual laws of algebra to obtain the rules of algebra above (with polar coordinates).

2 Structures on spaces

A space is another word for a set, which itself is just a collection of objects. Here, we will introduce various *structures* we can place on sets. We will first introduce the concept of a linear structure and study *linear spaces* (sometimes called *vector spaces*). We also briefly introduce a metric structure to introduce distance between objects into a space (not necessarily linear). We then explore further structure that we can place on linear spaces: a norm structure, which gives us the concept of size of objectives, and an inner product structure, which gives us the concept of angles. We will find that every inner product space is a normed space, which, in turn, is a metric space. Note that a metric space need not be linear while normed and an inner product spaces are defined on a linear spaces. Finally, we introduce topology, which gives us a way to think about neighbourhoods of elements of the space.

2.1 Linear space

We have already seen that $\mathbb{R}^n$ can be endowed with an *order structure* so that any two elements can be meaningfully compared. We also saw that $\mathbb{R}$ has a "linear" (i.e., algebraic) structure meaning that we can add and multiply any two elements.[11] However, we saw that $\mathbb{R}^n$ is not a field in general which makes us wonder whether $\mathbb{R}^n$ has a linear structure. We will see here that $\mathbb{R}^n$ is also endowed with a linear structure in the sense that we can "add" and "scale" any two elements. We will formalise the "linearity" using the concept of a linear space (of which $\mathbb{R}^n$ would be an example).

A *linear space* (or a *vector space*) over a field F is a 4-tuple $(V, (F, +_F, \cdot_F), +, \cdot)$, where V is the set of elements called *vectors*, $(F, +_F, \cdot_F)$ is a field, a binary operation $+ : V \times V \to V$ called *vector addition*, and a binary operation $\cdot : F \times V \to V$ called *scalar multiplication* such that

- (*Closure*) $\mathbf{x} + \mathbf{y} \in V \ \forall \mathbf{x}, \mathbf{y} \in V$ and $\lambda \cdot \mathbf{x} \in V \ \forall \mathbf{x} \in V \ \forall \lambda \in F$;
- (*Associativity*) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}) \ \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, and $(\lambda \cdot_F \beta) \cdot \mathbf{x} = \lambda \cdot (\beta \cdot \mathbf{x}) \ \forall \lambda, \beta \in F \ \forall \mathbf{x} \in V$;
- (*Commutativity*) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \ \forall \mathbf{x}, \mathbf{y} \in V$;
- (*Distributivity*) $\lambda \cdot (\mathbf{y} + \mathbf{z}) = \lambda \cdot \mathbf{y} + \lambda \cdot \mathbf{z} \ \forall \mathbf{x}, \mathbf{y} \in V \ \forall \lambda \in F$, and $(\lambda +_F \beta) \cdot \mathbf{x} = \lambda \cdot \mathbf{x} + \beta \cdot \mathbf{x} \ \forall \lambda, \beta \in F \ \forall \mathbf{x} \in V$;
- (*Existence of identity*) $\exists! \mathbf{0} \in V$, $\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x}$ and $\exists 1 \in F$, $1 \cdot \mathbf{x} = \mathbf{x} \ \forall \mathbf{x} \in V$;
- (*Existence of additive inverse*) $\forall \mathbf{x} \in V$, $\exists! (-\mathbf{x}) \in V$, $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.

We express elements of linear space V in bold,[12] and say that V is a linear space without specifying F, $+_F$ and $\cdot_F$. We also often do not distinguish between $+_F$ and $+$ or $\cdot_F$ and $\cdot$. We refer to $\mathbf{0}$ as the *origin* or the *zero* vector, and refer to any $\mathbf{x} \in V \backslash \{\mathbf{0}\}$ as a *nonzero vector*.

We can define *vector subtraction* as $- : V \times V \to V$ such that $\mathbf{x} - \mathbf{y} \equiv -(\mathbf{x}, \mathbf{y}) \coloneqq \mathbf{x} + (-\mathbf{y})$ and scalar division as $\frac{\cdot}{\cdot} : F \times V \to V$ such that $\frac{\mathbf{x}}{\lambda} \equiv \frac{\cdot}{\cdot}(\lambda, \mathbf{x}) \coloneqq \lambda^{-1} \cdot \mathbf{x}$. Note also that: (i) $\mathbf{0}$ is unique (why?); (ii) $(-\mathbf{x})$ is unique for all $\mathbf{x} \in V$; (iii) $-\mathbf{x} = (-1) \cdot \mathbf{x} \ \forall \mathbf{x} \in V$; and (iv) $0 \cdot \mathbf{x} = \mathbf{0} \ \forall \mathbf{x} \in V$ and $\lambda \cdot \mathbf{0} = \mathbf{0} \ \forall \lambda \in F$.

Example 7. A trivial example of a linear space is the set $\{\mathbf{0}\}$, which is a single set consisting of the origin. Any linear space that contains more than one vector is *nontrivial*.

[11] By linear, we mean that we can add and scale elements.
[12] You might see $\vec{x}$ instead of $\mathbf{x}$.

Example 8. A familiar example of a linear space is the Euclidean space, $(\mathbb{R}^n, +, \cdot)$, where the vector addition and scalar multiplication are defined in a component-by-component manner. i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$,

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) \equiv \mathbf{x} + \mathbf{y} \coloneqq (x_1 + y_1, \ldots, x_n + y_n)\,,$$
$$\lambda \cdot (x_1, \ldots, x_n) \equiv \lambda \cdot \mathbf{x} \coloneqq (\lambda \cdot x_1, \ldots, \lambda \cdot x_n)\,.$$

However, note that $\mathbb{R}^n_{++}$ is not a linear space (does not contain the origin) nor is $\mathbb{R}^n_{+}$ a linear space (since, under the usual operations, it does not contain additive inverse for nonzero vectors).

Example 9. Given a nonempty set X, and the set of all real-valued functions on X,

$$V \coloneqq \left\{ f : f \in \mathbb{R}^X \right\},$$

is a linear space with

$$(f + g)(x) \coloneqq f(x) + g(x)\,,$$
$$(\lambda f)(x) \coloneqq \lambda \cdot f(x)\,,$$
$$(\mathbf{0})(x) \coloneqq 0,$$
$$(-f)(x) \coloneqq -f(x)\,.$$

We will see that sequences and and matrices are both real-valued functions, it follows that the set of all sequences and the set of all matrices are both linear spaces.

Given a linear space $(V, (F, +_F, \cdot_F), +, \cdot)$, a (resp. *proper*) subset S of V is a (resp. *proper*) *linear subspace* of $(V, (F, +_F, \cdot_F), +, \cdot)$ if $(S, (F, +_F, \cdot_F), +|_S, \cdot|_S)$ is a linear space. For an example of $S \subseteq V$ that is not a linear subspace, let $V = \mathbb{R}^n$ and S be any subset such that $\mathbf{0} \notin S$. The following gives us a quick way to check whether a subspace is linear.

Proposition 13. *Let $(V, (F, +_F, \cdot_F), +, \cdot)$ be a linear space and S be a nonempty (resp. proper) subset of V. Then, $(S, (F, +_F, \cdot_F), +|_S, \cdot|_S)$ is a (resp. proper) linear subspace of V if and only if $\alpha \cdot \mathbf{x} + \beta \cdot \mathbf{y} \in S$ $\forall \alpha, \beta \in F$ $\forall \mathbf{x}, \mathbf{y} \in S$.*

From the proposition above, it is immediate that $[0, 1]$ is not a linear subspace of $\mathbb{R}$ (why?) whereas $\{\mathbf{x} \in \mathbb{R}^2 : x_1 + x_2 = 0\}$ is a linear subspace of $\mathbb{R}^2$.

Given Y and Z that are subspaces of a linear space $(V, (F, +_F, \cdot_F), +, \cdot)$, the *sum space* of Y and Z, denoted $Y + Z$, is given by

$$Y + Z \coloneqq \{\mathbf{y} + \mathbf{z} : \mathbf{y} \in Y,\ \mathbf{z} \in Z\}\,.$$

Further, if $Y \cap Z = \{\mathbf{0}\}$, then the sum space is called a *direct sum of Y and Z* and is denoted $Y \oplus Z$. Any $\mathbf{x} \in Y \oplus Z$ has a unique representation as $\mathbf{x} = \mathbf{y} + \mathbf{z}$ where $\mathbf{y} \in Y$ and $\mathbf{z} \in Z$.

Proposition 14. *Intersections and sum space of linear subspaces are a linear subspaces.*

Exercise 11. Show that a union of linear subspaces need not be a linear subspace.

From now on, unless otherwise specified, assume if $F = \mathbb{R}$, $+_{\mathbb{R}}$ and $\cdot_{\mathbb{R}}$ are the usual addition and multiplication operators on $\mathbb{R}$. Whenever the underlying field of a linear space is $\mathbb{R}$ (i.e., a linear space is over $\mathbb{R}$), we will simply write $(V, +, \cdot)$.

2.1.1 Linear, affine and convex combinations

Given a linear space $(V, +, \cdot)$ and $(\mathbf{x}_i) \in V^\infty$, define, $\forall n \in \mathbb{N}$,

$$\sum_{i=1}^{n} \mathbf{x}_i \coloneqq \mathbf{x}_1 + \cdots + \mathbf{x}_n, \ \sum_{i=k}^{n} \mathbf{x}_i \coloneqq \sum_{i=1}^{n-k+1} \mathbf{x}_{i+k-1}.$$

Given a linear space $(V, +, \cdot)$ and vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in V$ for some $n \in \mathbb{N}$, a *linear combination of* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a vector $\lambda_1 \cdot \mathbf{x}_1 + \cdots + \lambda_n \cdot \mathbf{x}_n$ for some *coefficients* of linear combination $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. An *affine combination* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for some coefficients $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ (i.e., weights can be negative) such that $\sum_{i=1}^n \lambda_i = 1$. A *convex combination* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for some coefficients $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_+$ (weights must be nonnegative) such that $\sum_{i=1}^n \lambda_i = 1$.

2.1.2 Span, affine, and convex hull

Let $(V, +, \cdot)$ be a linear space. Fix a nonempty subset S of V.

The *span of* S, denoted $\operatorname{span}(S)$, is the set of all linear combinations of (finitely many) elements of S; i.e.,[13]

$$\operatorname{span}(S) \coloneqq \left\{ \sum_{i=1}^{n} \lambda_i \mathbf{x}_i : n \in \mathbb{N} \text{ and } (\mathbf{x}_i, \lambda_i) \in S \times \mathbb{R} \ \forall i \in \{1, \ldots, n\} \right\}.$$

Note that $\operatorname{span}(\varnothing) = \operatorname{span}(\{\mathbf{0}\}) = \{\mathbf{0}\}$, $\operatorname{span}(V) = V$ and $\operatorname{span}(\{x\}) = \{\lambda \cdot x : \lambda \in \mathbb{R}\}$. Say that a set S spans a set E if $E = \operatorname{span}(S)$. It follows from Proposition 13 that $\operatorname{span}(S)$ is a linear subspace (why?) and we refer to it as the *linear subspace spanned* (or *generated*) *by* S.

The *affine hull of* S, denoted $\operatorname{aff}(S)$, is the set of all affine combinations of (finitely many) elements of S; i.e.,

$$\operatorname{aff}(S) \coloneqq \left\{ \sum_{i=1}^{n} \lambda_i \mathbf{x}_i : n \in \mathbb{N} \text{ and } (\mathbf{x}_i, \lambda_i) \in S \times \mathbb{R} \ \forall i \in \{1, \ldots, n\} \text{ with } \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

Note $\operatorname{aff}(\varnothing) = \{\mathbf{0}\}$. Note that $\operatorname{aff}(S)$ need not be a linear subspace. However, for all $\mathbf{x} \in S$,

$$\begin{aligned} \operatorname{aff}(S) &= \operatorname{span}(S - \{\mathbf{x}\}) + \{\mathbf{x}\} \\ &\equiv \{\mathbf{y} + \mathbf{x} : \mathbf{y} \in \operatorname{span}(\{\mathbf{s} - \mathbf{x} : \mathbf{s} \in S\})\}. \end{aligned}$$

That is, if we translate vectors in S by a vector $-\mathbf{x} \in S$, then its span is equal to the affine hull of S translated by $-\mathbf{x}$. Since $\operatorname{span}(S - \{\mathbf{x}\})$ is a linear subspace, we also realise that $\operatorname{aff}(S)$ is a translation away from being a linear subspace.

The *convex hull of* S, denoted $\operatorname{co}(S)$, is the set of all convex combinations of (finitely many) elements of S; i.e.,

$$\operatorname{co}(S) \coloneqq \left\{ \sum_{i=1}^{n} \lambda_i \mathbf{x}_i : n \in \mathbb{N} \text{ and } (\mathbf{x}_i, \lambda_i) \in S \times \mathbb{R}_+ \ \forall i \in \{1, \ldots, n\} \text{ with } \sum_{i=1}^{n} \lambda_i = 1 \right\}.$$

[13] Equivalently, $\operatorname{span}(S)$ is the smallest linear subspace that contains S; i.e.,

$$\operatorname{span}(S) \equiv \bigcap \{W : (W \text{ is a linear subspace of } V) \wedge (S \subseteq W)\}.$$

2.1.3 Convex sets

A subset S of a linear space $(V, +, \cdot)$ is *convex* if it contains all convex combinations of pairs in S; i.e.,

$$S \subseteq V \text{ convex} \Leftrightarrow \lambda S + (1-\lambda) S = S \; \forall \lambda \in [0,1].$$

Here are some properties of convex sets.

Proposition 15. *Suppose $(S_\alpha)_{\alpha \in A}$ is a collection of convex sets in a linear space $(V, +, \cdot)$. Then, $\bigcap_{\alpha \in A} S_\alpha$ is convex.*

Proof. Take any $\mathbf{s}_1, \mathbf{s}_2 \in S \coloneqq \bigcap_{\alpha \in A} S_\alpha$ so that $\mathbf{s}_1, \mathbf{s}_2 \in S_\alpha$ for all $\alpha \in A$. Since each S_α is convex, any convex combination of $\mathbf{s}_1$ and $\mathbf{s}_2$ are also in each S_α and are also in S. ■

Exercise 12. Suppose S_1 and S_2 are both convex subsets in a linear space $(V, +, \cdot)$. Then, $S_1 + S_2$ is convex.

Proposition 16. *Let $S \subseteq X$ be a subset in a linear space $(V, +, \cdot)$. S is convex if and only if it contains all convex combinations of elements in S.*

Proof. We prove this by induction on n. For $n = 1$, then since $\mathbf{s} = 1 \cdot \mathbf{s} \in S$ for all $\mathbf{s} \in S$, convex combination of $n = 1$ element is contained in S. Make the induction hypothesis that S contains convex combinations of n elements. We wish to show that S also contains convex combinations of $n + 1$ elements. The Principle of Mathematical Induction will take care of the rest.

$$\mathbf{s} = \sum_{i=1}^{n+1} \lambda_i \mathbf{s}_i$$

be a convex combination of $n + 1$ elements in S so that $\sum_{i=1}^{n+1} \lambda_i = 1$. Since

$$\overline{\mathbf{s}} \coloneqq \sum_{i=2}^{n+1} \frac{\lambda_i}{\sum_{j=2}^{n+1} \lambda_j} \mathbf{s}_i$$

is a convex combination of n elements in S, $\overline{\mathbf{s}} \in S$ by the induction hypothesis. Then,

$$\begin{aligned} \mathbf{s} &= \lambda_1 \mathbf{s}_1 + \left(\sum_{i=2}^{n+1} \lambda_i \right) \sum_{j=2}^{n+1} \frac{\lambda_j}{\sum_{i=2}^{n+1} \lambda_i} \mathbf{s}_j \\ &= \lambda_1 \mathbf{s}_1 + \left(\sum_{i=2}^{n+1} \lambda_i \right) \overline{\mathbf{s}} = \lambda_1 \mathbf{s}_1 + (1 - \lambda_1) \overline{\mathbf{s}}. \end{aligned}$$

Thus, by definition of a convex set, $\mathbf{s} \in S$. ■

2.1.4 Algebraic and relative interior

Consider the interval $(0, 1)$ in $\mathbb{R}$ and a point $x \in (0, 1)$. Observe that some part of the line segment between x and any other point in $\mathbb{R}$ (excluding the endpoint x) is contained $(0, 1)$; i.e., for any $x \in (0, 1)$ and any $y \in \mathbb{R}$, there exists $z \in \text{co}(\{x, y\}) \backslash \{x\}$ such that $\text{co}(\{x, z\}) \subseteq (0, 1)$. In this sense, all points of $(0, 1)$ are in the "interior" of $(0, 1)$. However, we could not do the same for $(0, 1]$—the line segment between $x = 1$ and $z = 2$, say, intersects $(0, 1]$ at only at the end point 1.

To formalise this idea, let S be a subset of a linear space $(V, +, \cdot)$. An element $\mathbf{x} \in S$ is an *algebraic interior point* of S (in V) if, for any $\mathbf{y} \in V$, there exists an $\alpha_y > 0$ such that

$$(1 - \alpha) \mathbf{x} + \alpha \mathbf{y} \in S \; \forall 0 \leq \alpha \leq \alpha_y.$$

The set of all algebraic interior points of S in V is called the *algebraic interior* of S (in V) and is denoted al-int$_V(S)$. If $S \subseteq$ al-int$_V(S)$, then we say that S is *algebraically open* in V. If $\mathbf{x} \in$ al-int$_V(S)$, then one can move linearly from $\mathbf{x}$ towards any direction in the linear space V without leaving the set S immediately.

The natural "mother space" of a convex set is its affine hull. In particular, if S is a subset of a linear space V with $\mathbf{0} \in S$, then we do not need vectors in $V \backslash \text{span}(S)$. But nothing changes if we drop the assumption that $\mathbf{0} \in S$ except that we should now consider aff(S) (why?). Thus, the algebraic interior of a subset S of a linear space relative to its affine hull, aff(S), is of primary importance in convex analysis.

An element $\mathbf{x} \in S$ is an *relative interior point* of S if, for any $\mathbf{y} \in \text{aff}(S)$, there exists an $\alpha_y > 0$ such that

$$(1-\alpha)\,\mathbf{x} + \alpha\mathbf{y} \in S \;\forall 0 \leq \alpha \leq \alpha_y.$$

The set of all relative interior points of S in V is called the *relative interior* of S (in V) and is denoted ri(S). If $S \subseteq \text{ri}(S)$, then we say that S is *relatively open*. If $\mathbf{x} \in \text{ri}(S)$, then one can move from $\mathbf{x}$ towards any direction in aff(S) without leaving the set S immediately.

Note that

$$\text{al-int}_V\,(S) = \begin{cases} \text{ri}\,(S) & \text{if aff}\,(S) = V \\ \varnothing & \text{if aff}\,(S) \subset V \end{cases}.$$

The first case is obvious (is it?) and so we focus on the latter case. So suppose aff$(S) \subset V$ and take any $\mathbf{x}^* \in S$ and let $Y \coloneqq \text{span}(S - \{\mathbf{x}^*\}) = \text{aff}(S) - \{\mathbf{x}^*\}$. That aff$(S)$ is a proper subset of V means that Y is a proper linear subspace of V. Thus, we can find $\mathbf{w} \in V \backslash Y$ and notice that

$$(1-\alpha)\,\mathbf{z} + \alpha\mathbf{w} \notin Y \;\forall \mathbf{z} \in Y \;\forall \alpha \in (0,1];$$

otherwise, we would contradict $\mathbf{w} \notin Y$. But then

$$(1-\alpha)\,(\mathbf{z}+\mathbf{x}^*) + \alpha\,(\mathbf{w}+\mathbf{x}^*) \notin Y + \{\mathbf{x}^*\} = \text{aff}\,(S) \;\; \forall \mathbf{z} \in Y \;\forall \alpha \in (0,1].$$

Letting $\mathbf{y} = \mathbf{w} + \mathbf{x}^*$, therefore, we find $(1-\alpha)\mathbf{x} + \alpha\mathbf{y} \notin Y$ for any $\mathbf{x} \in S$ and any $0 < \alpha \leq 1$; i.e., al-int$_V(S) = \varnothing$.

2.1.5 Ray, half lines, cones, and conical hull

A subset of a linear space $(V, +, \cdot)$ is a *ray* if it is a set of nonnegative multiples of some vector; i.e., $\{\alpha\mathbf{x} : \alpha \in \mathbb{R}_+)$ for some $\mathbf{x} \in V$. A *half line* is a translation of a ray; i.e., $\{\alpha\mathbf{x} : \alpha \in \mathbb{R}_+\} + \{\mathbf{x}_0\}$ for some $\mathbf{x}, \mathbf{x}_0 \in V$. A *cone* is a set that contains all of its rays; i.e.,

$$S \subseteq V \text{ cone} \Leftrightarrow \alpha\mathbf{x} \in S \;\forall \alpha \in \mathbb{R}_+ \;\forall \mathbf{x} \in S.$$

A set $S \subseteq V$ is *convex cone* if it is both convex and cone.

Proposition 17. *A convex cone contains all nonnegative (finite) linear combinations of its elements.*

Proof. Let S be a convex cone in a linear space $(V, +, \cdot)$. Take an arbitrary nonnegative linear combinations of elements in S,

$$\mathbf{x} = \sum_{i=1}^{n} \lambda_i \mathbf{s}_i$$

for some $n \in \mathbb{N}$, $(\lambda_i)_{i=1}^n \in \mathbb{R}_+^n$ and $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subseteq S$. We wish to show that $\mathbf{x} \in S$. If $\lambda_1 = \cdots = \lambda_n = 0$, then $\mathbf{x} = \mathbf{0} = \lambda\mathbf{s}$ for any $\mathbf{s} \in S$; i.e., $\mathbf{x} \in S$. If $(\lambda_i)_{i=1}^n$ is a nonzero vector, then

$\lambda \coloneqq \sum_{i=1}^n \lambda_i > 0$. Then,

$$\mathbf{x} = \sum_{i=1}^n \frac{\lambda_i}{\lambda} (\lambda \mathbf{s}_i).$$

Since S is a cone, $\lambda \mathbf{s}_i \in S$ and, by construction, $\sum_{i=1}^n \lambda_i/\lambda = 1$. Hence, $\mathbf{x}$ is a convex combination of elements in S; i.e., $\mathbf{x} \in S$. ■

The *convex conical hull of* S, denoted $\operatorname{coni}(S)$, is the set of all nonnegative linear combinations of (finitely many) elements of S; i.e.,

$$\operatorname{coni}(S) \coloneqq \left\{ \sum_{i=1}^n \lambda_i \mathbf{x}_i : n \in \mathbb{N} \text{ and } (\mathbf{x}_i, \lambda_i) \in S \times \mathbb{R}_+ \ \forall i \in \{1, \ldots, n\} \right\}.$$

2.1.6 Linear dependence and independence

Fix a linear space $(V, +, \cdot)$. A subset $S \subseteq V$ is *linearly dependent in* V if it either equals $\{\mathbf{0}\}$ or at least one of the vectors, say $\mathbf{x} \in S$, can be expressed as a linear combination of finitely many vectors in $S \backslash \{\mathbf{x}\}$. For any $n \in \mathbb{N}$, any distinct vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in V$ are *linearly dependent* if $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is linearly dependent in V. The following is immediate from the definition of span.

Proposition 18. *Given a linear space $(V, +, \cdot)$, a subset S of $V \backslash \{\mathbf{0}\}$ is linearly dependent in V if and only if there exists an $\mathbf{x} \in S$ such that $\mathbf{x} \in \operatorname{span}(S \backslash \{\mathbf{x}\})$.*

A subset $S \subseteq V$ is *linearly independent* in V if no finite subset of it is linearly dependent in V. For any $n \in \mathbb{N}$, the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in V$ are *linearly independent* if $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is linearly independent in V.

Proposition 19. *Given a linear space $(V, +, \cdot)$, for any $n \in \mathbb{N}$, vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in V$ are linearly independent if and only if $\forall (\lambda_i)_{i=1}^n \in \mathbb{R}^n$,*

$$\sum_{i=1}^n \lambda_i \mathbf{x}_i = \mathbf{0} \Rightarrow \lambda_1 = \cdots = \lambda_n = 0.$$

Proof. Fix some $n \in \mathbb{N}$. (Sufficiency) Suppose that vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in V$ are linearly independent. By way of contradiction, suppose there exists $(\lambda_i)_{i=1}^n \in \mathbb{R}^n$ such that $\sum_{i=1}^n \lambda_i \mathbf{x}_i = \mathbf{0}$ but $\lambda_j \neq 0$ for some $j \in \{1, \ldots n\}$. Then,

$$\lambda_j \mathbf{x}_j = -\sum_{i \neq j} \lambda_i \mathbf{x}_i \Leftrightarrow \mathbf{x}_j = \sum_{i \neq j} \left(-\frac{\lambda_i}{\lambda_j} \right) \mathbf{x}_i.$$

That is, $\mathbf{x}_j$ can be expressed as a linear combination of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \backslash \{\mathbf{x}_j\}$, which contradicts the hypothesis that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly independent.

(Necessity) Suppose that $\forall (\lambda_i)_{i=1}^n \in \mathbb{R}^n$, $\sum_{i=1}^n \lambda_i \mathbf{x}_i = \mathbf{0}$ implies that $\lambda_1 = \cdots = \lambda_n = 0$. Toward a contradiction, suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is not linearly independent. Without loss of generality, we may assume that there exists $(\mu_i)_{i=1}^{n-1} \in \mathbb{R}^{n-1}$ such that

$$\mathbf{x}_n = \sum_{i=1}^{n-1} \mu_i \mathbf{x}_i \Leftrightarrow \mathbf{0} = \sum_{i=1}^{n-1} \mu_i \mathbf{x}_i - \mathbf{x}_n = \sum_{i=1}^{n} \mu_i \mathbf{x}_i,$$

where the last equality follows by defining $\mu_n \coloneqq -1$. But above contradicts the hypothesis that implies that $\mu_n = 0$. ■

Note that the empty set $\varnothing$ and any singleton set except for $\{\mathbf{0}\}$ are linearly independent set in any linear space.

The proposition above tells us that there is at most one way of expressing vectors as a linear combination of linearly independent vectors.

Corollary 3. *Given a linear space $(V,+,\cdot)$, suppose that $\mathbf{x}_1,\dots,\mathbf{x}_n \in V$ are linearly independent. Let $(\lambda_i)_{i=1}^n, (\mu_i)_{i=1}^n \in \mathbb{R}^n$.*

$$\sum_{i=1}^n \lambda_i \cdot \mathbf{x}_i = \sum_{i=1}^n \mu_i \cdot \mathbf{x}_i \Rightarrow \lambda_i = \mu_i \ \forall i \in \{1,\dots,n\}.$$

Proof. Follows from Proposition 19 and the observation that

$$\sum_{i=1}^n \lambda_i \cdot \mathbf{x}_i = \sum_{i=1}^n \mu_i \cdot \mathbf{x}_i \Leftrightarrow \sum_{i=1}^n (\lambda_i - \mu_i) \cdot \mathbf{x}_i = \mathbf{0}. \qquad \blacksquare$$

The following fundamental result about linear dependency is that there cannot be more than n linearly independent vectors in a linear space spanned by n vectors.

Proposition 20. *Suppose $(V,+,\cdot)$ is a linear space and $X, Y \subseteq V$. If Y is linearly independent in V and $Y \subseteq \operatorname{span}(X)$, then $|Y| \leq |X|$.*

Proof. Suppose $(V,+,\cdot)$ is a linear space and $X, Y \subseteq V$ and that Y is linearly independent in V and $Y \subseteq \operatorname{span}(X)$. If $X = \varnothing$, then $\operatorname{span}(X) = \{\mathbf{0}\}$ and so $Y \in \{\varnothing, \{\mathbf{0}\}\}$. Since Y must be linearly independent, it must be $Y = \varnothing \Rightarrow |Y| = 0$ so the claim follows. Similarly if $|X| = \infty$, then there is nothing to prove. Suppose that $n \coloneqq |X| \in \mathbb{N}$ and write $X \equiv \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$. Toward a contradiction, suppose that $|Y| > n$. Take any $\mathbf{y}_1 \in Y$. By linear independence of $Y \equiv \{\mathbf{y}_1,\dots,\mathbf{y}_m\}$ $(m > n)$, $\mathbf{y}_1 \neq \mathbf{0}$. Since $\mathbf{y}_1 \in \operatorname{span}(X)$, there exists nonzero $(\lambda_i^1)_{i=1}^n \in \mathbb{R}^n$ such that

$$\mathbf{y}_1 = \sum_{i=1}^n \lambda_i^1 \mathbf{x}_i.$$

Let λ_1^1 be the nonzero coefficient, then

$$\mathbf{x}_1 = \frac{1}{\lambda_1^1}\mathbf{y}_1 + \sum_{i=2}^n \left(-\frac{\lambda_i^1}{\lambda_1^1}\right)\mathbf{x}_i$$

so that $\mathbf{x}_1$ can be expressed as a linear combination of $\mathbf{y}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, and any linear combination of $\mathbf{x}_1,\dots,\mathbf{x}_n$ can be expressed as a linear combination of $\mathbf{y}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ by replacing $\mathbf{x}_1$ with the expression above. Thus,

$$\operatorname{span}(\{\mathbf{y}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}) = \operatorname{span}(X).$$

Next, we take $\mathbf{y}_2 \in \operatorname{span}(X)$ which is nonzero by the linear independence of Y and

$$\begin{aligned}\mathbf{y}_2 = \sum_{i=1}^n \lambda_i^2 \mathbf{x}_i &= \lambda_1^2 \left(\frac{1}{\lambda_1^1}\mathbf{y}_1 + \sum_{i=2}^n \left(-\frac{\lambda_i^1}{\lambda_1^1}\right)\mathbf{x}_i\right) + \sum_{i=2}^n \lambda_i^2 \mathbf{x}_i \\ &= \frac{\lambda_1^2}{\lambda_1^1}\mathbf{y}_1 + \sum_{i=2}^n \lambda_1^2 \left(\lambda_i^2 - \frac{\lambda_i^1}{\lambda_1^1}\right)\mathbf{x}_i\end{aligned}$$

for some nonzero $(\lambda_i^2)_{i=1}^n \in \mathbb{R}^n$. Since Y is linearly independent, it cannot be that the only nonzero coefficient is that for $\mathbf{y}_1$. Hence, we may pick a nonzero coefficient and call it the coefficient on $\mathbf{x}_2$. Rearranging the expression above shows that $\mathbf{x}_2$ can be expressed as a linear combination of

$\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n$ so that

$$\text{span}\left(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\}\right) = \text{span}\left(X\right).$$

Repeating the same process successively, we can find vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ such that

$$\text{span}\left(\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}\right) = \text{span}\left(X\right).$$

This means that, any $\mathbf{y} \in Y \backslash \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ (and such $\mathbf{y}$ exists since $|Y| > n$) can be written as a linear combination of $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$; contradicting the linear independence of Y. ■

2.1.7 Basis

We want to think of a minimum set of vectors that allows us to construct a given linear space. We call such a set a basis. Formally, a (Hamel) *basis* of a linear space $(V, +, \cdot)$ is the smallest subset of V that spans V; i.e., S is a basis of $(V, +, \cdot)$ if (i) $V = \text{span}(S)$; and (ii) $V = \text{span}(T)$ implies $T \subset S$ is false. The following proposition allows us to replace condition (ii) with linear independence, and this is the definition you often find.

Proposition 21. *A subset S of a linear space $(V, +, \cdot)$ is a basis for V if and only if S is linearly independent and $V = \text{span}(S)$.*

Proof. (Sufficiency) Suppose that $S \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is linearly independent and $V = \text{span}(S)$. To show that S is a basis for V, it suffices to show that for any $T \equiv \{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ such that $V = \text{span}(T)$, we have that $\neg(T \subset S)$. Toward a contradiction, suppose that $\mathbf{t}_i \in S$ for all $i = \{1, \ldots, m\}$ and suppose for some $\mathbf{s}_j \notin T$—without loss, we may assume that $\mathbf{s}_1 \notin T$. These imply that every $\mathbf{t}_i$ equals some $\mathbf{s}_j \neq \mathbf{s}_1$; i.e., there exists $f : \{1, \ldots, m\} \to \{2, \ldots, n\}$ such that $\mathbf{t}_i = \mathbf{s}_{f(i)}$ for all $i \in \{1, \ldots, m\}$. Since $\text{span}(S) = V = \text{span}(T)$, $\mathbf{s}_1 \in \text{span}(T)$; i.e., there exists nonzero $(\gamma_i)_{i=1}^n \in \mathbb{R}^n$ such that

$$\mathbf{s}_1 = \sum_{i=1}^{m} \gamma_i \mathbf{t}_i = \sum_{i=1}^{m} \gamma_i \mathbf{s}_{f(i)}.$$

But this implies that $\mathbf{s}_1$ is a linear combination of $\{\mathbf{s}_2, \ldots, \mathbf{s}_n\} = S \backslash \{\mathbf{s}_1\}$, which contradicts the fact that S is linearly independent.

(Necessity) Suppose that S is a basis for V. Then, by definition, $V = \text{span}(S)$ and so it remains to show that S is linearly independent. Toward a contradiction, suppose that S is linearly dependent; i.e., there exists $\mathbf{x} \in S$ such that $\mathbf{x} \in \text{span}(S \backslash \{\mathbf{x}\})$. But this implies that $\text{span}(S \backslash \{\mathbf{x}\}) = \text{span}(S) = V$ and we also have $S \backslash \{\mathbf{x}\} \subset S$; the latter contradicts condition (ii) in the definition of basis. ■

Since $V = \text{span}(S)$, any $\mathbf{x} \in V$ can be expressed as a linear combination of vectors in $S \coloneqq \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. That is,

$$\forall \mathbf{x} \in V,\ \exists (\lambda_i(\mathbf{x}))_{i=1}^n \in \mathbb{R}^n,\ \mathbf{x} = \sum_{i=1}^{n} \lambda_i(\mathbf{x}) \mathbf{s}_i.$$

We think of $(\mathbf{x})_S \coloneqq (\lambda_i(\mathbf{x}))_{i=1}^n$ as the coordinate of the vector $\mathbf{x}$ relative to the basis S. Moreover, the fact that S is linearly independent means, by Corollary 3, that coordinate of each $\mathbf{x} \in V$ is unique.

For any $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$, the ith unit vector in $\mathbb{R}^n$, denoted by, $\mathbf{e}_i$, is the n-vector whose ith term equals 1 and whose all other terms equal 0. Since $S = \{\mathbf{e}_i : i \in \{1, \ldots, n\}\}$ is linearly independent and $\text{span}(S) = \mathbb{R}^n$, it is a basis, and in fact, we refer to it as the *canonical basis* of the linear space $(\mathbb{R}^n, +, \cdot)$. However, basis of $\mathbb{R}^n$ is not unique, and, in fact, any linearly independent set of vectors of cardinality n is a basis of $\mathbb{R}^n$. Nevertheless, the following result tells

us that any basis has the same number of elements; i.e., coordinate of the vector any $\mathbf{x} \in V$ has the same number of elements.

Exercise 13. Any two bases of a finite-dimensional linear space have the same number of elements.

We define the *rank* of $S \subseteq V$, denoted $\text{rank}(S)$, as the cardinality of any one if its basis. Rank of the whole linear space is the *dimension*, denoted $\dim(V) \coloneqq \text{rank}(V)$. We say that a linear space $(V, +, \cdot)$ is $\dim(V)$ dimensional. If $\dim(V)$ is not finite, we refer to V as being *infinite dimensional* and write $\dim(X) = \infty$. The trivial linear space $\{\mathbf{0}\}$ has a dimension of 0.

Theorem 1. *Every linear space has a basis.*

Proof. When the linear space has a finite dimension, the result follows from the definition. The proof is a little more complicated when the linear space is infinite dimensional and relies on Zorn's lemma (Lemma 1). Let (*) denote the property of a set that any finite subset of itself is linearly independent. Define a partial orderer $(P, \supseteq)$ as

$$P \coloneqq \left\{ S \in 2^V : S \text{ satisfies } (*) \right\}.$$

To be able to apply Zorn's lemma, we show that every totally ordered subset of P has an upper bound (in P). Fix any $T \subseteq P$ that is a totally ordered set. Define $M \coloneqq \bigcup_{S \in T} S$. Then, M is an upper bound of T since

$$M \supseteq B \ \forall B \in T.$$

It remains to show that $M \in P$; i.e., we must verify that M satisfies (*). Take any $m_1, \ldots, m_n \in M$ for some $n \in \mathbb{N}$. By definition of M, there exists $B_1, \ldots, B_n \in T$ such that

$$m_i \in B_i \ \forall i \in \{1, \ldots, n\}.$$

Since T is a totally ordered set, there exists B_k for some $k \in \{1, \ldots, n\}$ such that

$$B_k \supseteq B_i \ \forall i \in \{1, \ldots, n\} \Rightarrow m_1, \ldots, m_n \in B_k.$$

Moreover, since $B_k \in T \subseteq P$, B_k satisfies (*). Thus, $m_1, \ldots, m_n$ are linearly independent. Since $m_1, \ldots, m_n$ was arbitrary, it follows that $M \in P$.

By Zorn's lemma, we may conclude that there exists $S \in P$ that is maximal and, by construction, S satisfies (*). To show that S is a basis, it suffices to show that any vector in V can be written as a finite linear combination of vectors in S. Fix any $\mathbf{x} \in V$ and let $\tilde{S} = S \cup \{\mathbf{x}\}$.

- Case 1: $\mathbf{x} \in S$. Then, $\mathbf{x}$ is a linear combination of elements of S trivially and we are done.
- Case 2: $\mathbf{x} \in V \backslash S$. Then, $S \subset \tilde{S}$, which implies that $\tilde{S} \notin P$ by the maximality of S. Thus, there exists a finite subset of S that is not linearly independent. By Proposition 19, this means that there exists $\mathbf{s}_1 \ldots, \mathbf{s}_n \in \tilde{S}$ such that

$$\sum_{i=1}^{n} \lambda_i \mathbf{s}_i = \mathbf{0} \tag{2.1}$$

  for some $(\lambda_i)_{i=1}^n \in \mathbb{R}^n$ with $\lambda_j \neq 0$ for some $j \in \{1, \ldots, n\}$.

  - Subcase a: $\mathbf{s}_1 \ldots, \mathbf{s}_n \in S$. This is not possible since S satisfies (*) so that (2.1) implies $\lambda_1 = \cdots = \lambda_n = 0$ by Proposition 19.
  - Subcase b: for some $\mathbf{s}_j$, $\mathbf{x} = \mathbf{s}_j \notin S$ and $\mathbf{s}_i \in S$ for all $i \in \{1, \ldots, n\} \backslash \{j\}$. Without loss of generality, suppose $j = 1$. Note that $\lambda_1 \neq 0$, otherwise, we would have $\sum_{i=2}^n \lambda_i \mathbf{s}_i =$

0 from (2.1) and Proposition 19 together with the fact that $\mathbf{s}_2, \ldots, \mathbf{s}_n \in S$ (so that $\mathbf{s}_2, \ldots, \mathbf{s}_n$ are linearly independent) implies $\lambda_2 = \cdots = \lambda_n = 0$, which contradicts the fact that $\lambda_j \neq 0$ for some $j \in \{1, \ldots, n\}$. Since $\lambda_1 \neq 0$, we may rearrange (2.1) to obtain

$$\mathbf{x} = \mathbf{s}_1 = \sum_{i=2}^{n} \left(-\frac{\lambda_i}{\lambda_1} \right) \mathbf{s}_i;$$

i.e., $\mathbf{x}$ is a linear combination of vectors in S.

Note that a Hamel basis need not be countable. ■

Corollary 4. *Let $(V, +, \cdot)$ be a finite-dimensional linear space and let $(S, +|_S, \cdot|_S)$ be its linear subspace. Then,*

$$\dim S \leq \dim V$$

and the inequality is strict if S is a proper subspace.

Proof. Suppose $S \subseteq V$. By Theorem 1, V and S have bases and denote them, respectively, by B_V and B_S. Since B_S is linearly independent by Proposition 21 and every $\mathbf{s} \in S$ can be expressed as a linear combination of B_V, we have $B_S \subseteq \text{span}(B_V)$. Then, by Proposition 20,

$$\dim(S) = |B_S| \leq |B_V| = \dim(V).$$

Now suppose that $S \subset V$. Thus, there exists $\mathbf{x} \in V$ such that $\mathbf{x} \notin S$. Letting B_S denote a basis for S. The claim is that $B_S \cup \{\mathbf{x}\}$ is linearly independent. Suppose not, then $\mathbf{x}$ must be a linear combination of vectors in B_S, which implies that $\mathbf{x} \in \text{span}(B_S) = S$, contradicting that $\mathbf{x} \notin S$. Hence, $\dim(S) = |B_S| \leq |B_V| = \dim(V)$ in this case. ■

Proposition 22. *Suppose $(V, +, \cdot)$ is a linear space with $\dim V < \infty$. Then, any linearly independent set $W \subseteq V$ with $|W| = \dim V$ is a basis of V.*

Proof. Suppose that $(V, +, \cdot)$ is a linear space with $n = \dim V < \infty$. Let $S \coloneqq \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be a basis of V which exists by Theorem 1. Let $W \coloneqq \{\mathbf{w}_1, \ldots, \mathbf{w}_n\} \subseteq V$ be a linearly independent set. By Proposition 21, it suffices to show that $\text{span}(W) = V$. We follow the argument outlined in the proof of Proposition 20. Since $\mathbf{w}_1 \in \text{span}(S) = V$, there exists nonzero $(\lambda_i^1)_{i=1}^n \in \mathbb{R}^n$ such that

$$\mathbf{w}_1 = \sum_{i=1}^{n} \lambda_i^1 \mathbf{s}_i.$$

Let λ_1^1 be the nonzero coefficient, then

$$\mathbf{s}_1 = \frac{1}{\lambda_1^1} \mathbf{w}_1 + \sum_{i=2}^{n} \left(-\frac{\lambda_i^1}{\lambda_1^1} \right) \mathbf{s}_i$$

so that $\mathbf{s}_1$ can be expressed as a linear combination of $\mathbf{w}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, and any linear combination of $\mathbf{s}_1, \ldots, \mathbf{s}_n$ can be expressed as a linear combination of $\mathbf{w}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ by replacing $\mathbf{s}_1$ with the expression above. Thus,

$$\text{span}(\{\mathbf{w}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}) = \text{span}(S) = V.$$

Next, we take $\mathbf{w}_2 \in \text{span}(X)$ which is nonzero by the linear independence of W and

$$\begin{aligned}\mathbf{w}_2 &= \sum_{i=1}^{n} \lambda_i^2 \mathbf{s}_i = \lambda_1^2 \left(\frac{1}{\lambda_1^1} \mathbf{w}_1 + \sum_{i=2}^{n} \left(-\frac{\lambda_i^1}{\lambda_1^1} \right) \mathbf{s}_i \right) + \sum_{i=2}^{n} \lambda_i^2 \mathbf{s}_i \\ &= \frac{\lambda_1^2}{\lambda_1^1} \mathbf{w}_1 + \sum_{i=2}^{n} \lambda_1^2 \left(\lambda_i^2 - \frac{\lambda_i^1}{\lambda_1^1} \right) \mathbf{s}_i\end{aligned}$$

for some nonzero $(\lambda_i^2)_{i=1}^n \in \mathbb{R}^n$. Since W is linearly independent, it cannot be that the only nonzero coefficient is that for $\mathbf{w}_1$. Hence, we may pick a nonzero coefficient and call it the coefficient on $\mathbf{s}_2$. Rearranging the expression above shows that $\mathbf{s}_2$ can be expressed as a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \mathbf{s}_3, \ldots, \mathbf{s}_n$ so that

$$\text{span}\left(\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{s}_3, \ldots, \mathbf{s}_n\}\right) = \text{span}\left(S\right) = V.$$

Repeating the same process successively, we realise that

$$\text{span}\left(W\right) = \text{span}\left(S\right) = V. \qquad \blacksquare$$

From now on, I will generally refer to a linear space $(V, +, \cdot)$ as V.

2.1.8 Linear transformations

In a linear space, we talk about linear combination of vectors. An important class of functions are ones that maps from one linear space into another linear space in a way that preserves linear combinations. We refer to them as *linear transformations* (or *linear operators* or *linear maps*). A real-valued linear transformation is a *linear functional on* V.

Formally, let $(V, +_V, \cdot_V)$ and $W = (W, +_W, \cdot_W)$ be two linear spaces. A function $L : V \to W$ is a *linear transformation* if

$$L\left(\lambda \cdot_V \mathbf{x} +_V \mathbf{y}\right) = \lambda \cdot_W L\left(\mathbf{x}\right) +_W L\left(\mathbf{y}\right) \ \ \forall \mathbf{x}, \mathbf{y} \in V \ \forall \lambda \in \mathbb{R}. \tag{2.2}$$

For brevity, we will not be explicit about the sets that $+$ and $\cdot$ operators relate. The set of all linear transformations from a linear space V to a linear space W is denoted $\mathcal{L}(V, W)$. A linear maps preserves the origin since

$$L\left(\mathbf{0}_V\right) = L\left(0 \cdot_V \mathbf{v}\right) = 0 \cdot_W L\left(\mathbf{v}\right) = \mathbf{0}_W,$$

where $\mathbf{0}_V$ and $\mathbf{0}_W$ are the zeros in linear spaces V and W respectively.

Exercise 14. Show that $L : V \to W$ is a linear transformation if and only if, for any $\mathbf{x}, \mathbf{y} \in V$ and $\lambda \in \mathbb{R}$,

$$L\left(\mathbf{x} + \mathbf{y}\right) = L\left(\mathbf{x}\right) + L\left(\mathbf{y}\right) \ \text{ and } L\left(\lambda \mathbf{x}\right) = \lambda L\left(\mathbf{x}\right).$$

Example 10. Let $V = W = \mathbb{R}^2$ and $f : V \to W$.

(i) $f(x_1, x_2) \coloneqq (x_1 + 2x_2, x_2 - x_1)$ is a linear transformation (check).

(ii) $f(x_1, x_2) \coloneqq (2, -1)$ is *not* a linear transformation because $(0, 0)$ is mapped to $(2, -1)$.

(iii) $f(x_1, x_2) \coloneqq (0, 0)$ is a linear transformation (check).

(iv) $f(x_1, x_2) \coloneqq (x_1^2, x_2)$ is not a linear transformation because

$$\begin{aligned} f(\lambda(x_1, x_2) + (y_1, y_2)) &= f(\lambda x_1 + y_1, \lambda x_2 + y_2) \\ &= \left(\lambda^2 x_1^2 + y_1^2 + 2\lambda x_1 y_1, \lambda x_2 + y_2\right) \\ &\neq \left(\lambda x_1^2 + y_1^2, \lambda x_2 + y_2\right) \\ &= \lambda\left(x_1^2, x_2\right) + \left(y_1^2, y_2\right) = \lambda f(x_1, x_2) + f(y_1, y_2). \end{aligned}$$

Example 11. Let V be a linear space of all functions with continuous first derivatives defined on $\mathbb{R}$ (why is this a linear space?), and W the linear space of all real-valued functions defined on $\mathbb{R}$. Let $D : V \to W$ be the function that maps a function $f \in V$ into its derivative; i.e.,

$$D(f) \coloneqq f'.$$

You should know that

$$D(f + g) = D(f) + D(g) \text{ and } D(\lambda f) = \lambda D(f).$$

Hence, D is a linear transformation.

Example 12. Let V be the linear space of all continuous functions defined on $\mathbb{R}$, and W be the linear space of functions with continuous first derivatives defined on $\mathbb{R}$. For any $x \in \mathbb{R}$, Let $J_x : V \to W$ be the function that maps a function $f \in V$ into the integral; i.e.,

$$J(f) \coloneqq \int_0^x f(t)\,\mathrm{d}t.$$

You should know that

$$J(f + g) \equiv \int_0^x (f(t) + g(t))\,\mathrm{d}t = \int_0^x f(t)\,\mathrm{d}t + \int_0^x g(t)\,\mathrm{d}t \equiv J(f) + J(g)$$

and

$$J(\lambda f) = \int_0^x \lambda f(t)\,\mathrm{d}t = \lambda \int_0^x f(t)\,\mathrm{d}t \equiv \lambda J(f).$$

Hence, J is a linear transformation.

The *image* of $L \in \mathcal{L}(V, W)$ is given by

$$\operatorname{im}(L) \coloneqq \{\mathbf{y} \in W : \exists \mathbf{x} \in V,\ L(\mathbf{x}) = \mathbf{y}\}.$$

As usual, it is the values in the codomain that the linear transformation L can take values in. The *null space* (or the *kernel*) is the subset of the domain that is mapped to the zero element in the codomain:

$$\operatorname{null}(L) \equiv \ker(L) \coloneqq \{\mathbf{x} \in V : L(\mathbf{x}) = \mathbf{0}\}.$$

The null space is a linear subspace of V since $L(\lambda\mathbf{x} + \mathbf{y}) = \lambda L(\mathbf{x}) + L(\mathbf{y}) = 0$ so that $\lambda\mathbf{x} + \mathbf{y} \in S$ for all $\lambda \in \mathbb{R}$ and all $\mathbf{x}, \mathbf{y} \in S$.

Exercise 15. Consider $f : \mathbb{R}^3 \to \mathbb{R}^2$ defined as $f(x_1, x_2, x_3) \coloneqq (x_1 + x_2, x_3)$. Find its image and null space.

The *rank* of L is given by

$$\operatorname{rank}(L) \coloneqq \dim(\operatorname{im}(L)).$$

The nullity of L is the rank of the null space; i.e.,

$$\text{nullity}\,(L) \coloneqq \dim\left(\text{null}\,(L)\right)$$

Proposition 23. *Let V and W be linear spaces. Then, $L \in \mathcal{L}(V,W)$ is injective if and only if* $\text{null}(L) = \{\mathbf{0}_V\}$.

Proof. Suppose that $L \in \mathcal{L}(V,W)$ is injective; i.e., every element $\mathbf{x} \in V$ is mapped to a unique element in W. As noted above, linear maps preserves the origin so $L(\mathbf{0}_V) = \mathbf{0}_W$. Thus, $\text{null}(L) = \{\mathbf{0}_V\}$.

Conversely, suppose that $\text{null}(L) = \{\mathbf{0}_V\}$. Suppose that there exist $\mathbf{x}, \mathbf{y} \in V$ such that $L(\mathbf{x}) = L(\mathbf{y})$. Since L is linear, $L(\mathbf{x}) = L(\mathbf{y})$ implies that

$$L\left(\mathbf{x} - \mathbf{y}\right) = L\left(\mathbf{x}\right) - L\left(\mathbf{y}\right) = \mathbf{0}_W.$$

Since $\text{null}(L) = \{\mathbf{0}_V\}$, it must be that $\mathbf{x} - \mathbf{y} = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$. ■

Theorem 2 (Rank-nullity theorem)**.** *Let V and W be two finite-dimensional linear space and $L \in \mathcal{L}(V,W)$. Then,*

$$\dim V = \text{nullity}\,(L) + \text{rank}\,(L)\,.$$

Corollary 5. *Suppose* $\dim V = \dim W$*. Then, $L \in \mathcal{L}(V,W)$ is surjective if and only if L is injective. In particular,* $\text{null}(L) = \{\mathbf{0}_V\}$ *if and only if L is bijective.*

Proof. Suppose $n = \dim V = \dim W$. By Proposition 23 and the definition of nullity, L is injective if and only if $\text{nullity}(L) = 0$. Then, by the Rank-nullity theorem (Theorem 2), L is injective if and only if $n = \dim V = \text{rank}(L) \equiv \dim(\text{im}(L))$. That is, the image of L is a n-dimensional subspace of the n-dimensional linear space W. But the only full-dimensional subspace of a finite-dimensional linear space is itself and so this happens if and only if the image is all of W (i.e., when $\text{im}(L) \equiv L(V) = W$). Thus, for any $\mathbf{w} \in W$, there must exist $\mathbf{x} \in V$ such that $L(\mathbf{x}) = \mathbf{w}$; i.e., L is surjective. ■

A linear transformation $L \in \mathcal{L}(V,W)$ is *invertible* if there exists $L^{-1} : W \to V$ such hat

$$L^{-1}\left(L\left(\mathbf{x}\right)\right) = \mathbf{x}\ \forall \mathbf{x} \in V,$$
$$L\left(L^{-1}\left(\mathbf{w}\right)\right) = \mathbf{w}\ \forall \mathbf{w} \in W;$$

i.e., L is invertible if it is surjective and injective.

Exercise 16. Let V and W be two linear spaces. If $L \in \mathcal{L}(V,W)$ is invertible, then $L^{-1} \in \mathcal{L}(W,V)$; i.e., L^{-1} is linear.

2.1.9 Isomorphism

Two finite-dimensional linear spaces $(V, +_V, \cdot_V)$ and $W = (W, +_W, \cdot_W)$ are *isomorphic* if there is an invertible linear transformation $L \in \mathcal{L}(V,W)$ called *isomorphism.*

Isomorphic linear spaces are essentially indistinguishable. We can think of them as relabelling of each other because L being invertible means that L is surjective and injective; i.e., every element of V is mapped to a unique element in W, and every element of W is mapped to a unique element in V.

Proposition 24. *Two finite-dimensional linear spaces $(V, +_V, \cdot_V)$ and $W = (W, +_W, \cdot_W)$ are isomorphic if and only if* $\dim V = \dim W$.

The following tells us that any finite-dimensional linear space over $\mathbb{R}$ is essentially indistinguishable from $\mathbb{R}^n$.

Theorem 3. *Every n-dimensional linear space $(V, (\mathbb{R}, +, \cdot)_{+V}, \cdot_V)$ is isomorphic to $\mathbb{R}^n$.*

Proof. Let V be an n-dimensional linear space over $\mathbb{R}$ and let $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ denote a basis of V. Recalling that S is linearly independent (Proposition 22), by Corollary 3, for each $\mathbf{x} \in V$, there exists unique $(\lambda_i)_{i=1}^n \in \mathbb{R}^n$ such that $\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{s}_i$. Define $L : V \to \mathbb{R}^n$ by

$$L(\mathbf{x}) = L\left(\sum_{i=1}^n \lambda_i \mathbf{s}_i\right) \coloneqq (\lambda_1, \ldots, \lambda_n).$$

Since $(\lambda_i)_{i=1}^n$ is unique for each $\mathbf{x} \in V$, L is injective. To show that L is surjective, take any $(\alpha_i)_{i=1}^n \in \mathbb{R}^n$, then since V is a linear space, $\sum_{i=1}^n \alpha_i \mathbf{s}_i \in V$. ■

2.1.10 Dual spaces

Let $(V, +, \cdot)$ be a linear space (over the field $\mathbb{R}$). The *dual* (linear) *space* of V is given by $(\mathcal{L}(V, \mathbb{R}), +, \cdot)$; i.e., the collection of all linear functionals on V along with addition and multiplication operator in the field $\mathbb{R}$. The dual space of V is often denoted V^*. When applied to vector spaces of functions (which are typically infinite-dimensional), dual spaces are used to describe measures and distributions. Consequently, the dual space is an important concept in (some parts of) economics.

2.1.11 Convex, concave, quasiconcave and quasiconvex functions

Let $(V, +, \cdot)$ be a linear space (over the field $\mathbb{R}$). Given a convex $X \subseteq V$, a function $f : X \to \mathbb{R}$ is

- *concave* if $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in X$ and $\alpha \in [0, 1]$.
- *convex* if $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in X$ and $\alpha \in [0, 1]$.
- *strictly concave* if $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) > \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$ for any two distinct $\mathbf{x}, \mathbf{y} \in X$ and $\alpha \in (0, 1)$.
- *strictly convex* if $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < M\alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$ for any two distinct $\mathbf{x}, \mathbf{y} \in X$ and $\alpha \in (0, 1)$.
- *quasi-concave* if $f(\alpha x + (1-\alpha)y) \geq \min\{f(x), f(y)\}$ for any $x, y \in X$ and $\alpha \in [0, 1]$.
- *quasi-convex* if $f(\alpha x + (1-\alpha)y) \leq \max\{f(x), f(y)\}$ for any $x, y \in X$ and $\alpha \in [0, 1]$.
- *strictly* quasi-concave if $f(\alpha x + (1-\alpha)y) > \min\{f(x), f(y)\}$ for all distinct $x, y \in X$ and $\alpha \in (0, 1)$.
- *strictly* quasi-convex if $f(\alpha x + (1-\alpha)y) < \max\{f(x), f(y)\}$ for all distinct $x, y \in X$ and $\alpha \in (0, 1)$.

If f is concave and convex, then f is affine. Concavity implies quasiconcavity, and convexity implies quasiconvexity. Sums of concave functions are concave and sums of convex functions are convex; however, sums of quasiconcave (resp. quasiconvex) functions need not be quasiconcave (resp. quasiconvex).

2.2 Metric spaces

A vector $\mathbf{x} \in V$ represents a point in the linear space V. Given two vectors $\mathbf{x}, \mathbf{y} \in V$, we can define a notion of distance, called a *metric*, between these two points. A *metric* on an nonempty set X is a function $\rho : X^2 \to \mathbb{R}$ that satisfies the following conditions:

- (*nonnegativity*) $\rho(x, y) \geq 0 \ \forall x, y \in X$;
- (*identity of indiscernibles*) $\rho(x, y) = 0 \Leftrightarrow x = y$;
- (*symmetry*) $\rho(x, y) = \rho(y, x) \ \forall x, y \in X$;
- (*triangle inequality*) $\rho(x, y) \leq \rho(x, z) + \rho(y, z) \ \forall x, y, z \in X$.

A *metric space* is a pair (X, ρ), where X is a nonempty set and ρ is a metric on X. Observe that metrics can be defined on any nonempty set X; in particular, X need not be a linear space.

Proposition 25. *If (X, ρ) is a metric space, then any subspace $(Y, \rho|_Y)$ with $\varnothing \neq Y \subseteq X$ is a metric space and is called a metric subspace of (X, ρ).*

Example 13 (Examples and non-examples of metric spaces)**.**

(i) Euclidean metric space. $(\mathbb{R}^n, d)$, where d is the *Euclidean distance* given by

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

(ii) Discrete metric space. (X, d_{discrete}), where $X \neq \varnothing$ and

$$d_{\text{discrete}}(x, y) = \mathbb{1}_{\{x \neq y\}}.$$

(iii) Product metric spaces. If (X, ρ_X) and (Y, ρ_Y) are metric spaces, then $(X \times Y, \rho)$ is a metric space where

$$\rho((x_1, y_1), (x_2, y_2)) \coloneqq \left[(\rho_X(x_1, x_2))^p + (\rho_Y(y_1, y_2))^p\right]^{1/p} \tag{2.3}$$

for any $p \geq 1$.

(iv) Sequence (ℓ^p) spaces. For any $0 < p \leq \infty$, ℓ^p spaces are the set of all sequences in $\mathbb{R}$ that is p-summable; i.e.,

$$\ell^p \coloneqq \left\{(x_n) \in \mathbb{R}^\infty : \sum_{n=1}^{\infty} |x_n|^p < \infty\right\} \ \forall 0 < p < \infty,$$
$$\ell^\infty \coloneqq \left\{(x_n) \in \mathbb{R}^\infty : \sup_n |x_n| < \infty\right\}.$$

In particular, we call ℓ^∞ the space of all bounded sequences. For any $0 < p < \infty$, (ℓ^p, ρ) with

$$\rho((x_n), (y_n)) \coloneqq \sum_{n=1}^{\infty} |x_n - y_n|^p$$

is a metric space. If $p = \infty$, then (ℓ^p, ρ) with $\rho((x_n), (y_n)) \coloneqq \sup_{n \in \mathbb{N}} |x_n - y_n|$ is a metric space.

(v) Bounded real-valued function space, $(\mathbf{B}(S), \rho_\infty)$, where

$$\mathbf{B}(S) \coloneqq \left\{ f \in \mathbb{R}^S : \sup\{|f(x)| : x \in S\} < \infty \right\}$$

and ρ_∞ is the *sup-metric* given by $\rho_\infty(f,g) \coloneqq \sup_{x \in S} |f(x) - g(x)|$.

(vi) Continuous real-valued function space, $(\mathbf{C}(S), \rho_\infty)$, where S is compact and

$$\mathbf{C}(S) \coloneqq \left\{ f \in \mathbb{R}^S : f \in \mathbf{B}(S) \text{ and } f \text{ is continuous on } S \right\}.$$

(vii) Variation of information. Let $P = \{P_1, P_2, \ldots, P_k\}$ and $Q = \{Q_1, Q_2, \ldots, Q_\ell\}$ be two partitions of a set X with $n \in \mathbb{N}$ elements. Then, the variation of information between P and Q is given by

$$VI(P,Q) \coloneqq -\sum_{i,j} \frac{|P_i \cap Q_j|}{n} \left[\log\left(\frac{\frac{|P_i \cap Q_j|}{n}}{\frac{|P_i|}{n}} \right) + \log\left(\frac{\frac{|P_i \cap Q_j|}{n}}{\frac{|Q_j|}{n}} \right) \right].$$

The variation of information is a metric.

(viii) The Kullback-Leibler (KL) divergence. Let P and Q denote the CDFs of two continuous random variables on $\mathbb{R}$. The KL divergence is given by

$$KL(P,Q) \coloneqq \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)} \right) \mathrm{d}x,$$

where p and q denote the probability densities of P and Q respectively. KL divergence is not a metric because it is not symmetric and does not satisfy the triangle inequality.

Proposition 26 (Equivalent metrics)**.** *Let (X, ρ_1) and (X, ρ_2) be two metric spaces. The following are equivalent.*

(i) ρ_1 *and* ρ_2 *are* equivalent;

(ii) $\exists \alpha, \beta \in \mathbb{R}_+$, $\alpha\rho_1(x,y) \leq \rho_2(x,y) \leq \beta\rho_1(x,y)$ $\forall x, y \in X$;

(iii) $\exists c > 0$ *such that* $c^{-1}\rho_1(x,y) \leq \rho_2(x,y) \leq c\rho_1(x,y)$ $\forall x, y \in X$.

Proof. Define (i) as (ii). We show equivalence between (ii) and (iii).
(ii) $\Rightarrow$ (iii). Suppose that (ii) holds and let α, β be the constant. We must have $\alpha \leq \beta$.

- Suppose $\alpha \leq \beta \leq 1$. Then, $1/\alpha \geq 1 \geq \beta$ so that we may set $c \coloneqq 1/\alpha$.
- Suppose $\alpha \leq 1 \leq \beta$. Then, set $c \coloneqq \beta/\alpha$. Observe that $c \geq \beta$ and $c^{-1} \leq \alpha$.
- Suppose $1 \leq \alpha \leq \beta$. Then, set $c \coloneqq \beta$. Then, $1/c \leq 1 \leq \alpha$.

(iii) $\Rightarrow$ (ii). Let $c > 0$ satisfy the inequality. Define $\alpha \coloneqq 1/c$ and $\beta \coloneqq c$. ■

For example, if $X = \mathbb{R}^2$, then $\rho_1(x,y) \coloneqq \max\{|x_1 - y_1|, |x_2 - y_2|\}$ is equivalent to $\rho_2(x,y) \coloneqq |(x_1, x_2) - (y_1, y_2)|$. However, these are not equivalent to the discrete metric.

Let (X, ρ_X) and (Y, ρ_Y) be two metric spaces. A function $f : X \to Y$ is an *isometry* if

$$\rho_X(z, z') = \rho_Y(f(z), f(z')) \ \forall z, z' \in X.$$

An isometric function can be thought of as relabelling of points in X . Note that isometry $f : X \to Y$ must be injective as, otherwise, two distinct points in X could be mapped to the same point which

would contradict the identity of indiscernibles property of the metric ρ_X. An isometry that is bijective is called an *isometric isomorphism.* Metric spaces X and Y are *isometric*, denoted $X = Y$, if there exists an isometric isomorphism between them.

2.3 Normed spaces

We now introduce a concept of "size" to vectors in a linear space. A *norm* on a linear space V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following conditions:

(i) (*positivity*) $\|\mathbf{x}\| \geq 0$ $\forall \mathbf{x} \in V$ and $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$;

(ii) (*homogeneity*) $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ $\forall \mathbf{x} \in V$ $\forall \alpha \in \mathbb{R}$;

(iii) (*triangle inequality*) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ $\forall \mathbf{x}, \mathbf{y} \in V$.

A *normed space* (sometimes called the *normed linear space*) is a pair $(V, \|\cdot\|)$, where V is a linear space and $\|\cdot\|$ is a norm on V. Let S be a linear subspace of V (recall Proposition 13). Then, $(S, \|\cdot\||_S)$ is a normed subspace of $(V, \|\cdot\|)$.

The *natural metric*, ρ, on a normed space $(V, \|\cdot\|)$ is defined as

$$\rho(\mathbf{x}, \mathbf{y}) \coloneqq \|\mathbf{x} - \mathbf{y}\|.$$

Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on the same linear space V are *equivalent* if the induce equivalent natural metrics. A linear transformation, $L \in \mathcal{L}(V, W)$ between two normed space $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ is *isometric* if

$$\|L(\mathbf{x})\|_W = \|\mathbf{x}\|_V \ \ \forall \mathbf{x} \in V. \tag{2.4}$$

Example 14 (Examples and non-examples of normed spaces)**.**

- Euclidean normed space is Euclidean space $\mathbb{R}^n$ together with the Euclidean norm defined on $\mathbb{R}^n$, where $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is given by

$$\|\mathbf{x}\| \coloneqq \sqrt{\sum_{i=1}^{n} x_i^2}.$$

  We can extend the Euclidean norm into any arbitrary finite-dimensional linear space.

- If $p \geq 1$, then the sequence space $(\ell^p, \|\cdot\|_p)$ is a normed space with $\|(x_n)\|_p \coloneqq (\sum_{n=1}^{\infty} |x_n|^p)^{1/p}$. Note the triangle inequality fails with $0 < p < 1$.

- C (continuous function) space: $(C(X), \|\cdot\|_{L^\infty})$, where, X is a complete metric space,[14] $C(X)$ is the set of all real-valued continuous functions on X and $\|\cdot\|_{L^\infty}$ is the sup norm (i.e., $\|f\|_{L^\infty} \coloneqq \sup\{|f(x)|\} : x \in X\}$).

- $C^{0,1}([0,1])$ (Lipschitz-continuous function on the unit interval) space: $(C^{0,1}([0,1]), \|\cdot\|_{C^{0,1}([0,1])})$, where

$$C^{0,1}([0,1]) \coloneqq \left\{ f : [0,1] \to \mathbb{R} \text{ s.t. } \|f\|_{C^{0,1}([0,1])} < \infty \right\},$$
$$\|f\|_{C^{0,1}([0,1])} \coloneqq \sup_{x \in [0,1]} |f(x)| + \sup_{x,y \in [0,1], x \neq y} \frac{|f(x) - f(y)|}{|x - y|}.$$

[14] As we will see in ECON 6701, completeness refers whether every Cauchy sequences converges.

- L^p (Lebesgue) space: $\left(L^p\left(E\right), \|\cdot\|_{L^p(E)}\right)$ with $1 \leq p \leq \infty$, where, for any measurable $E \subseteq \mathbb{R}^n$,

$$L^p\left(E\right) \coloneqq \left\{f : E \to \mathbb{R} \text{ s.t. } \|f\|_{L^p(E)} < \infty\right\},$$
$$\|f\|_{L^p(E)} \coloneqq \begin{cases} \left(\int_E |f|^p \, \mathrm{d}x\right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \sup_{x \in E} |f\left(x\right)| & \text{if } p = \infty. \end{cases}$$

Let V be an n-dimensional linear space and $S \subseteq V$ be a basis of V. Then, for each $\mathbf{x} \in V$, there exists $(\lambda_i(\mathbf{x}))_{i=1}^n \in \mathbb{R}^n$ such that $\mathbf{x} = \sum_{i=1}^n \lambda_i(\mathbf{x})\mathbf{s}_i$. Define $|\cdot|_V : V \to \mathbb{R}$ by

$$|\mathbf{x}|_V \coloneqq \left(\sum_{i=1}^n |\lambda_i\left(\mathbf{x}\right)|^2\right)^{1/2}.$$

Proposition 27. *Let V be an n-dimensional linear space. Then, $(V, |\cdot|_V)$, where $|\cdot|_V$ is as defined above, is a normed space.*

Proof. It suffices to verify that $|\cdot|_V$ is a norm on V.

Positivity. By construction $|\mathbf{x}|_V \geq 0$ for all $\mathbf{x} \in V$. Suppose first that $\mathbf{x} = \mathbf{0}$. Since $\mathbf{s}_1, \ldots, \mathbf{s}_n$ is a basis and thus linearly independent (Proposition 21), $\mathbf{0} = \mathbf{x} = \sum_{i=1}^n \lambda_i(\mathbf{x})\mathbf{s}_i$ implies $\lambda_1(\mathbf{x}) = \cdots = \lambda_n(\mathbf{x}) = 0$ by Proposition 21.. Thus, $|\mathbf{0}|_V = 0$. Conversely, suppose that $|\mathbf{x}|_V = 0$. Then, by construction, $\lambda_1(\mathbf{x}) = \cdots = \lambda_n(\mathbf{x}) = 0$ so that $\mathbf{x} = \sum_{i=1}^n \lambda_i(\mathbf{x})\mathbf{s}_i = \mathbf{0}$.

Homogeneity. Fix any $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$. Since $\alpha\mathbf{x} = \sum_{i=1}^n \alpha\lambda_i(\mathbf{x})\mathbf{s}_i$,

$$|\alpha\mathbf{x}|_V = \left(\sum_{i=1}^n |\alpha\lambda_i\left(\mathbf{x}\right)|_V^2\right)^{1/2} = |\alpha| \left(\sum_{i=1}^n |\lambda_i\left(\mathbf{x}\right)|_V^2\right)^{1/2} = |\alpha| \, |\mathbf{x}|_V .$$

Triangle inequality. Fix any $\mathbf{x}, \mathbf{y} \in V$. Then, since $|\lambda_i(\mathbf{x}) + \lambda_i(\mathbf{y})| \leq |\lambda_i(\mathbf{x})| + |\lambda_i(\mathbf{y})|$ for each $i \in \{1 \ldots, n\}$,

$$\begin{aligned} |\mathbf{x} + \mathbf{y}|_V &= \left(\sum_{i=1}^n |\lambda_i\left(\mathbf{x}\right) + \lambda_i\left(\mathbf{y}\right)|^2\right)^{1/2} \\ &\leq \left(\sum_{i=1}^n |\lambda_i\left(\mathbf{x}\right)|^2\right)^{1/2} + \left(\sum_{i=1}^n |\lambda_i\left(\mathbf{y}\right)|^2\right)^{1/2} = |\mathbf{x}|_V + |\mathbf{y}|_V . \end{aligned}$$

It follows that $|\cdot|_V$ is a norm on V. ■

Proposition 28. *All norms on finite-dimensional linear spaces are equivalent.*

2.3.1 Geometric interpretation

Consider a point in $\mathbf{z} \coloneqq (x, y) \in \mathbb{R}^2$. The length of this vector (from the origin) in Cartesian coordinate system, $\|\mathbf{z}\|$ is given by the Pythagorean theorem:

$$\|\mathbf{z}\|^2 = x^2 + y^2 \Leftrightarrow \|\mathbf{z}\| = \sqrt{x^2 + y^2}.$$

Observe that this corresponds exactly to the Euclidean norm of vector $\mathbf{z}$.

2.4 Inner product spaces

Consider the linear space $\mathbb{R}^2$. We can think of any vectors $\mathbf{x} \in \mathbb{R}^2$ as representing direction (from the origin) or lines that connects the $\mathbf{x}$ and the origin. Given any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$—or two lines $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$—we can think about the angle between them. For example, we think of the x-axis (represented by $(1, 0)$) and y-axis (represented by $(0, 1)$) as having an angle of 90 degrees between them, and call them as being orthogonal. The inner product captures the notion of angle between any two vectors of a linear space.

An *inner product* on a linear space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following conditions:

- (*positivity*) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \ \forall \mathbf{x} \in V$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$;
- (*symmetry*) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \ \forall \mathbf{x}, \mathbf{y} \in V$;
- (*linearity*) $\langle \alpha \cdot \mathbf{x} + \beta \cdot \mathbf{y}, \mathbf{z} \rangle = \alpha \cdot \langle \mathbf{x}, \mathbf{z} \rangle + \beta \cdot \langle \mathbf{y}, \mathbf{z} \rangle \ \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V \ \forall \alpha, \beta \in \mathbb{R}$.

An inner product space is a pair $(V, \langle \cdot, \cdot \rangle)$, where V is a linear space and $\langle \cdot, \cdot \rangle$ is an inner product on V. Two vectors $\mathbf{x}$ and $\mathbf{y}$ in an inner product space $(V, \langle \cdot, \cdot \rangle)$ are *orthogonal* if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Since

$$\langle \mathbf{x}, \mathbf{0} \rangle = \langle \mathbf{x}, 0 \cdot \mathbf{y} \rangle = 0 \cdot \langle \mathbf{x}, \mathbf{y} \rangle = 0,$$

any nonzero vector is orthogonal to the zero vector. Every inner product space $(V, \langle \cdot \rangle)$ is a normed space with the following induced norm:

$$\|\mathbf{x}\| \coloneqq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

Example 15. An example of a inner product space is the *Euclidean dot product* defined on $\mathbb{R}^n$, where $\cdot : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is given by

$$\mathbf{x} \cdot \mathbf{y} \coloneqq \sum_{i=1}^{n} x_i y_i.$$

The induced norm of the Euclidean dot product is indeed the Euclidean norm. Many inner product can be defined on the same linear space V. For example, with $V = \mathbb{R}^n$, with $(w_i)_{i=1}^n \in \mathbb{R}^n_{++}$,

$$\langle \mathbf{x}, \mathbf{y} \rangle \coloneqq \sum_{i=1}^{n} w_i x_i y_i$$

is an inner product on $\mathbb{R}^n$ (called weighted Euclidean inner product). Some other examples:

- $(L^2(X), \langle \cdot, \cdot \rangle)$ for some measurable X, where $\langle f, g \rangle = \int_X f(x) g(x) \mathrm{d}x$.
- $(\ell^2(\mathbb{N}), \langle \cdot, \cdot \rangle)$, where $\langle (x_n), (y_n) \rangle = \sum_{n=1}^{\infty} x_n y_n$.[15]

Theorem 4 (Cauchy-Schwarz Inequality)**.** *Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space and $\|\cdot\|$ be the induced norm. Then,*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\| \ \ \forall \mathbf{x}, \mathbf{y} \in V.$$

Exercise 17. Prove the Cauchy-Schwarz inequality.

Proposition 29 (Generalised Pythagorean theorem)**.** *Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space and $\|\cdot\|$ be the induced norm. Then,*

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2 \langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2 \ \ \forall \mathbf{x}, \mathbf{y} \in V.$$

[15] Holds also for $\ell^2(\mathbb{Z})$ which is a space of sequences of the form $(\ldots, b_{-2}, b_{-1}, b_0, b_1, b_2, \ldots)$.

In particular, if $\mathbf{x}$ *and* $\mathbf{y}$ *are orthogonal, then*

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 .$$

Proof. We use the linearity and symmetry property of the inner product:

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle \\ &= \langle \mathbf{x}, \mathbf{x}+\mathbf{y}\rangle + \langle \mathbf{y}, \mathbf{x}+\mathbf{y}\rangle \\ &= \langle \mathbf{x}, \mathbf{x}\rangle + \langle \mathbf{x}, \mathbf{y}\rangle + \langle \mathbf{y}, \mathbf{x}\rangle + \langle \mathbf{y}, \mathbf{y}\rangle \\ &= \|\mathbf{x}\|^2 + 2\cdot\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 .\end{aligned}$$

Finally, recall that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\langle \mathbf{x}, \mathbf{y}\rangle = 0$. ■

The following results says that sum of the squares of the lengths of the four sides of a parallelogram equals the sum of the squares of the lengths of the two diagonals.

Corollary 6 (Parallelogram identity)**.** *Let* $(V, \langle\cdot,\cdot\rangle)$ *be an inner product space and* $\|\cdot\|$ *be the induced norm. Then,*

$$\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in V.$$

Proof. By the previous proposition,

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|^2 &= \|\mathbf{x}+(-1)\cdot\mathbf{y}\|^2 \\ &= \|\mathbf{x}\|^2 + 2\cdot\langle \mathbf{x}, (-1)\cdot\mathbf{y}\rangle + \|(-1)\cdot\mathbf{y}\|^2 \\ &= \|\mathbf{x}\|^2 - 2\cdot\langle \mathbf{x}, \mathbf{y}\rangle + |-1|^2\|\mathbf{y}\|^2 \\ &= \|\mathbf{x}\|^2 - 2\cdot\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 .\end{aligned}$$

Thus,

$$\begin{aligned}\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{y}\|^2 &= \left(\|\mathbf{x}\|^2 - 2\cdot\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2\right) + \left(\|\mathbf{x}\|^2 + 2\cdot\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2\right) \\ &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 .\end{aligned}$$

■

2.4.1 Geometric interpretation

Let x, y and z represents the length of the three sides of a triangle and θ the angle between sides of lengths x and y. The law of cosine, which is a generalisation of the Pythagorean theorem to non-right triangles, says that

$$z^2 = x^2 + y^2 - 2xy\cos\theta.$$

Since "length" of vector in $\mathbb{R}^2$ is represented by the Euclidean norm $\|\cdot\|$, abusing notation slightly and letting $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$, we can rewrite the law of cosine as

$$\|\mathbf{z}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta.$$

Using the fact that $\mathbf{x} = \mathbf{y} + \mathbf{z} \Leftrightarrow \mathbf{z} = \mathbf{x} - \mathbf{y}$,

$$\begin{aligned}2\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}+(-\mathbf{y})\|^2 \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \left(\|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2\right) = 2\langle \mathbf{x}, \mathbf{y}\rangle \\ \Leftrightarrow \cos\theta &= \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle}}.\end{aligned}$$

Hence, the angle between two vectors can be computed using the inner product. Observe that if two vectors are orthogonal, i.e., $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, then the angle between them is $\cos\theta = 0$ so that θ is 90 degrees.

2.4.2 Orthogonal complements

Given an inner product space $(V, \langle \cdot, \cdot \rangle)$ and $W \subseteq V$, $(W, \langle \cdot, \cdot \rangle|_W)$ is a *subspace* of the inner product space and is itself an inner product space. A vector $\mathbf{x} \in V$ is *orthogonal to subspace* W if it is orthogonal to every vector in W; i.e.,

$$\langle \mathbf{x}, \mathbf{w} \rangle = 0 \ \forall \mathbf{w} \in W.$$

The set of all vectors in V that are orthogonal to W is the *orthogonal complement of* W and denoted

$$W^\perp \coloneqq \{\mathbf{x} \in V : \langle \mathbf{x}, \mathbf{w} \rangle = 0 \ \forall \mathbf{w} \in W\},$$

where $\perp$ is read "perp." Note $W^\perp$ is a subspace of V, $(W^\perp)^\perp = W$, and the only common vector between W and $W^\perp$ is $\mathbf{0}$. The last observation means that V is a direct sum of W and $W^\perp$; i.e.,

$$V = W \oplus W^\perp.$$

Hence, any $\mathbf{x} \in V$ can be uniquely expressed as $\mathbf{x} = \mathbf{w} + \mathbf{w}^\perp$, where $\mathbf{w} \in W$ and $\mathbf{w}^\perp = W^\perp$.

Proposition 30. *Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. Then, $\{\mathbf{0}\}^\perp = V$ and $V^\perp = \{\mathbf{0}\}$.*

Proof. Recall that $\langle \mathbf{0}, \mathbf{x} \rangle = 0$ for all $x \in V$ and hence $\{\mathbf{0}\}^\perp = V$ and $V^\perp = \{\mathbf{0}\}$. ■

Example 16. The orthogonal complement of a line W through the original in $\mathbb{R}^2$ is the line perpendicular to W. The orthogonal complement of a line W in $\mathbb{R}^3$ is the plane that is perpendicular to W.

2.4.3 Orthonormal bases

A set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq V$ of an inner product space $(V, \langle \cdot, \cdot \rangle)$ is an *orthogonal set* if all pairs of distinct vectors in the set are orthogonal. An orthogonal set in which each vector has a unit norm is called *orthonormal*. Since any vector $\mathbf{x} \in V$ can be normalised to have a unit norm by dividing by $\|\mathbf{x}\|$, it follows that any orthogonal set of nonzero vectors can always be converted to an orthonormal set via such normalisation. A basis consisting of orthonormal vectors in an inner product space is an *orthonormal basis*, and a basis consisting of orthogonal vectors is an *orthogonal basis*.

If $V = \mathbb{R}^n$, the canonical basis, $\{\mathbf{e}_i\}_{i=1}^n$ is orthonormal since $\|\mathbf{e}_i\| = 1$ under the Euclidean norm. In fact, it is an orthogonal basis since $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$ for any $i \neq j$.

We are often interested in finding orthonormal basis as it allows vectors to be expressed in a simple manner.

Proposition 31. *Suppose $S \coloneqq \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is an orthogonal basis for an inner product space $(V, \langle \cdot, \cdot \rangle)$ and let $\|\cdot\|$ be the induced norm. Then,*

$$\mathbf{x} = \sum_{i=1}^{n} \frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \mathbf{s}_i \ \forall \mathbf{x} \in V.$$

Proof. By definition of a basis, for any $\mathbf{x} \in V$, there exists $(\lambda_i)_{i=1}^n \in \mathbb{R}^n$ such that $\mathbf{x} = \sum_{j=1}^n \lambda_j \mathbf{s}_j$. The result then follows because

$$\frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} = \frac{1}{\|\mathbf{s}_i\|^2} \left\langle \sum_{j=1}^n \lambda_j \mathbf{s}_j, \mathbf{s}_i \right\rangle = \sum_{j=1}^n \lambda_j \frac{\langle \mathbf{s}_j, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} = \lambda_i \frac{\langle \mathbf{s}_i, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} = \lambda_i,$$

where the penultimate equality uses the fact that $\mathbf{s}_i$ and $\mathbf{s}_j$ for $i \neq j$ are orthogonal to each other and the last equality uses the fact that $\|\mathbf{s}_i\| = \sqrt{\langle \mathbf{s}_i, \mathbf{s}_i \rangle}$. ■

We can think of $(\mathbf{x})_S \coloneqq (\langle \mathbf{x}, \mathbf{s}_i \rangle)_{i=1}^n$ as the coordinates of the vector $\mathbf{x} \in V$ relative to the orthonormal basis $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Expressing vectors in this way allows simple computation of inner product, induced norm and induced metric associated with vector(s) in original coordinates.

Proposition 32. *Suppose that $S \coloneqq \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is an orthonormal basis for an inner product space $(V, \langle \cdot, \cdot \rangle)$ and let $\|\cdot\|$ and $d(\cdot, \cdot)$ be the induced norm and induced metric, respectively. Then, for any $\mathbf{x}, \mathbf{y} \in V$,*

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle &= \sum_{i=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle \cdot \langle \mathbf{y}, \mathbf{s}_i \rangle = \sum_{i=1}^n x_i^S y_i^S, \\
\|\mathbf{x}\| &= \sqrt{\sum_{i=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle^2} = \sum_{i=1}^n \left(x_i^S\right)^2, \\
d(\mathbf{x}, \mathbf{y}) &= \sqrt{\sum_{i=1}^n \langle \mathbf{x} - \mathbf{y}, \mathbf{s}_i \rangle^2} = \sqrt{\sum_{i=1}^n \left(x_i^S - y_i^S\right)^2}.
\end{aligned}$$

where $(x_i^S)_{i=1}^n$ and $(y_i^S)_{i=1}^n$ are such that $(\mathbf{x})_S = (x_i^S)_{i=1}^n$ and $(\mathbf{y})_S = (y_i^S)_{i=1}^n$.

Proof. By the linearity of $\langle \cdot, \cdot \rangle$,

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle &= \left\langle \sum_{i=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle \mathbf{s}_i, \sum_{i=1}^n \langle \mathbf{y}, \mathbf{s}_i \rangle \mathbf{s}_i \right\rangle \\
&= \sum_{i=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle \cdot \left\langle \mathbf{s}_i, \sum_{j=1}^n \langle \mathbf{y}, \mathbf{s}_j \rangle \mathbf{s}_j \right\rangle \\
&= \sum_{i=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle \cdot \left(\sum_{j=1}^n \langle \mathbf{y}, \mathbf{s}_j \rangle \cdot \langle \mathbf{s}_i, \mathbf{s}_j \rangle \right) \\
&= \sum_{i=1}^n \sum_{j=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle \cdot \langle \mathbf{y}, \mathbf{s}_j \rangle \cdot \langle \mathbf{s}_i, \mathbf{s}_j \rangle .
\end{aligned}$$

Since $\langle \mathbf{s}_i, \mathbf{s}_j \rangle = 0$ for all $i \neq j$ and $\langle \mathbf{s}_i, \mathbf{s}_i \rangle = 1$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n \langle \mathbf{x}, \mathbf{s}_i \rangle \cdot \langle \mathbf{y}, \mathbf{s}_i \rangle = \sum_{i=1}^n x_i^S y_i^S.$$

The expression for the induced norm follows from the fact that $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. The expression for the induced metric comes from the fact that $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. ■

Of course, the usefulness of orthonormal bases may be vacuous if they do not exist. Luckily, we have the following result. The proof involves constructing an orthonormal basis from an arbitrary basis (which we know exists by Theorem 19). The algorithm is called the *Gram-Schmidt process.*

We first show that orthogonality implies linear independence.

Proposition 33. *Suppose that $S \coloneqq \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is an orthogonal set of nonzero vectors in an inner product space $(V, \langle\cdot,\cdot\rangle)$. Then, S is linearly independent.*

Proof. Take any $(\lambda_i)_{i=1}^n \in \mathbb{R}^n$ such that $\sum_{i=1}^n \lambda_i \mathbf{s}_i = \mathbf{0}$. By Proposition 19, it suffices to show that $\lambda_1 = \cdots = \lambda_n = 0$. Since S is an orthogonal set, $\forall i \in \{1, \ldots, n\}$,

$$0 = \langle \mathbf{0}, \mathbf{s}_i \rangle = \left\langle \sum_{j=1}^n \lambda_j \mathbf{s}_j, \mathbf{s}_i \right\rangle = \sum_{j=1}^n \lambda_j \langle \mathbf{s}_j, \mathbf{s}_i \rangle = \lambda_i \langle \mathbf{s}_i, \mathbf{s}_i \rangle .$$

Since each $\mathbf{s}_i$ is a nonzero vector $\langle \mathbf{s}_i, \mathbf{s}_i \rangle > 0$ by positivity. Therefore, $\lambda_i = 0 \ \forall i \in \{1, \ldots, n\}$. ∎

Next we introduce the concept of orthogonal projections. This generalises the fact that each vector $\mathbf{x} \in \mathbb{R}^2$ can be expressed as sum of vectors along the x- and y-axes, which are orthogonal.

Proposition 34. *Suppose W is a finite-dimensional subspace of an inner product space $(V, \langle\cdot,\cdot\rangle)$ and that $S \coloneqq (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is an orthogonal basis for W. Then, for any $\mathbf{x} \in V$,*

$$\begin{aligned} \operatorname{proj}_W \mathbf{x} &\coloneqq \sum_{i=1}^n \frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \mathbf{s}_i \in W, \\ \operatorname{proj}_{W^\perp} \mathbf{x} &\coloneqq \mathbf{x} - \operatorname{proj}_W \mathbf{x} \in W^\perp. \end{aligned}$$

Proof. Take the premise of the proposition.

(i) $\operatorname{proj}_W \mathbf{x} \in W$. Since $\langle \mathbf{x}, \mathbf{s}_i \rangle / \|\mathbf{s}_i\|^2 \in \mathbb{R}$, $\operatorname{proj}_W \mathbf{x}$ is a linear combination of vectors in S. Thus,

$$\operatorname{proj}_W \mathbf{x} \in \operatorname{span}(S) = W,$$

where the last equality follows by the fact that S is a basis and Proposition 21.

(ii) $\operatorname{proj}_{W^\perp} \mathbf{x} \in W^\perp$. Define $\mathbf{x}^\perp \coloneqq \operatorname{proj}_{W^\perp} \mathbf{x}$. We wish to show that $\langle \mathbf{x}^\perp, \mathbf{w} \rangle = 0 \ \forall \mathbf{w} \in W$. Using definitions and linearity of $\langle\cdot,\cdot\rangle$,

$$\begin{aligned} \left\langle \mathbf{x}^\perp, \mathbf{w} \right\rangle &= \langle \mathbf{x}, \mathbf{w} \rangle - \left\langle \sum_{i=1}^n \frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \mathbf{s}_i, \mathbf{w} \right\rangle \\ &= \langle \mathbf{x}, \mathbf{w} \rangle - \sum_{i=1}^n \frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \langle \mathbf{s}_i, \mathbf{w} \rangle . \end{aligned}$$

By Proposition 31, $\mathbf{w} = \sum_{i=1}^n \langle \mathbf{w}, \mathbf{s}_i \rangle \mathbf{s}_i / \|\mathbf{s}_i\|^2$ so that

$$\begin{aligned} \left\langle \mathbf{x}^\perp, \mathbf{w} \right\rangle &= \left\langle \mathbf{x}, \sum_{i=1}^n \frac{\langle \mathbf{w}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \mathbf{s}_i \right\rangle - \sum_{i=1}^n \frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \langle \mathbf{s}_i, \mathbf{w} \rangle \\ &= \sum_{i=1}^n \frac{\langle \mathbf{w}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \langle \mathbf{x}, \mathbf{s}_i \rangle - \sum_{i=1}^n \frac{\langle \mathbf{x}, \mathbf{s}_i \rangle}{\|\mathbf{s}_i\|^2} \langle \mathbf{s}_i, \mathbf{w} \rangle = 0. \end{aligned}$$

∎

We refer to $\operatorname{proj}_W \mathbf{x}$ as an *orthogonal projection of* $\mathbf{x}$ *on* W and orthogonal projection of $\mathbf{x}$ on W as a *component of* $\mathbf{x}$ *orthogonal to* W. By Proposition 21 and Corollary 3, projections are unique relative to each orthogonal basis. Put differently, each distinct orthonormal basis of W gives rise to a unique projection of $\mathbf{x}$ on W.

Theorem 5. *Every nonzero finite dimensional inner product space has an orthonormal basis.*

Proof. Let $(V, \langle\cdot,\cdot\rangle)$ be a nonzero inner product space. Since V is a linear space, by Theorem 19, there exists a basis $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ so that V is n dimensional. The goal is to construct an orthonormal basis, $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, from S (note that, if it exists, V must have the same number of elements by Corollary 13.

Step 1 Define $\mathbf{v}_1 \coloneqq \mathbf{s}_1$.

Step 2 Define $\mathbf{v}_2$ that is orthogonal to $\mathbf{v}_1$ by computing the component of $\mathbf{s}_2$ that is orthogonal to $W_1 \coloneqq \text{span}(\mathbf{v}_1)$. Note that $\mathbf{v}_1$ is the (trivially orthogonal) basis of W_1 by construction, by Proposition 34,

$$\mathbf{v}_2 \coloneqq \text{proj}_{W_1^\perp}\mathbf{s}_2 = \mathbf{s}_2 - \frac{\langle \mathbf{s}_2, \mathbf{v}_1\rangle}{\|\mathbf{v}_1\|^2}\cdot\mathbf{v}_1 \in W_1^\perp$$

For $\mathbf{v}_2$ to be a basis, we must have $\mathbf{v}_2 \neq \mathbf{0}$. To show this, by way of contradiction, suppose that $\mathbf{v}_2 = \mathbf{0}$, then

$$\mathbf{s}_2 = \frac{\langle \mathbf{s}_2, \mathbf{v}_1\rangle}{\|\mathbf{v}_1\|^2}\cdot\mathbf{v}_1 = \frac{\langle \mathbf{s}_2, \mathbf{s}_1\rangle}{\|\mathbf{s}_1\|^2}\cdot\mathbf{s}_1.$$

This implies that $\mathbf{s}_2$ is a scalar multiple of $\mathbf{s}_1$ so that $\mathbf{s}_1$ and $\mathbf{s}_2$ are linearly dependent. However, this contradicts the fact that any basis is linearly independent (Proposition 21).

Step 3 Define $\mathbf{v}_3$ that is orthogonal to $\mathbf{v}_1$ and $\mathbf{v}_2$ by computing the component of $\mathbf{s}_3$ that is orthogonal to $W_2 = \text{span}(\{\mathbf{v}_1, \mathbf{v}_2\})$. Since $\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthogonal basis of W_2 by construction, by Proposition 34,

$$\mathbf{v}_3 \coloneqq \text{proj}_{W_2^\perp}\mathbf{s}_3 = \mathbf{s}_3 - \frac{\langle \mathbf{s}_3, \mathbf{v}_1\rangle}{\|\mathbf{v}_1\|^2}\cdot\mathbf{v}_1 - \frac{\langle \mathbf{s}_3, \mathbf{v}_2\rangle}{\|\mathbf{v}_2\|^2}\cdot\mathbf{v}_2 \in W_2^\perp.$$

If $\mathbf{v}_3 = \mathbf{0}$, then

$$\begin{aligned}\mathbf{s}_3 &= \frac{\langle \mathbf{s}_3, \mathbf{v}_1\rangle}{\|\mathbf{v}_1\|^2}\cdot\mathbf{v}_1 + \frac{\langle \mathbf{s}_3, \mathbf{v}_2\rangle}{\|\mathbf{v}_2\|^2}\cdot\mathbf{v}_2 \\ &= \frac{\langle \mathbf{s}_3, \mathbf{v}_1\rangle}{\|\mathbf{v}_1\|^2}\cdot\mathbf{s}_1 + \frac{\langle \mathbf{s}_3, \mathbf{v}_2\rangle}{\|\mathbf{v}_2\|^2}\cdot\left(\mathbf{s}_2 - \frac{\langle \mathbf{s}_2, \mathbf{v}_1\rangle}{\|\mathbf{v}_1\|^2}\cdot\mathbf{s}_1\right)\end{aligned}$$

so that $\mathbf{s}_3$ is a linear combination of $\mathbf{s}_1$ and $\mathbf{s}_2$, contradicting the fact that any basis is linearly independent (Proposition 21).

Continuing in this way, after n steps, we will obtain an orthogonal set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. To conclude the proof, we must show that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis. By Proposition 21, it suffices to show that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly independent and that $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = V$. Linear independence follows from the fact that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ are orthogonal and Proposition 33. That $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = V$ follows from Proposition 22 because $\dim V = n$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is linearly independent. ■

We can, in fact, apply the Gram-Schmidt process to any linearly independent set of vectors in an inner product space to obtain an orthonormal set of vectors.

Corollary 7 (Projection Theorem)**.** *Suppose W is a finite dimensional subspace of an inner product space $(V, \langle\cdot,\cdot\rangle)$. Then, for any $\mathbf{x} \in V$, there exists an orthogonal projection of $\mathbf{x}$ on W, denoted $\text{proj}_W\mathbf{x} \in W$, and component of $\mathbf{x}$ orthogonal to W, denoted $\text{proj}_{W^\perp}\mathbf{x} \in W^\perp$, such that*

$$\mathbf{x} = \text{proj}_W\mathbf{x} + \text{proj}_{W^\perp}\mathbf{x}.$$

Proof. Given Proposition 34, it suffices to show that W has an orthogonal basis. But we know that W has an orthogonal basis from Theorem 5. ■

Remark 3. Note that we can write

$$\mathbf{x} = \text{proj}_W \mathbf{x} + (\mathbf{x} - \text{proj}_W \mathbf{x}) .$$

Does this look familiar?

Theorem 6 (Approximation Theorem)**.** *Suppose W is a finite dimensional subspace of an inner product space $(V, \langle \cdot, \cdot \rangle)$. For any $\mathbf{x} \in V$, $\text{proj}_W \mathbf{x}$ is the best approximation to $\mathbf{x}$ from W in the sense that it is the element in W that minimises the distance between $\mathbf{x}$ and itself; i.e.,*

$$\|\mathbf{x} - \text{proj}_W \mathbf{x}\| < \|\mathbf{x} - \mathbf{w}\| \ \ \forall \mathbf{w} \in W \backslash \{\text{proj}_W \mathbf{x}\} .$$

Proof. Adding and subtracting $\text{proj}_W \mathbf{x}$, we can write

$$\mathbf{x} - \mathbf{w} = (\mathbf{x} - \text{proj}_W \mathbf{x}) + (\text{proj}_W \mathbf{x} - \mathbf{w}) .$$

By construction, $\text{proj}_\text{W} \mathbf{x} \in \text{W}$ so that $\text{proj}_W \mathbf{x} - \mathbf{w} \in W$, and $\mathbf{x} - \text{proj}_W \mathbf{x} = \text{proj}_{W^\top} \mathbf{x} \in W^\top$. That is, the two terms in the expression above are orthogonal. Then, by the Generalised Pythagorean Theorem (Theorem 29),

$$\|\mathbf{x} - \mathbf{w}\|^2 = \|\mathbf{x} - \text{proj}_W \mathbf{x}\|^2 + \|\text{proj}_W \mathbf{x} - \mathbf{w}\|^2 .$$

By positivity, the second term is strictly positive unless $\text{proj}_W \mathbf{x} = \mathbf{w}$. Thus, for any $\mathbf{w} \in W \backslash \{\text{proj}_W \mathbf{x}\}$,

$$\|\mathbf{x} - \mathbf{w}\|^2 > \|\mathbf{x} - \text{proj}_W \mathbf{x}\|^2 ,$$

which implies the desired result. ■

Example 17 (Fourier series)**.** A function of the form

$$\begin{aligned} \tau(x) = c_0 &+ c_1 \cos(x) + c_2 \cos(2x) + \cdots + c_n \cos(nx) \\ &+ d_1 \sin(x) + d_2 \sin(2x) + \cdots + d_n \sin(nx) \end{aligned}$$

is called a *trigonometric polynomial*, and if c_n and d_n are not both zero, then $\tau(x)$ is said to have order n. It can be shown that

$$S \coloneqq \{1, \cos(x), \cos(2x), \ldots, \cos(nx), \sin(x), \sin(2x), \ldots, \sin(nx)\}$$

are linearly independent so that, for any interval $[a, b] \subseteq \mathbb{R}$, they form a basis for a $(2n + 1)$-dimensional subspace of $C([a, b])$ (i.e., continuous functions in the interval $[a, b]$). Let $\mathbf{f} \in C([0, 2\pi])$, where $(C(0, 2\pi), \langle \cdot, \cdot \rangle)$ is an inner product space with

$$\langle \mathbf{f}, \mathbf{g} \rangle \coloneqq \int_0^{2\pi} f(x)\, g(x)\, \mathrm{d}x.$$

We wish to find the best approximation to $\mathbf{f}$ from W, where $W \subseteq C([0, 2\pi])$ is the space of trigonometric polynomial functions of order n. By the Approximation Theorem (Theorem 6), the best approximation is given by $\text{proj}_W \mathbf{f}$, which requires to find orthonormal basis $\{\mathbf{g}_1, \ldots, \mathbf{g}_{2n}\}$ from W (see Proposition 34):

$$\text{proj}_W \mathbf{f} = \sum_{i=1}^{2n} \langle \mathbf{f}, \mathbf{g}_i \rangle \mathbf{g}_i.$$

Since S is a basis for W, we can obtain an orthonormal basis by applying the Gram-Schmidt process.

The result is that

$$\mathbf{g}_0 = \frac{1}{\sqrt{2\pi}},\ \mathbf{g}_i = \begin{cases} \frac{1}{\sqrt{\pi}} \cos(ix) & \text{if } i \in \{1, \ldots, n\} \\ \frac{1}{\sqrt{\pi}} \sin((i-n)x) & \text{if } i \in \{n+1, \ldots, 2n\} \end{cases}.$$

Define

$$\begin{aligned} c_0 &= \frac{2}{\sqrt{2\pi}} \langle \mathbf{f}, \mathbf{g}_0 \rangle = \frac{1}{\pi} \int_0^{2\pi} f(x)\, \mathrm{d}x \\ c_i &= \frac{1}{\sqrt{\pi}} \langle \mathbf{f}, \mathbf{g}_i \rangle = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(ix)\, \mathrm{d}x \ \forall i \in \{1, \ldots, n\}, \\ d_i &= \frac{1}{\sqrt{\pi}} \langle \mathbf{f}, \mathbf{g}_{n+i} \rangle = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(ix)\, \mathrm{d}x \ \forall i \in \{1, \ldots, n\}, \end{aligned}$$

then

$$\mathrm{proj}_W \mathbf{f} = \frac{a_0}{2} + \left(\sum_{i=1}^{n} c_i \cos(ix) \right) + \left(\sum_{i=1}^{n} d_i \sin(ix) \right).$$

The coefficients $(c_0, c_1, \ldots, c_n, d_1, \ldots, d_n)$ are called the *Fourier coefficients* of $\mathbf{f}$. It can be shown that approximation improves as $n \to \infty$ so that

$$f(x) = \frac{a_0}{2} + \sum_{i=1}^{\infty} (c_i \cos(ix) + d_i \sin(ix)).$$

The right-hand side of this equation is called the *Fourier series* for f over the interval $[0, 2\pi]$.

2.5 Topologies

Let X be a nonempty set. A *topology* $\mathcal{T}$ for X is a collection of subsets of X, called *open sets*, that have the following properties.

(i) The entire set, X, and the empty set, $\varnothing$, are open—i.e. $\varnothing, X \in \mathcal{T}$.

(ii) The intersection of any finite collection of open sets is open—i.e. for any $\mathcal{O} = \{O_1, O_2, \ldots, O_n\} \subseteq \mathcal{T}$, $\bigcap_{i=1}^{n} O_i \in \mathcal{T}$;

(iii) The union of any collection of any open set is open—i.e. for any $\mathcal{O} \subseteq \mathcal{T}$, $\bigcup_{O \in \mathcal{O}} O \in \mathcal{T}$.

The pair $(X, \mathcal{T})$, where X is an nonempty set and $\mathcal{T}$ is a topology for X, is a *topological space*. For any $x \in X$, an open set that contains x is a *neighbourhood* of x. A collection of neighbourhoods of x, $\mathcal{B}_x$, is called a *base for the topology at* x if, for any neighbourhood O of x, there exists a set $B \in \mathcal{B}_x$ such that $B \subseteq O$. A collection of open sets $\mathcal{B}$ is called a *base for the topology* $\mathcal{T}$ if it contains a base for the topology at each point in X. A base determines a unique topology on X that consists of $\varnothing$, and the union of sets belonging to the base.

Proposition 35. *A subset $E \subseteq X$ is open if and only if, for each point $x \in E$, there is a neighbourhood of x that is contained in E. That is,*

$$E \in \mathcal{T} \Leftrightarrow \forall x \in E,\ \exists O_x \in \mathcal{T},\ x \in O_x \subseteq E.$$

Proof. Let $E \subseteq X$, where $(X, \mathcal{T})$ is a topological space. Suppose that $\forall x \in E$, $\exists O_x \in \mathcal{T}$, $x \in O_x \subseteq E$. Then,

$$x \in O \coloneqq \bigcup_{x' \in E} O_{x'}\ \forall x \in E \Rightarrow E \subseteq O.$$

Moreover, since $O_x \subseteq E$ for all $x \in E$, it follows that $O \subseteq E$. That is, $O = E$. Finally, since O is a union of open sets, by (iii) of the definition, $O = E$ is an open set. Conversely, suppose that E is open. Then, E is a neighbourhood of any x that is contained in E. ■

Proposition 36. *Every open set is a union of neighbourhoods.*

Proof. Let $E \subseteq X$ be an open set, where $(X, \mathcal{T})$ is a topological space. For each $x \in E$, let N_x be the open neighbourhood in E containing x, which must exist by proposition above. Consider $N \coloneqq \bigcup_{x \in E} N_x$. Since for every x, $x \in N_x \subseteq N$, it follows that $U \subseteq N$. Since every $N_x \subseteq E$, it also follows that $N \subseteq E$. Thus, $U = N$. ■

For a subset $E \subseteq X$, a point $x \in X$ is a *point of closure* of E provided that every neighbourhood of x contains a point in E. The collection of points of closure of E is the *closure* of E, denoted by $\overline{E}$. Since $E \subseteq \overline{E}$, if E contains all of its points of closure, i.e., $E = \overline{E}$, then E is *closed*.

Proposition 37. *A subset of a topological space X is open if and only if its complement in X is closed.*

Proof. First, suppose that $E \subseteq X$ is open. Let $x \in X$ be a point of closure of $E^c = X \backslash E$; i.e., every neighbourhood of x contains a point in E^c. Then, x cannot belong to E because, otherwise, there would be a neighbourhood of x that is contained in E and therefore does not intersect E^c. Thus, x belongs to E^c and hence E^c is closed. Now suppose that E^c is closed; i.e., E^c contains all of its points of closure. Let $x \in E$. Then, there must be a neighbourhood of x that is contained in E, for otherwise, every neighbourhood of x would contain points in E^c and therefore x would be a point of closure of E^c. Since E^c is closed, x would belong to E^c, which is a contradiction. ■

Remark 4. Above proposition combined with De Morgan's law gives an alternative characterisation of topological spaces using closed sets. A topology $\mathcal{T}$ for X is a collection of subsets of X, called closed sets, that have the following properties.

(i) The entire set, X, and the empty set, $\varnothing$, are closed;

(ii) The union of any finite collection of closed sets is closed—i.e. for any $\mathcal{C} = \{C_1, C_2, \ldots, C_n\} \subseteq \mathcal{T}$, $\bigcup_{i=1}^{n} C_i$ is closed;

(iii) The intersection of any collection of any closed set is closed—i.e. for any $\mathcal{C} \subseteq \mathcal{T}$, $\bigcap_{C \in \mathcal{C}} C$ is closed.

Given a topology $(X, \mathcal{T})$, the *interior of* $S \subseteq X$, denoted $\text{int}(S)$, is the union of all subsets of S that are open in X. A point that is in the interior of S is an *interior point* of S. If X is in fact a metric space, then $x \in X$ is an interior point of S if there exists $r > 0$ such that y is in S whenever $\rho(x, y) < r$. The *boundary of* $S \subseteq X$, denoted $\text{bd}(S)$ or ∂S, is the set of points in the closure of S not belonging to the interior of S. A point that is in the boundary of S is a *boundary point* of S.

Example 18 (Examples of topologies)**.**

- *Discrete topology.* Let X be any nonempty set. The discrete topology for X is the collection of all subsets of X. For discrete topology, every set containing a point is a neighbourhood of that point. The discrete topology is induced by the discrete metric.
- *Trivial topology.* let X be any nonempty set. Then, $\{\varnothing, X\}$ is the *trivial topology* for X. The only neighbourhood of a point is the whole set X.
- *Topological subspaces.* Given a topological space $(X, \mathcal{T})$ and a nonempty subset $E \subseteq X$, we define the *inherited topology* $\mathcal{S}$ for E to consists of all sets of the form $E \cap O$,where $O \in \mathcal{T}$. We call the topological space $(E, \mathcal{S})$ a subspace of $(X, \mathcal{T})$.

- *Product topology.* Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be two topological spaces. In $X \times Y$, consider the collection of sets $\mathcal{B}$ consisting of $O_1 \times O_2$, where O_1 is open in X and O_2 is open in Y. Then, $(X \times Y, \mathcal{B})$ is the *product topology* for $X \times Y$.
- *Metric topology.* Let (X, ρ) be a metric space. Define an ball centred at $x \in X$ with radius $r > 0$ as
$$B_r(x) \coloneqq \{y \in X : \rho(z, x) < r\}.$$
Define a subset $O \subseteq X$ to be open if, for every $x \in O$, $B_r(x) \subseteq O$ for some $r > 0$. The collection of such open sets is the *metric topology* for X induced by the metric ρ.
- *Strong* (or *norm*) *topology.* Let V be a normed space. The metric topology for V induced by the natural metric ρ is called the *strong* (or no*rm*) *topology* for X.

Let X and Y be topological spaces. A function $f : X \to Y$ is *continuous* if, for any open set $V \subseteq Y$, $f^{-1}(V) \equiv \{x \in X : f(x) \in V\}$ is an open subset of X.[16]

2.5.1 Order over topologies

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologies for a nonempty set X. We say that $\mathcal{T}_1$ is *weaker* than $\mathcal{T}_2$ (or, equivalently, $\mathcal{T}_2$ is stronger than $\mathcal{T}_1$) if $\mathcal{T}_1 \subseteq \mathcal{T}_2$. Thus, any open set in the weaker topology is also open in the stronger topology. The converse is not necessarily true.

2.5.2 Compactness

Let $(X, \mathcal{T})$ be a topological space. A collection of sets $\mathcal{S} = \{S_\lambda\}_{\lambda \in \Lambda}$ is a *cover* of a set $S \subseteq X$ if $S \subseteq \bigcup_{\lambda \in \Lambda} S_\lambda$. Let $\mathcal{S}$ be a cover of S. A *subcover* is a subset of $\mathcal{S}$ that is also a cover of S. If each set S_λ is open, then $\mathcal{S}$ is an *open cover*. If Λ is a finite set, then $\mathcal{S}$ is a *finite cover* (note that each set S_λ need not be finite). A subset $S \subseteq X$ is *compact* (in X) if every open cover of S has a finite subcover.

If $X = \mathbb{R}^n$, then a subset $S \subseteq X$ is compact if and only if S is closed and bounded (*Heine-Borel* theorem). In a metric space, while compactness implies closedness and boundedness, the converse only holds if the space is *complete* (i.e., if every Cauchy sequence is convergent). The usefulness of compactness comes from the fact that any finite set has a maximum—thus, existence of finite covers allows us to take a maximum among the finite subcovers.

2.5.3 Homeomorphism

Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be two topological spaces. A function $f : X \to Y$ is a *homeomorphism* if f is a bijection, f is continuous, and the inverse function f^{-1} is continuous. If such a homeomorphism exists between X and Y, then they are *homeomorphic* and the two spaces have the same topological properties (e.g., if a set is compact in one space, it is also compact in the other). In fact, homeomorphic spaces induces an equivalence relation on topological spaces.

Example 19. An open interval (a, b) is homeomorphic to $\mathbb{R}$. To see this, define $f : (a, b) \to \mathbb{R}$ as $f(x) \coloneqq \frac{1}{a-x} + \frac{1}{b-x}$. Note that $\mathbb{R}^m$ and $\mathbb{R}^n$ are not homeomorphic if $m \neq n$.

[16] Equivalently, f is continuous if inverse image of any closets set in Y are closed in X.

3 Linear algebra

3.1 Matrices

For any $m, n \in \mathbb{N}$, an $m \times n$ *matrix* (in $\mathbb{R}$), denoted X, is a real-valued function

$$X : \{1, \ldots, m\} \times \{1, \ldots, n\} \to \mathbb{R},$$

which we often represent as

$$X \equiv [x_{ij}] \equiv \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix},$$

where $x_{ij} \coloneqq X(i, j)$ for each $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$. We may write $[x_{ij}]_{m \times n}$ to stress its dimension. Thus, x_{ij} represents the ith row and jth column—or the ijth entry—of the matrix X. Matrices are usually written with uppercase letters and its element with corresponding lowercase letters.

Real numbers are 1×1 matrices, and a vector in $\mathbb{R}^n$ can be seen as either $n \times 1$ or $1 \times n$ matrices. We often refer to $n \times 1$ matrices as *column vectors* and $1 \times n$ matrices as *row vectors*. Unless otherwise specified, a vector is taken to be a column vector. A matrix whose entries all zero is called a *zero matrix*. We call any matrices with $m = n$ as $(n \times n)$ *square matrices*. A square matrix X is *diagonal* if all its nondiagonal elements are zero; i.e., X is a diagonal matrix if $a_{ij} = 0$ $\forall (i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\} : i \neq j$, and we write $X = \operatorname{diag}(a_{11}, \ldots, a_{nn})$. An $n \times n$ square matrix X is an *identity matrix* if $X = \operatorname{diag}(1, \ldots, 1)$ and we denote it as $\mathbb{I}_n$. A square matrix X is *upper* (resp. *lower*) *triangular* if all its elements below its diagonals are zeros; i.e., $a_{ij} = 0$ $\forall (i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\} : i > j$ (resp. $a_{ij} = 0$ $\forall (i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\} : i < j$). A diagonal matrix is both upper and lower triangular.

3.1.1 Space of matrices as a linear space

The set of all $m \times n$ matrices is denoted

$$\mathbb{R}^{m \times n} \coloneqq \left\{ X : X \in \mathbb{R}^{\{1, \ldots, m\} \times \{1, \ldots, n\}} \right\}.$$

We will define matrix addition and scalar multiplications so that $(\mathbb{R}^{m \times n}, (\mathbb{R}, +_{\mathbb{R}}, \cdot_{\mathbb{R}}), +, \cdot)$ is a linear space.

Proposition 38. *Given $X, Y \in \mathbb{R}^{m \times n}$ and $\lambda \in \mathbb{R}$, define $+ : \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ and $\cdot : \mathbb{R} \times \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ as*

$$X + Y \equiv + (X, Y) \coloneqq [x_{ij} + y_{ij}],$$
$$\lambda \cdot X \equiv \cdot (\lambda, X) \coloneqq [\lambda \cdot_{\mathbb{R}} x_{ij}].$$

Then, $(\mathbb{R}^{m \times n}, (\mathbb{R}, +_{\mathbb{R}}, \cdot_{\mathbb{R}}), +, \cdot)$ is a linear space of dimension $m \times n$ with zero given by the zero matrix, the additive inverse of any $X \in \mathbb{R}$ is $-1 \cdot X$, and the multiplicative identity is 1.

Let $E_{ij} \in \mathbb{R}^{m \times n}$ be a matrix whose entries are all zero except the ijth entry which is equal to 1 . The set $\{E_{ij}\}_{i=1,\ldots,m, j=1,\ldots,n}$ is a basis of $\mathbb{R}^{m \times n}$ and is the *canonical basis* of the linear space $\mathbb{R}^{m \times n}$.

Given any $X \in \mathbb{R}^{m \times n}$, a matrix Xcan be viewed as either a collection column vectors, $\{\mathbf{x}^j\}_{j \in \{1\ldots, n\}}$,

or a collection of row vectors, $\{\mathbf{x}_i\}_{i\in\{1,\dots,m\}}$; i.e.,

$$X \equiv \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x}^1 \cdots \mathbf{x}^n \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{bmatrix},$$

where

$$\mathbf{x}^j \coloneqq \begin{bmatrix} x_{1j} \\ \vdots \\ x_{mj} \end{bmatrix}_{m\times 1}, \ \mathbf{x}_i \coloneqq \begin{bmatrix} x_{i1} & \cdots & x_{in} \end{bmatrix}_{1\times n}.$$

3.1.2 Multiplication

Given $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times r}$, their *matrix multiplication* (or *product*), denoted XY, is defined as

$$XY \coloneqq \left[\sum_{k=1}^{n} x_{ik}y_{kj}\right]_{m\times r} = \begin{bmatrix} \sum_{k=1}^n x_{1k}y_{k1} & \cdots & \sum_{k=1}^n x_{1k}y_{kr} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^n x_{mk}y_{k1} & \cdots & \sum_{k=1}^n x_{mk}y_{kr} \end{bmatrix}.$$

Thus, matrix multiplication is defined only for matrices whose number of columns and rows are the same, called *conformable* matrices. Matrix multiplication operation is not commutative; i.e.,

$$XY \neq YX.$$

We therefore distinguish between *pre-* and *post-multiplications* (or *left-* and *right-multiplication*)—*pre-multiplying* (or *left-multiplying*) X by Y means YX and *post-multiplying* (or *right-multiplying*) X by Y means XY.

Given a matrix $X \in \mathbb{R}^{m\times n}$, let $D^k \equiv (d_{ii}) \in \mathbb{R}^{k\times k}$ denote a $k \times k$ diagonal matrix. Then,

$$XD^n = \begin{bmatrix} d_{11}\mathbf{x}^1 & \cdots & d_{nn}\mathbf{x}^n \end{bmatrix}, \ D^mX = \begin{bmatrix} d_{11}\mathbf{x}_1 \\ \vdots \\ d_{mm}\mathbf{x}_m \end{bmatrix}.$$

Let $\mathbf{1}^j$ denote a column vector whose jth entry is 1 and all other entries are zeros. Similarly, let $\mathbf{1}_i$ denote a row vector whose ith entry is 1 and all other entires are zeros. Then,

$$X\mathbf{1}^j = \mathbf{x}^j, \ \mathbf{1}_iX = \mathbf{x}_i$$

and so

$$\mathbf{1}_iX\mathbf{1}^j = x_{ij}.$$

These are useful to keep in mind in econometrics (and especially when coding).

Another important fact is that $XY = 0$ need not imply that $A = 0$ or $B = 0$ unlike multiplication of real numbers.

Proposition 39. *Suppose $X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times r}$, $Z \in \mathbb{R}^{r\times s}$ and $\lambda \in \mathbb{R}$.*

- *Matrix product is distributive with respect to addition; i.e., $X(Y + Z) = XY + XZ$ and $(X + Y)Z = XZ + YZ$.*
- *Matrix product is associative; i.e., $X(YZ) = (XY)Z$.*

- $X(\lambda Y) = \lambda(XY)$.
- *Conformable identity matrix is the multiplicative identity; i.e.,* $\mathbb{I}_m X = X\mathbb{I}_n = X$.
- *Zero matrix is absorbent; i.e.,* $[0]_{k\times m}X = [0]_{k\times n}$ *and* $X[0]_{n\times k} = [0]_{m\times k}$.
- *If* X *and* Y *are upper (resp. lower) triangular, then* $XY = YX$ *is also upper (resp. lower) triangular.*

For square matrices, we can define repeated products or *powers* of matrices. Given any $X \in \mathbb{R}^{n\times n}$ and any $k \in \mathbb{N}$,

$$X^k \coloneqq X\overbrace{\times\cdots\times}^{k \text{ times}}X,\ X^0 \coloneqq \mathbb{I}_n.$$

A matrix X is *idempotent* if $X^2 = X$.

Matrix product as a linear transformation Let $V = \mathbb{R}^{n\times r}$ and $W = \mathbb{R}^{m\times r}$ and define $L_X : \mathbb{R}^{n\times r} \to \mathbb{R}^{m\times r}$ as $L_X(Y) \coloneqq XY$ for some $X \in \mathbb{R}^{m\times n}$; i.e., we pre-multiply "vectors" in $\mathbb{R}^{n\times r}$ by X. Then, for any $Y, Z \in V$ and $\alpha \in \mathbb{R}$, we have

$$L_X(\alpha\cdot Y + Z) = X(\alpha\cdot Y + Z) = \alpha XY + XZ = \alpha\cdot L_X(Y) + L_X(Z).$$

Thus, matrix $X \in \mathbb{R}^{m\times n}$ is a linear transformation that maps vectors from linear space $\mathbb{R}^{n\times r}$ to linear space $\mathbb{R}^{m\times r}$. Since we can write XY as a linear combination of columns, $\mathbf{x}^j$, of X,

$$L_X(Y) = XY = \begin{bmatrix} \sum_{j=1}^n y_{j1}\mathbf{x}^j & \cdots & \sum_{j=1}^n y_{jr}\mathbf{x}^j \end{bmatrix}_{m\times r},$$

we can interpret pre-multiplication as a column transformation.

Suppose we now define $M_X : \mathbb{R}^{r\times m} \to \mathbb{R}^{r\times n}$ as $M_X(Z) \coloneqq ZX$ for some $X \in \mathbb{R}^{m\times n}$; i.e., we post-multiply "vectors" in $\mathbb{R}^{r\times m}$ by X. Since we can write ZX as a linear combination of rows, $\mathbf{x}_i$, of X,

$$M_X(Z) = ZX = \begin{bmatrix} \sum_{i=1}^m z_{ri}\mathbf{x}_i \\ \vdots \\ \sum_{i=1}^m z_{ri}\mathbf{x}_i \end{bmatrix}_{r\times n},$$

we can interpret post-multiplication as a row transformation.

Linear transformations as matrices Let V and W be linear spaces (over $\mathbb{R}$) with $\dim V = n$ and $\dim W = m$, respectively. Fix bases $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ for V and $T = \{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ for W. Take any $L \in \mathcal{L}(V, W)$. Then, for each $j = \{1, \ldots, n\}$, since $L(\mathbf{s}_j) \in W$, there exists $(\alpha_{ij})_{i=1}^m \in \mathbb{R}^m$ such that

$$L(\mathbf{s}_j) = \sum_{i=1}^m \alpha_{ij}\mathbf{t}_i.$$

Define

$$M_{V,W}(L) \coloneqq [\alpha_{ij}] = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{m1} & \cdots & \alpha_{mn} \end{bmatrix}_{m\times n}$$

so that jth column of $M_{V,W}(L)$ is the coordinates of $L(\mathbf{s}_j)$ with respect to T. That is,

$$L(\mathbf{x}) = M_{V,W}(L)\,\mathbf{x}.$$

This tells us that any linear transformation from V to W is equivalent to matrices once we fixed the two bases.

Proposition 40. *Let V and W be linear spaces with $\dim V = n$ and $\dim W = m$. Then, the space of linear transformations from V to W, $\mathcal{L}(V,W)$ is isomorphic to the space of $m \times n$ matrices, $\mathbb{R}^{m\times n}$. Moreover, if $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is a basis for V and $T = \{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ is a basis for W, then*

$$M_{V,W} \in \mathcal{L}\left(\mathcal{L}\left(V,W\right), \mathbb{R}^{m\times n}\right)$$

and $M_{V,W}$ is an isomorphism from $\mathcal{L}(V,W)$ to $\mathbb{R}^{m\times n}$.

Geometric interpretation of matrix product Any $(z_1, z_2) \in \mathbb{R}^2$ can be expressed as a point on a graph. Suppose we consider linear transformations $L : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$L\left(\mathbf{z}\right) \coloneqq X\mathbf{z} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} x_{11}z_1 + x_{12}z_2 \\ x_{21}z_1 + x_{22}z_2 \end{bmatrix}.$$

Consider the linear transformation of the unit square by pre-multiplication using matrix X. The vertices in the unique square is then mapped to a parallelogram:

$$\begin{aligned} &(0,0) \mapsto (0,0), && (1,0) \mapsto (x_{11}, x_{21}), \\ &(1,1) \mapsto (x_{11}+x_{12}, x_{21}+x_{22}), && (0,1) \mapsto (x_{12}, x_{22}). \end{aligned}$$

Thus, linear transformation here can be thought of as a change from the Cartesian coordinate system to a new grid system with axes that are generally not at right angels, and each unit square of the grid is transformed into a parallelogram.

The following are some standard transformations.

- Identity map: $X = \mathbb{I}_2$.
- Rotation through angle θ: $X = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$.
- Reflection about y axis: $X = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$.
- Reflection about x axis: $X = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.
- Reflection about $y = x$: $X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

3.1.3 Span

Recall that any matrix $X \in \mathbb{R}^{m\times n}$ can be viewed as a collection of (row or column) vectors. Thus, we can consider the linear space that the vectors span. The linear space spanned by the columns of X is the *column space* (or *image*) of X; i.e.,

$$\mathrm{im}\left(X\right) \coloneqq \mathrm{span}\left(\left\{\mathbf{x}^1, \ldots, \mathbf{x}^n\right\}\right) \equiv \left\{ \sum_{j=1}^{n} \lambda_j \mathbf{x}^j : \lambda_j \in \mathbb{R}\ \forall j \in \{1, \ldots, n\} \right\}.$$

The *column rank* of X is the rank of the column space of X (recall that rank of a linear space equals the cardinality of (any) basis of that space). The linear space spanned by the rows of X, i.e., $\mathrm{span}(\{\mathbf{x}_1, \ldots, \mathbf{x}_m\})$, is the *row space* of X. The *row rank* of X is the rank of the row space of X.

Proposition 41. *For any $X \in \mathbb{R}^{m\times n}$,* $\text{rank}(\text{span}(\{\mathbf{x}^1, \ldots, \mathbf{x}^n\})) = \text{rank}(\text{span}(\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}))$.

The proposition above means that it is unambiguous to the rank of matrix X without reference to row or column ranks. Hence, we simply write $\text{rank}(X)$ to mean the rank of the column (or row) space of matrix X.

Observe that $\text{rank}(\mathbb{I}_n) = n$ and each column (resp. row) are bases of column (resp. row) spaces.

Proposition 42. *For any $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times r}$,*

$$\text{rank}\,(XY) \leq \min\{\text{rank}\,(X), \text{rank}\,(Y)\}.$$

Proof. Suppose $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{n\times r}$ and recall that each column of $Z \coloneqq XY$ is a linear combination of columns of X. Hence, $\text{im}(Z) \subseteq \text{im}(X)$ and $\text{im}(Z)$ is a linear subspace of $\text{im}(X)$. Therefore, by Corollary 4,

$$\text{rank}\,(XY) = \dim(Z) \leq \dim(X) = \text{rank}\,(X).$$

Similarly, recall that each row of XY is a linear combination of the rows in Y. Hence,

$$\text{span}\,(\{\mathbf{z}_1, \ldots, \mathbf{z}_m\}) \subseteq \text{span}\,(\{\mathbf{y}_1, \ldots, \mathbf{y}_m\})$$

and $\text{span}(\{\mathbf{z}_1, \ldots, \mathbf{z}_m\})$ is a linear subspace of $\text{span}(\{\mathbf{y}_1, \ldots, \mathbf{y}_m\})$. Therefore, by Corollary 4,

$$\text{rank}\,(XY) = \dim(Z) \leq \dim(Y) = \text{rank}\,(Y).$$

Together, these imply that the desired inequality. ■

Corollary 8. *Rank of an $m \times n$ matrix is always smaller than both m and n.*

Proof. By the previous proposition and the fact that $\text{rank}(\mathbb{I}_n) = n$,

$$\text{rank}\,(X) = \text{rank}\,(X\mathbb{I}_n) \leq \min\{\text{rank}\,(X), \text{rank}\,(\mathbb{I}_n)\} = \min\{\text{rank}\,(X), n\},$$
$$\text{rank}\,(X) = \text{rank}\,(\mathbb{I}_m X) \leq \min\{\text{rank}\,(\mathbb{I}_m), \text{rank}\,(X)\} = \min\{m, \text{rank}\,(X)\}.$$

Together, these imply that

$$\text{rank}\,(X) \leq \min\{m, n\}.$$ ■

We say that matrix $X \in \mathbb{R}^{m\times n}$ has *full rank* if $\text{rank}(X) = \min\{m, n\}$.

Proposition 43. *Suppose $X \in \mathbb{R}^{m\times n}$ has rank $k \leq \min\{m, n\}$. Then, X has k linearly independent columns and rows.*

The *null space* (or the kernel) of matrix $X \in \mathbb{R}^{m\times n}$ is the solutions to the homogenous system of equations, $X\mathbf{z} = \mathbf{0}$; i.e.,

$$\text{null}\,(X) = \left\{\mathbf{z} \in \mathbb{R}^{n\times 1} : X\mathbf{z} = \mathbf{0}\right\}.$$

The dimension of the null space of X is the *nullity of X*, denoted $\text{nullity}(X) \coloneqq \dim(\text{null}(X))$.

Theorem 7 (Rank-nullity theorem)**.** *Suppose $X \in \mathbb{R}^{m\times n}$, then*

$$\text{rank}\,(X) + \text{nullity}\,(X) = n.$$

3.1.4 Inverse

An $n \times n$ square matrix X is *invertible* (or *nonsingular*) if there exists a matrix $Y \in \mathbb{R}^{n\times n}$ such that

$$XY = YX = \mathbb{I}_n.$$

If it exists, we call Y the *inverse matrix of* X and denote it as X^{-1}. Inverse matrices are unique whenever they exist and has the property that $(X^{-1})^{-1} = X$.[17] Matrices that are not invertible are *singular*.

Proposition 44. *Suppose $X, Y \in \mathbb{R}^{n\times n}$ are invertible, then their matrix product is also invertible and*

$$(XY)^{-1} = Y^{-1}X^{-1}.$$

We will not go over the general method of computing inverse matrices.[18]

Proposition 45. *A triangular matrix (upper or lower) is nonsingular if and only if its diagonal entires are nonzero. The inverse of an upper (resp. lower) triangular matrix is also upper (resp. lower) triangular whenever they exist.*

Above implies that a diagonal matrix is invertible if and only if the diagonal elements are nonzero. Thus, $\mathbb{I}$ is always invertible.

Proposition 46. *Suppose $X \in \mathbb{R}^{n\times n}$ is an invertible, diagonal matrix. Then,*

$$X^{-1} = \operatorname{diag}\left(\frac{1}{a_{11}}, \ldots, \frac{1}{a_{nn}}\right).$$

Proposition 47. *A square matrix $X \in \mathbb{R}^{n\times n}$ is invertible if and only if* $\operatorname{rank}(X) = n$.

3.1.5 Transpose

Given $X \in \mathbb{R}^{m\times n}$, the *transpose of* X, denoted $X^\top$ or X', is the matrix obtained by placing the ijth entries into the jith positions instead; i.e.,

$$X^\top \coloneqq [a_{ji}]_{n\times m}.$$

Note $(X^\top)^\top = X$, transpose of a real number (i.e., a 1×1 matrix) is itself, and transpose of a row vector is a column vector (and vice versa). We say that a matrix X is *symmetric* if it equals its transpose; i.e., $X = X^\top$. Symmetric matrices must be square matrices. A diagonal matrix is symmetric. Symmetries is preserved under matrix addition (and subtraction).

Proposition 48. *Suppose $X \in \mathbb{R}^{m\times n}$, $Y \in \mathbb{R}^{n\times r}$ and $\lambda \in \mathbb{R}$.*

- $(\lambda X)^\top = \lambda X^\top$.
- $(X + Y)^\top = X^\top + Y^\top$.
- $(XY)^\top = Y^\top X^\top$.
- $(X^{-1})^\top = (X^\top)^{-1}$ *whenever X is invertible.*
- $XX^\top$ *and* $X^\top X$ *are symmetric matrices.*
- *If X is upper (resp. lower) triangular, then $X^\top$ is lower (resp. upper) triangular.*

[17] Toward a contradiction, suppose X has two distinct inverses Y and Z. But $Y = Y\mathbb{I} = Y(XZ) = (YX)Z = \mathbb{I}Z = Z$; a contradiction.

[18] If you need to go over this, take a look at Sundaram §1.3.5.

3.1.6 Trace

Given $X \in \mathbb{R}^{n\times n}$, the *trace of* X, denoted $\text{tr}(X)$ is the sum of its diagonals; i.e.,

$$\text{tr}\,(X) \coloneqq \sum_{i=1}^{n} x_{ii}.$$

Proposition 49. *Suppose* $X, Y \in \mathbb{R}^{n\times n}$, $\lambda \in \mathbb{R}$,

- *Trance is a linear transformation:* $\text{tr}(\lambda X + Y) = \lambda \text{tr}(X) + \text{tr}(Y)$.
- $\text{tr}(X) = \text{tr}(X^\top)$.
- *If* AB *and* BA *are square matrices (but not necessarily* A *and* B*), then* $\text{tr}(AB) = \text{tr}(BA)$.
- *Invariant under cyclic permutation:* $\text{tr}(ABCD) = \text{tr}(BCDA) = \text{tr}(CDAB) = \text{tr}(DABC)$.

3.1.7 Determinant

The *determinant* of a square matrix $X \in \mathbb{R}^{n\times n}$, denoted $\det(X)$ (or $|X|$), is an a matrix defined inductively in the following way:

(i) for a 1×1 matrix $X = x_{11}$, define its determinant as $\det(X) \coloneqq x_{11}$;

(ii) for an $n \times n$ matrix with $n \geq 2$, define its determinant via the *cofactor of expansion* of X along the first row)

$$\det\,(X) \coloneqq \sum_{j=1}^{n} (-1)^{1+j}\, a_{ij} \det\,(X_{-1,-j})\,,$$

where $X_{-i,-j}$ is the matrix X with the ith row and jth columns eliminated.[19]

This formula means that

$$X = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \Rightarrow \det\,(X) = a_{11}a_{22} - a_{12}a_{21}$$

and

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

$$\Rightarrow \det\,(X) = a_{11} \det\left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}\right) - a_{12} \det\left(\begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix}\right) + a_{13} \det\left(\begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}\right)$$

$$= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

Proposition 50. *Suppose* $X \in \mathbb{R}^{n\times n}$.

- $\det(\mathbb{I}_n) = 1$.
- *If* X *is triangular, then* $\det(X) = \prod_{i=1}^{n} x_{ii}$.
- $\det(XY) = \det(X)\det(Y)$.
- *If* X *is invertible, then* $\det(X^{-1}) = 1/\det(X)$.
- *If any two columns of* X *are equal, then* $\det(X) = 0$.

Proposition 51. *A square matrix* $X \in \mathbb{R}^{n\times n}$ *is invertible if and only if* $\det(X) \neq 0$.

[19] In fact, we can define determinant using cofactor expansion along any arbitrary rows or columns of X.

Geometric interpretation of determinant Recall that a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$ reshapes a unit square to a parallelogram in Cartesian coordinate system. It turns out that the area of the parallelogram produced by transformation the unit square via matrix X is equal to $|\det(X)|$ (the sign of $\det(X)$ tells us about the orientation of the parallelogram compared to the original square). All of the standard transformations in the example are associated with matrices with determinant of one. This means that these transformations preserve area. In contrast, consider the following matrix, for some $k_1, k_2 \in \mathbb{R}$,

$$X = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix} \Rightarrow |\det(X)| = |k_1 k_2|.$$

This linear transformation scales area by a factor $k_1 k_2$; it scales in the x-axis direction by a factor k_1 and in the y-axis direction of a factor k_2.

Observe also that the matrices associated with the aforementioned standard linear transformations are all invertible/nonsingular. This means that each point is mapped to a unique point so that the transformations are reversible. In contrast, linear transformation via singular matrices cannot be reversed uniquely. For example, consider the matrix

$$X = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \Rightarrow |\det(X)| = 0.$$

Observe that every point in (x, y) is mapped to some point on the same line $y = 2x$. This means that there are many points that are mapped to the same point on the line $y = 2x$. For example,

$$(2, 2), \left(3, \frac{1}{2}\right) \mapsto (4, 8).$$

3.1.8 Kronecker product

Kronecker product is a function $\otimes : \mathbb{R}^{m\times n} \times \mathbb{R}^{r\times s} \to \mathbb{R}^{rm\times ns}$ defined by

$$X \otimes Y \coloneqq \begin{bmatrix} x_{11}Y & \cdots & x_{1n}Y \\ \vdots & \ddots & \vdots \\ x_{m1}Y & \cdots & x_{mn}Y \end{bmatrix}.$$

(More generally, Kronecker product is a special case of a linear transformation, called *tensor product*, applied to linear space of matrices.)

Example 20. Consider a panel data with dimensions $i \in \{1, \ldots, N\}$ and $t \in \{1, \ldots, T\}$:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\gamma} + \alpha_i + \delta_t + u_{it},$$

where $y_{it} \in \mathbb{R}$ is the dependent variable, $\mathbf{x}_{it} \in \mathbb{R}^k$ is a column vector of independent variables, $\boldsymbol{\gamma} \in \mathbb{R}^k$ is the column vector of coefficients, $\alpha_i \in \mathbb{R}$ is the fixed effect, δ_t is the "time dummy" and u_{it} is the error term. To express the data in matrix form, we stack observations—let's say first over

$t \in \{1, \ldots, T\}$ for each $i \in \{1, \ldots, N\}$:

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{2T} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix}_{NT\times 1} = \underbrace{\begin{bmatrix} \mathbf{x}_{11}' \\ \vdots \\ \mathbf{x}_{1T}' \\ \mathbf{x}_{21}' \\ \vdots \\ \mathbf{x}_{2T}' \\ \vdots \\ \mathbf{x}_{N1}' \\ \vdots \\ \mathbf{x}_{NT}' \end{bmatrix}_{NT\times k}}_{:=\mathbf{X}_1} \boldsymbol{\gamma}_{k\times 1} + \underbrace{\begin{bmatrix} \iota_T & \mathbf{0}_T & \cdots & \cdots & \mathbf{0}_T \\ \mathbf{0}_T & \iota_T & \mathbf{0}_T & \cdots & \mathbf{0}_T \\ \mathbf{0}_T & \mathbf{0}_T & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mathbf{0}_T \\ \mathbf{0}_T & \mathbf{0}_T & \cdots & \mathbf{0}_T & \iota_T \end{bmatrix}_{NT\times N}}_{:=\mathbf{X}_2} \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}_{N\times 1}}_{:=\boldsymbol{\alpha}}$$

$$+ \underbrace{\begin{bmatrix} \mathbb{I}_T \\ \mathbb{I}_T \\ \vdots \\ \mathbb{I}_T \end{bmatrix}_{NT\times T}}_{:=\mathbf{X}_3} \underbrace{\begin{bmatrix} \delta_1 \\ \vdots \\ \delta_T \end{bmatrix}_{T\times 1}}_{:=\boldsymbol{\delta}} + \underbrace{\begin{bmatrix} u_{11} \\ \vdots \\ u_{1T} \\ u_{21} \\ \vdots \\ u_{2T} \\ \vdots \\ u_{N1} \\ \vdots \\ u_{NT} \end{bmatrix}_{NT\times 1}}_{:=\mathbf{u}},$$

where $\iota_T \in \mathbb{R}^T$ column vector of ones, $\mathbf{0}_T \in \mathbb{R}^T$ is a column vector of zeros, and $\mathbb{I}_T \in \mathbb{R}^{T\times T}$ is an identity matrix. We can write above equivalently as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\gamma} + \underbrace{(\mathbb{I}_N \otimes \iota_T)}_{=\mathbf{X}_2}\boldsymbol{\alpha} + \underbrace{(\iota_N \otimes \mathbb{I}_T)}_{=\mathbf{X}_3}\boldsymbol{\delta} + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]_{NT\times(k+N+T)},$$
$$\boldsymbol{\beta} = \left[\boldsymbol{\gamma}', \boldsymbol{\alpha}', \boldsymbol{\delta}'\right]'_{(k+N+T)\times 1}.$$

Note that, given $K \coloneqq k+N+T$,

$$(\mathbf{X}'\mathbf{X})_{K\times K} = [\tilde{\mathbf{x}}_{11}\cdots\tilde{\mathbf{x}}_{1T}\cdots\cdots\tilde{\mathbf{x}}_{N1}\cdots\tilde{\mathbf{x}}_{NT}]\begin{bmatrix}\tilde{\mathbf{x}}'_{11}\\ \vdots\\ \tilde{\mathbf{x}}'_{1T}\\ \vdots\\ \vdots\\ \tilde{\mathbf{x}}_{N1}\\ \vdots\\ \tilde{\mathbf{x}}_{NT}\end{bmatrix} = \sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it},$$

$$(\mathbf{X}'\mathbf{u})_{K\times 1} = [\tilde{\mathbf{x}}_{11}\cdots\tilde{\mathbf{x}}_{1T}\cdots\cdots\tilde{\mathbf{x}}_{N1}\cdots\tilde{\mathbf{x}}_{NT}]\begin{bmatrix}u_{11}\\ \vdots\\ u_{1T}\\ \vdots\\ \vdots\\ u_{N1}\\ \vdots\\ u_{NT}\end{bmatrix} = \sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}u_{it}.$$

3.1.9 Space of matrices as an inner product space

Suppose $\mathbf{y},\mathbf{z}\in\mathbb{R}^{n\times 1}$ and $X\in\mathbb{R}^{n\times n}$ is invertible. First, observe that Euclidean dot product can be written as matrix product:

$$\mathbf{y}\cdot\mathbf{z} = \sum_{i=1}^{n} y_i z_i = \begin{bmatrix} z_1 & \cdots & z_n \end{bmatrix}\begin{bmatrix} y_1\\ \vdots\\ y_n\end{bmatrix} = \mathbf{z}^\top\mathbf{y} = \mathbf{y}^\top\mathbf{z}.$$

Since $X\mathbf{y}, X\mathbf{z}\in\mathbb{R}^{n\times 1}$, we may consider the Euclidean dot product between the two:

$$X\mathbf{y}\cdot X\mathbf{z} = (X\mathbf{y})^\top X\mathbf{z} = \mathbf{y}^\top X^\top X\mathbf{z}.$$

The function $\langle\cdot,\cdot\rangle_X : \mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ defined by

$$\langle\mathbf{y},\mathbf{z}\rangle_X \coloneqq X\mathbf{y}\cdot X\mathbf{z} = \mathbf{y}^\top X^\top X\mathbf{z}.$$

is an inner product and is called an *inner product on* $\mathbb{R}^n$ *generated by* X. The Euclidean dot product is a special case in which $X=\mathbb{I}_n$. Weighted Euclidean dot product is a special case in which $X=\operatorname{diag}(w_1,\ldots,w_n)$.

Proposition 52. $(\mathbb{R}^{m\times n},\langle\cdot,\cdot\rangle)$ *is an inner product space where*

$$\langle X,Y\rangle \coloneqq \operatorname{tr}\left(Y^\top X\right).$$

The following provides a geometric connection between the null space and row/column spaces of a matrix.

Proposition 53. *Given $X \in \mathbb{R}^{m\times n}$,*

$$\begin{aligned}\operatorname{null}(X) &= (\operatorname{span}(\{\mathbf{x}_1,\dots,\mathbf{x}_m\}))^{\perp},\\ \operatorname{null}\left(X^{\top}\right) &= \left(\operatorname{span}\left(\left\{\mathbf{x}^1,\dots,\mathbf{x}^n\right\}\right)\right)^{\perp} = (\operatorname{im}(X))^{\perp},\end{aligned}$$

where the orthogonal complement is taken with respect to the Euclidean inner product.

Proof. Suppose $X \in \mathbb{R}^{m\times n}$ and the we take the inner product to be the Euclidean dot product.

Take any $\mathbf{y} \in \operatorname{null}(X)$. We first show that $\mathbf{y}$ is orthogonal to every vector in the row space of X. By definition of null space,

$$X\mathbf{y} = \mathbf{0} \Leftrightarrow \mathbf{x}_i \cdot \mathbf{y} = \langle \mathbf{x}_i, \mathbf{y}\rangle = 0 \ \forall i \in \{1,\dots,m\}.$$

Thus, each $\mathbf{x}_i$ (i.e., each row of matrix X) is orthogonal to $\mathbf{y}$. By definition, for any $\mathbf{x} \in \operatorname{span}(\{\mathbf{x}_1,\dots,\mathbf{x}_m\})$, there exists $(\lambda_i)_{i=1}^m \in \mathbb{R}^m$ such that $\mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{x}_i$. Thus,

$$\langle \mathbf{x}, \mathbf{y}\rangle = \left\langle \sum_{i=1}^m \lambda_i \mathbf{x}_i, \mathbf{y} \right\rangle = \sum_{i=1}^m \lambda_i \langle \mathbf{x}_i, \mathbf{y}\rangle = 0.$$

Hence, $\mathbf{y}$ is orthogonal to any element in the row space of X. Sine $\mathbf{y}$ was arbitrary, it follows that every element of $\operatorname{null}(X)$ is orthogonal to the row space of X. Conversely, take any $\mathbf{y}^{\perp}$ that is orthogonal to every row vector of X; i.e.,

$$\left\langle \mathbf{x}_i, \mathbf{y}^{\perp}\right\rangle = \mathbf{x}_i \cdot \mathbf{y}^{\perp} = 0 \ \forall i \in \{1,\dots,m\}.$$

This implies that $\mathbf{y}^{\perp} \in \operatorname{null}(X)$ since

$$X\mathbf{y}^{\perp} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} y_1^{\perp} \\ \vdots \\ y_n^{\perp} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m x_{1i} y_i^{\perp} \\ \vdots \\ \sum_{i=1}^m x_{mi} y_i^{\perp} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \cdot \mathbf{y}^{\perp} \\ \vdots \\ \mathbf{x}_m \cdot \mathbf{y}^{\perp} \end{bmatrix} = \mathbf{0}.$$

Since $\mathbf{y}^{\perp}$ was chosen arbitrary, it follows that any vector that is orthogonal to every tow vector of X is in the null space of X. We have now shown the first equality. To show the second equality, observe that the column space of X is equal to the row space of $X^{\top}$. ■

Recall that in any inner product space V, the zero space ($\{\mathbf{0}\}$) and V are orthogonal complement of one another. Thus, if $X \in \mathbb{R}^{n\times n}$, to say that $A\mathbf{x} = 0$ has only the trivial solution (of $\mathbf{x} = 0$) is equivalent to saying that the orthogonal complement of the null space of X is all of $\mathbb{R}^n$, or, equivalently (by the proposition above), that the row space of X is all of $\mathbb{R}^n$. This gives the following characterisation.

Proposition 54. *The following are equivalent.*

- *A matrix $X \in \mathbb{R}^{n\times n}$ is invertible.*
- $(\operatorname{null}(X))^{\perp} = \mathbb{R}^n$.
- $(\operatorname{span}(\{\mathbf{x}_1,\dots,\mathbf{x}_m\}))^{\perp} = \{\mathbf{0}\}$.

Theorem 8 (QR-decomposition)**.** *Suppose $X \in \mathbb{R}^{m\times n}$ has linearly independent column vectors; i.e., $\{\mathbf{x}^1,\dots,\mathbf{x}^n\}$ are linearly independent. Then, X can be decomposed as*

$$X = QR,$$

where $Q \in \mathbb{R}^{m\times n}$ is a matrix with orthonormal column vectors and $R \in \mathbb{R}^{n\times n}$ is an invertible upper triangular matrix.

Proof. Let $X \in \mathbb{R}^{m\times n}$ be a matrix with linearly independent column vectors $\{\mathbf{x}^1, \dots, \mathbf{x}^n\}$. Then, $(\mathrm{im}(X), \langle\cdot,\cdot\rangle)$ is an inner product space with $\langle\cdot,\cdot\rangle$ being the Euclidean dot product; moreover, $\{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ is a basis of $\mathrm{im}(X)$. Since $\{\mathbf{x}^1, \dots, \mathbf{x}^n\}$ is linearly independent, we can apply the Gram-Schmidt process to obtain orthonormal an basis, $S \coloneqq \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$. By Proposition 31, we can write

$$\mathbf{x}^i = \sum_{j=1}^{n} \left\langle \mathbf{x}^i, \mathbf{s}_j \right\rangle \mathbf{s}_j \ \forall i \in \{1, \dots, n\}.$$

Thus,

$$X = \begin{bmatrix} \mathbf{x}^1 & \cdots & \mathbf{x}^n \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{bmatrix}}_{\coloneqq Q} \underbrace{\begin{bmatrix} \langle \mathbf{x}^1, \mathbf{s}_1 \rangle & \cdots & \langle \mathbf{x}^n, \mathbf{s}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{x}^1, \mathbf{s}_n \rangle & \cdots & \langle \mathbf{x}^n, \mathbf{s}_n \rangle \end{bmatrix}}_{\coloneqq R}.$$

By construction, for any $j \geq 2$, the vector $\mathbf{s}_j$ is orthogonal to $\mathbf{x}^1, \dots, \mathbf{x}^{j-1}$ so that

$$R = \begin{bmatrix} \langle \mathbf{x}^1, \mathbf{s}_1 \rangle & \langle \mathbf{x}^2, \mathbf{s}_1 \rangle & \cdots & \cdots & \langle \mathbf{x}^n, \mathbf{s}_1 \rangle \\ 0 & \langle \mathbf{x}^2, \mathbf{s}_2 \rangle & \cdots & \cdots & \langle \mathbf{x}^n, \mathbf{s}_2 \rangle \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \langle \mathbf{x}^n, \mathbf{s}_n \rangle \end{bmatrix}.$$

That is, R is an upper triangular matrix. ∎

Since $X \in \mathbb{R}^{n\times n}$ is invertible if and only if it has linearly independent columns, it follows that every invertible matrix has a QR-decomposition.

3.2 System of linear equations

A system of a linear equations take the form:

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

where $a_{ij}, b_j \in \mathbb{R}$ for all $i, j \in \{1, \dots, n\} \times \{1, \dots, m\}$ and the unknowns are $x_1, \dots, x_n \in \mathbb{R}$. The system of linear equations can be written succinctly as

$$A\mathbf{x} = \mathbf{b},$$

where $A = [a_{ij}]_{m\times n}$, $\mathbf{b} = (b_1, \dots, b_m)$ and $\mathbf{x} = (x_1, \dots, x_n)$ (remember the convention that unspecified vectors are row vectors). The following proposition characterised the existence as well as the unique of the solutions to a system of linear equations.

Proposition 55. *Suppose $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} = \mathbb{R}^{m\times 1}$. Then, the system of equation $A\mathbf{x} = \mathbf{b}$ has a solution if and only if*

$$\mathrm{rank}\left([A|\mathbf{b}]\right) = \mathrm{rank}\left(A\right), \tag{3.1}$$

where $[A|\mathbf{b}] \coloneqq [\mathbf{a}^1, \ldots, \mathbf{a}^n, \mathbf{b}] \in \mathbb{R}^{m\times n+1}$ *is the augmentation of matrix* A *by* $\mathbf{b}$ *(called the augmented matrix).*

If the system has a solution, then (i) the solution is unique if and only if A *is full rank; (ii) the system has infinitely many solutions if and only if* $\text{rank}(A) < n$.

Recall that $A\mathbf{x}$ is a linear combination of columns of X. The condition (3.1) means that $\mathbf{b}$ is a linear combination of the columns of A; i.e., the existence of a solution is equivalent to $\mathbf{b}$ belonging in the column space of A. Recalling Proposition 7, uniqueness of a solution (when one exists) depends on the dimension of the null space of A. In fact, one can think of $\text{nullity}(A)$ as the number of free parameters in the system of equations.

Proposition 56 (General solution of a linear equation)**.** *Given* $A \in \mathbb{R}^{m\times n}$ *and* $\mathbf{b} \in \mathbb{R}^{m\times 1}$*, suppose* $x^* \in \mathbb{R}^{n\times 1}$ *satisfies* $A\mathbf{x}^* = \mathbf{b}$*. Then, the set of all solutions satisfying the* $A\mathbf{x} = \mathbf{b}$ *is given by*

$$\{\mathbf{x}^* + \mathbf{x}_h : \mathbf{x}_h \in \text{null}\,(A)\}\,.$$

A system of linear equations $A\mathbf{x} = \mathbf{b}$ is *homogenous* if $\mathbf{b} = \mathbf{0}$ and *non-homogenous* if $\mathbf{b} \neq \mathbf{0}$. A vector $\mathbf{x}^*$ satisfying $A\mathbf{x}^* = \mathbf{b}$ is a *particular solution* of the linear equation $A\mathbf{x} = \mathbf{b}$. The result above tells us that the *general solution* of a non-homogenous system of linear equations, $\{\mathbf{x} : A\mathbf{x} = \mathbf{b}\}$. is the sum of any particular solution ($\mathbf{x}^*$) and the general solution of the related homogenous system ($\mathbf{x}_h$ such that $A\mathbf{x}_h = \mathbf{0}$).

When solving linear equations by hand, you can use the Gauss-Jordan elimination, which we will not cover here. In the special case in which A is an invertible square matrix, the (unique) solution is given by $\mathbf{x}^* \coloneqq A^{-1}\mathbf{b}$ and we have the following theorem to characterise the solution.

Theorem 9 (Cramer's Rule)**.** *Suppose* $A \in \mathbb{R}^{n\times n}$ *is invertible and* $\mathbf{b} \in \mathbb{R}^{n\times 1}$*. The* i*th entry of the vector* $\mathbf{x}^* \coloneqq A^{-1}\mathbf{b} \in \mathbb{R}^{n\times 1}$ *is given by*

$$\mathbf{x}_i^* = \frac{\det(A_i)}{\det(A)} \;\forall i \in \{1, \ldots, n\}\,,$$

where $A_i \in \mathbb{R}^{n\times n}$ *is the matrix obtained by replacing the* i*th column of* A *by* $\mathbf{b}$*.*

3.3 Least square solution as orthogonal projections

Consider a linear system of equations $X\mathbf{b} = \mathbf{y}$, where $X \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^{n\times 1}$ and $\mathbf{y} \in \mathbb{R}^{m\times 1}$. Suppose that this system of equation has no solution. That is, by our discussion about Proposition 55, $\mathbf{y} \notin \text{im}(X)$. We might find ourselves in such situation if $X\mathbf{b} = \mathbf{y}$ represents a theoretical relationship that is only observed in real data with some noise. One way to solve such an *inconsistent* system of equation is to look for $\mathbf{b}$ that comes "as close as possible" to being a solution in the sense that it minimises the the Euclidean distance between $X\mathbf{b} \in \text{im}(X)$ and $\mathbf{y}$. Thinking of $\mathbf{e^b} \coloneqq X\mathbf{b} - \mathbf{y}$ as errors, we wish to minimise $\|\mathbf{e^b}\| = \sqrt{(e_1^b)^2 + \cdots + (e_n^b)^2}$ with respect to $\mathbf{b}$. The solution to such a problem is called the *least squares solution* of $X\mathbf{b} = \mathbf{y}$ since it minimises the squares errors.

Define W to be the column space of X; i.e., $W \coloneqq \text{im}(X)$. For any $\mathbf{b} \in \mathbb{R}^{n\times 1}$ then $X\mathbf{b}$ is a linear combination of column vectors of X; i.e., $X\mathbf{b} \in W$ for any $\mathbf{b} \in \mathbb{R}^{n\times 1}$. By the Approximation Theorem (Theorem 6), the best approximation to $\mathbf{y}$ in W is the orthogonal projection of $\mathbf{y}$ on W, $\hat{\mathbf{y}} \coloneqq \text{proj}_W\mathbf{y}$. We now wish to find $\hat{\mathbf{b}} \in \mathbb{R}^{n\times 1}$ such that

$$X\hat{\mathbf{b}} = \text{proj}_W\mathbf{y}.$$

We will do so without computing $\text{proj}_W\mathbf{y}$ explicitly. Observe that

$$\mathbf{y} - X\hat{\mathbf{b}} = \mathbf{y} - \text{proj}_W\mathbf{y} = \text{proj}_{W^\top}\mathbf{y} \in W^\perp.$$

That is, $\mathbf{y} - X\hat{\mathbf{b}}$ is orthogonal to W, which is the column space of X. Then, by Proposition 53,

$$\mathbf{y} - X\hat{\mathbf{b}} \in \text{null}\left(X^\top\right) \Leftrightarrow X^\top\left(\mathbf{y} - X\hat{\mathbf{b}}\right) = \mathbf{0} \Leftrightarrow X^\top\mathbf{y} = X^\top X\hat{\mathbf{b}}.$$

The last expression is called the *normal system* associated with $X\mathbf{b} = \mathbf{y}$. Recall that we are concerned with a linear system of equations $X\mathbf{b} = \mathbf{y}$ with no solutions. The point of all of this is the following result.

Proposition 57. *A solution to a normal system associated with $X\mathbf{b} = \mathbf{y}$ always exists.*

We first prove the following lemma.

Lemma 3. *For any matrix $X \in \mathbb{R}^{m\times n}$,*

$$\text{null}\left(X^\top X\right) = \text{null}\left(X\right),\ \text{im}\left(X^\top X\right) = \text{im}\left(X^\top\right).$$

Proof. For any $\mathbf{b} \in \text{null}(X)$, then $X\mathbf{b} = \mathbf{0}$ so that $\mathbf{X}^\top\mathbf{X}\mathbf{b} = \mathbf{X}^\top\mathbf{0} = \mathbf{0}$ and $\mathbf{b} \in \text{null}(X^\top X)$. Thus, $\text{null}(X) \subseteq \text{null}(X^\top X)$. Conversely, take any $\mathbf{b} \in \text{null}(X^\top X)$, then

$$X^\top X\mathbf{b} = \mathbf{0} \Rightarrow \mathbf{b}^\top X^\top X\mathbf{b} = \mathbf{b}^\top\mathbf{0} = \mathbf{0} \Rightarrow (X\mathbf{b})^\top (X\mathbf{b}) = \langle X\mathbf{b}, X\mathbf{b}\rangle = \|X\mathbf{b}\|^2 = \mathbf{0}.$$

By positivity, it follows that $X\mathbf{b} = 0$. Hence, $\text{null}(X^\top X) \subseteq \text{null}(X)$. Together, we have $\text{null}(X^\top X) = \text{null}(X)$. Since $X^\top X$ is symmetric,

$$\text{im}\left(X^\top X\right) = \text{im}\left(\left(X^\top X\right)^\top\right) = \left(\text{null}\left(\left(X^\top X\right)^\top\right)\right)^\perp = \left(\text{null}\left(X^\top X\right)\right)^\perp.$$

where the penultimate equality uses Proposition 53. Since $\text{null}(X^\top X) = \text{null}(X)$,

$$\text{im}\left(X^\top X\right) = \left(\text{null}\left(X\right)\right)^\perp = \text{im}\left(X^\top\right),$$

where the last equality uses Proposition 53 again. ■

Proof of Proposition 57. By Proposition 55, we know that a solution to the normal system exists if $X^\top\mathbf{y} \in \text{im}(X^\top X)$. Since $X^\top\mathbf{y}$ is a linear combination of columns of $X^\top$, $X^\top\mathbf{y} \in \text{im}(X^\top)$. The lemma above tells us that $X^\top\mathbf{y} \in \text{im}(X^\top X)$. Hence, the solution to the normal system must exist. ■

Although we know that a solution to normal system exists, we do not know whether the solution is unique. But, by Proposition 55 again, if $X^\top X \in \mathbb{R}^{n\times n}$ is invertible, we know that the solution must be unique. In this case, we have

$$\hat{\mathbf{b}} = \left(X^\top X\right)^{-1} X^\top\mathbf{y}$$

and it also follows that

$$\text{proj}_W\mathbf{y} = X\left(X^\top X\right)^{-1} X^\top\mathbf{y}.$$

You may recall that $\hat{\mathbf{b}}$ is the OLS estimator. We see here that OLS gives us the vector in the column space of X that minimises the distance (under the metric induced by the Euclidean norm) between it and vector $\mathbf{y}$.

3.4 Eigenvalues and eigenvectors

Recall that $\mathbb{C}$ is the set of complex numbers. Let $\mathbb{C}^{m\times n}$ be the set of all $m \times n$ matrices whose entries are complex numbers. Given any $X \in \mathbb{C}^{n\times n}$, a scalar $\lambda \in \mathbb{C}$ is an *eigenvalue of* X if

$$\exists \mathbf{z} \in \mathbb{C}^n \backslash \{\mathbf{0}\},\ X\mathbf{z} = \lambda \mathbf{z}.$$

A vector $\mathbf{z} \in \mathbb{C}^n \backslash \{\mathbf{0}\}$ is an *eigenvector of* X if

$$\exists \lambda \in \mathbb{C},\ X\mathbf{z} = \lambda \mathbf{z}.$$

Proposition 58. *$\lambda \in \mathbb{C}$ is an eigenvalue of $X \in \mathbb{C}^{n\times n}$ if and only if*

$$\det(\lambda \mathbb{I}_n - X) = 0.$$

Proof. By definition, $\lambda \in \mathbb{C}$ is an eigenvalue of X if $X\mathbf{z} = \lambda\mathbf{z} \Leftrightarrow (X - \lambda\mathbb{I}_n)\mathbf{z} = \mathbf{0}$ has a nonzero solution $\mathbf{z}$. Since $(X-\lambda\mathbb{I}_n)\mathbf{z}$ is a linear combination of the columns of $X-\lambda\mathbb{I}_n$, that there is a nonzero solution means that the columns of the matrix $\lambda\mathbb{I}_n - X$ are linearly dependent by Proposition 19. By Proposition 43, this means that $\text{rank}(X) < n$, which implies that $\lambda\mathbb{I}_n - X$ is not invertible (Proposition 47). This, in turn, is equivalent to $\det(\lambda\mathbb{I}_n - X) = 0$ by Proposition 51. ■

Since

$$\lambda\mathbb{I}_n - X = \begin{bmatrix} \lambda - x_{11} & -x_{12} & \cdots & -x_{1n} \\ -x_{21} & \lambda - x_{22} & \cdots & -x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -x_{n1} & -x_{n2} & \cdots & \lambda - x_{nn} \end{bmatrix},$$

the determinant of $\lambda\mathbb{I}_n - X$ is a polynomial in λ of degree n; i.e., $\exists (c_i)_{i=0}^n \in \mathbb{C}^{n+1}$ with $c_n \neq 0$,

$$\det(\lambda\mathbb{I}_n - X) = \sum_{i=0}^{n} c_i \cdot (\lambda)^i.$$

We refer to $P_X(\lambda) \coloneqq \det(\lambda\mathbb{I}_n - X)$ as the *characteristic polynomial* of X. The Fundamental Theorem of Algebra tells us that we can find all eigenvalues of X by setting the characteristic polynomial of X to 0 and solving for all its roots.

Theorem 10 (Fundamental Theorem of Algebra)**.** *Let $P : \mathbb{C} \to \mathbb{C}$ be a polynomial of degree n; i.e., $P(\lambda) = \sum_{i=0}^n c_i \cdot (\lambda)^i$ for $(c_i)_{i=0}^n \in \mathbb{C}^{n+1}$ with $c_n \neq 0$. Then, P has exactly n roots in $\mathbb{C}$ (counted with multiplicity); i.e.,*

$$\exists (\lambda_i)_{i=1}^n \in \mathbb{C}^n,\ P(\lambda) = c_n \prod_{i=1}^{n} (\lambda - \lambda_i)$$

Proposition 59. *Suppose matrix $X \in \mathbb{C}^{n\times n}$ has eigenvalues $(\lambda_i)_{i=1}^n \in \mathbb{C}^n$. Then,*

$$P_X(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i),$$
$$\det(X) = \prod_{i=1}^{n} \lambda_i,$$
$$\text{tr}(X) = \sum_{i=1}^{n} \lambda_i.$$

It follows from above that a matrix is singular (not invertible) if at least one eigenvalue is 0.

3.4.1 Geometric interpretation

Recall that $X \in \mathbb{R}^{2\times 2}$ can be thought of as mapping points to new points (except for the origin). However, it may be that some straight lines remains fixed before and after the transformation. That is, suppose what happens if $X\mathbf{z} = \lambda\mathbf{z}$ for some $\lambda \in \mathbb{R}$ so that every point (z_1, z_2) on a given line is mapped to another point on the same straight line through the origin. The equation or the direction of the straight line, $\mathbf{z}$, is the eigenvector, and the corresponding value of the constant λ (the scaling in the direction of the associated eigenvector) is the eigenvalue.

For example, consider the following variant of the matrix that reflects about the y axis:

$$X = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}.$$

This matrix has eigenvalues of -2 and 1 and corresponding eigenvectors of $(1, 0)$ and $(0, 1)$ respectively. This means that X scales by a factor of 2 in the direction of the x-axis but leaves the scale in the y-axis unchanged. Indeed, observe that

$$(0, 1) \mapsto (0, 1)\,, (1, 0) \mapsto (-2, 0)\,,\ (1, 1) \mapsto (-2, 1)\,.$$

Thus, this linear mapping doubles the area and indeed $|\det(X)| = 2$.

Note that eigenvalue of zero means that all points on the line associated with the eigenvector is mapped to the origin.

3.5 Diagonalisation

A matrix $X \in \mathbb{R}^{n\times n}$ is *diagonalisable* (in $\mathbb{R}$) if there exists an invertible matrix $P \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ such that

$$\Lambda = P^{-1}XP.$$

The diagonal entries of $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ are the n eigenvalues of X since

$$\begin{aligned}
\det(\lambda\mathbb{I}_n - X) &= \det\left(P^{-1}\right)\det(\lambda\mathbb{I}_n - X)\det(P) \\
&= \det\left(P^{-1}(\lambda\mathbb{I}_n - X)P\right) \\
&= \det\left(P^{-1}\lambda\mathbb{I}_n P - P^{-1}XP\right) \\
&= \det(\lambda\mathbb{I}_n - \Lambda) \\
&= \prod_{i=1}^{n}(\lambda - \lambda_i)\,,
\end{aligned}$$

where, in first line, we used the fact that $\det(P^{-1}) = 1/\det(P)$. Moreover, a diagonalisable matrix X can be decomposed as follows:

$$X = \left(PP^{-1}\right)X\left(PP^{-1}\right) = P\left(P^{-1}XP\right)P^{-1} = P\Lambda P^{-1}.$$

Proposition 60. *A matrix $X \in \mathbb{C}^{n\times n}$ is diagonalisable in $\mathbb{C}$ if and only if X has n linearly independent eigenvectors.*

Proposition 61. *Suppose matrix $X \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues $(\lambda_i)_{i=1}^n \in \mathbb{C}^n$. Then, the corresponding eigenvectors are linearly independent.*

Proposition 62. *A matrix $X \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues in $\mathbb{C}$, then X is diagonalisable in $\mathbb{C}$ so that $X = P\Lambda P^{-1}$, where P is a matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues.*

Proof. Let $\mathbf{p}^i$ denote eigenvectors corresponding to eigenvalue λ_i; i.e.,

$$X\mathbf{p}^i = \lambda_i \mathbf{p}^i \ \forall i \in \{1, \dots, n\}.$$

Define $P \coloneqq \begin{pmatrix} \mathbf{p}^1 & \dots & \mathbf{p}^n \end{pmatrix}$ and $\Lambda \coloneqq \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, we can succinctly write the above set of equation as

$$XP = P\Lambda.$$

By the previous proposition, eigenvectors of X are linearly independent, which implies that P is invertible. Thus, post-multiplying both sides by P^{-1} yields $X = P\Lambda P^{-1}$ as desired. ■

An matrix $X \in \mathbb{R}^{n\times n}$ is orthogonal if $X^\top X = \mathbb{I}_n$, which, in turn, implies that $X^{-1} = X^\top$.

Proposition 63. *A symmetric matrix $X \in \mathbb{R}^{n\times n}$ has all real eigenvalues, and there exists an orthogonal matrix $P \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ such that $\Lambda = P^{-1}XP = P^\top XP$.*

Remark 5. The characterisation of when a matrix can be diagonalised is an example of *spectral theorem*. More generally, a spectral theorem characterises when a linear transformation can be diagonalised. Note that a *spectrum* of a linear transformation is a generalisation of the set of eigenvalues of a matrix.

Proposition 64. *A matrix X is idempotent if and only if its eigenvalues are either 0 or 1, and the number of eigenvalues equal to 1 is $\operatorname{tr}(X)$.*

Proof. Since X is idempotent it has to be symmetric. Suppose that X is idempotent, λ is an eigenvalue and $\mathbf{z} \neq \mathbf{0}$ the corresponding eigenvector, i.e., $\lambda\mathbf{z} = X\mathbf{z}$, then

$$\lambda\mathbf{z} = X\mathbf{z} = XX\mathbf{z} = X(\lambda\mathbf{z}) = \lambda X\mathbf{z} = \lambda^2\mathbf{z}.$$

Since $\mathbf{z} \neq \mathbf{0}$, above implies $\lambda = \lambda^2$ and hence either $\lambda = 0$ or $\lambda = 1$.

Conversely, by Proposition 63, we may write $X = P^\top \Lambda P$, where Λ is a matrix of eigenvalues and P is the matrix of corresponding eigenvectors. Then,

$$X^2 = P^\top \Lambda P P^\top \Lambda P = P^\top \Lambda P P^{-1} \Lambda P = P^\top \Lambda^2 P,$$

where we used the fact that P is orthogonal. Sine each diagonal entry on Λ is either 0 or 1, it follows that $\Lambda^2 = \Lambda$. That is,

$$X^2 = P^\top \Lambda P = X.$$

Finally,

$$\operatorname{tr}(X) = \operatorname{tr}\left(P^\top \Lambda P\right) = \operatorname{tr}\left(\Lambda P P^\top\right) = \operatorname{tr}(\Lambda).$$

Since all diagonal entries in Λ are 0 or 1, the number of eigenvalues equal to 1 is $\operatorname{trace}(X)$. ■

Linear dynamic system Suppose we have a linear dynamic system; i.e.,

$$\mathbf{z}_t = X\mathbf{z}_{t-1}$$

for some $\mathbf{z}_t \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times n}$. Repeated substitution yields that

$$\mathbf{z}_t = X(X\mathbf{z}_{t-2}) = \cdots = X^t\mathbf{z}_0.$$

Computing X^t can be difficult generally. However, if X was diagonal, the computation is easy because

$$(\operatorname{diag}(x_1, \dots, x_n))^t = \operatorname{diag}\left(x_1^t, \dots, x_n^t\right).$$

Even if X was not diagonal, diagonalisation helps computation. To see this, suppose that matrix X is diagonalisable and that there exists a matrix of eigenvectors $P \in \mathbb{R}^{n\times n}$ and a corresponding diagonal matrix of distinct eigenvalues $\Lambda = \text{diag}(\lambda_1^t, \ldots, \lambda_n^t) \in \mathbb{R}^{n\times n}$ such that $X = P\Lambda P^{-1}$. Then,

$$X^t = \left(P\Lambda P^{-1}\right) \times \cdots \times \left(P\Lambda P^{-1}\right) = P\Lambda^t P^{-1}.$$

Hence,

$$\mathbf{z}_t = P\Lambda^t P^{-1}\mathbf{z}_0$$

so that $\mathbf{z}_t$ is a linear combination of the initial values $\mathbf{z}_0$ with the weights given by the eigenvalues and eigenvectors.

Defining $\hat{\mathbf{z}}_t \coloneqq P^{-1}\mathbf{z}_t$, we can also write

$$\hat{\mathbf{z}}_t = \Lambda^t \hat{\mathbf{z}}_0 \Leftrightarrow \hat{z}_{ti} = \lambda_i^t \hat{z}_{0i} \ \forall i \in \{1, \ldots, n\}.$$

That is, each component of $\hat{\mathbf{z}}_t$, $\hat{z}_{ti}$, is described by independent equations so that it depends only on $\hat{z}_{0,i}$ (and λ_i).

We are also often interested in the long-run behaviour of a linear dynamic system; i.e., given $\mathbf{z}_t = X\mathbf{z}_{t-1}$, we wish to know

$$\lim_{t\to\infty} \mathbf{z}_t = \lim_{t\to\infty} X^t \mathbf{z}_0 = \lim_{t\to\infty} P\Lambda^t P^{-1}\mathbf{z}_0 = P\left(\lim_{t\to\infty} \Lambda^t\right) P^{-1}\mathbf{z}_0.$$

Since Λ is a diagonal matrix whose diagonal entries are the eigenvalues of X, it follows that the long-run behaviour of the dynamic system depends on the eigenvalues of X, $(\lambda_i)_{i=1}^n$. Since $\Lambda^t = \text{diag}(\lambda_1^t, \ldots, \lambda_n^t)$, for the long-run limit to exist, we must have that $|\lambda_i| < 1$ for any $z_{0i} \neq 0$, where $\mathbf{z}_0 = (z_{0i})_{i=1}^n$.

Example 21. Suppose we have a system of simultaneous first-order difference equations:

$$\begin{aligned} x_{t+1} &= 4x_t + 2y_t, \\ y_{t+1} &= -x_t + y_t, \end{aligned}$$

with initial values x_0 and y_0. We can write this in matrix form, $\mathbf{x}_{t+1} = A\mathbf{x}_t$:

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix}.$$

The eigenvalues of A are given by k that solves $|A - k\mathbb{I}| = 0$; i.e.

$$\begin{aligned} \begin{vmatrix} 4-k & 2 \\ -1 & 1-k \end{vmatrix} &= 0 \\ \Rightarrow (4-k)(1-k) + 2 &= 0 \\ \Rightarrow 6 - 5k + k^2 &= 0 \\ \Rightarrow (k-2)(k-3) &= 0. \end{aligned}$$

Thus, eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = 3$. The eigenvector corresponding to λ_1 is

$$\begin{aligned} \begin{bmatrix} 4-2 & 2 \\ -1 & 1-2 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} &= 0 \\ \begin{bmatrix} 2x_t + 2y_t \\ -x_t - y_t \end{bmatrix} &= 0 \end{aligned}$$

so that the corresponding eigenvector is $(1, -1)$. For λ_2,

$$\begin{bmatrix} 4-3 & 2 \\ -1 & 1-3 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} x_t + 2y_t \\ -x_t - 2y_t \end{bmatrix} = 0$$

so that the corresponding eigenvector is $(2, -1)$. Hence

$$\begin{aligned} P &= \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}, \\ \Lambda &= \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \\ P^{-1} &= \frac{1}{1} \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}. \end{aligned}$$

To verify

$$P\Lambda P^{-1} = \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ -1 & 1 \end{bmatrix} = A.$$

Notice that

$$\mathbf{x}_t = A^t \mathbf{x}_0 = P^{-1} \Lambda^t P \mathbf{x}_0.$$

Define $\mathbf{z}_t = P\mathbf{x}_t$, then

$$\mathbf{z}_t = \begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} = \begin{bmatrix} 2^t & 0 \\ 0 & 3^t \end{bmatrix} \begin{bmatrix} z_{1,0} \\ z_{2,0} \end{bmatrix} = \Lambda^t \mathbf{z}_0$$

That is,

$$z_{1,t} = \lambda_1^t z_{1,0}, \; z_{2,t} = \lambda_2^t z_{2,0}.$$

3.6 Definiteness

A function $Q : \mathbb{R}^n \to \mathbb{R}$ has a quadratic form if it can be represented by

$$\begin{aligned} Q(\mathbf{z}) &= \mathbf{z}^\top X \mathbf{z} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} z_i z_j \\ &= \left(\sum_{i=1}^{n} x_{ii} x_i^2 \right) + \left((x_{12} + x_{21}) z_1 z_2 + \cdots + (x_{n-1,n} + x_{n,n-1}) z_{n-1} z_n \right) \end{aligned}$$

for some $X \in \mathbb{R}^{n \times n}$. There are many ways to represent a quadratic form using a matrix; e.g., if X presents a quadratic form Q, then $X + Y$ also represents Q for any antisymmetric matrix Y (i.e., $y_{ij} = -y_{ji}$). However, each quadratic form can be represented by a unique symmetric matrix because such representation is equivalent to sharing the coefficient on $z_i z_j$ with $i \neq j$ equally between x_{ij} and x_{ji}.

A symmetric matrix $X \in \mathbb{R}^{n \times n}$ is

- *positive definite* if $\mathbf{z}^\top X \mathbf{z} > 0 \; \forall \mathbf{z} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$;
- *negative definite* if $\mathbf{z}^\top X \mathbf{z} < 0 \; \forall \mathbf{z} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$;
- *positive semidefinite* if $\mathbf{z}^\top X \mathbf{z} \geq 0 \; \forall \mathbf{z} \in \mathbb{R}^n$;

- *negative semidefinite* if $\mathbf{z}^\top X\mathbf{z} \leq 0\ \forall \mathbf{z} \in \mathbb{R}^n$;
- *indefinite* if they are none of the above (i.e., if $\exists \mathbf{z}, \mathbf{z}' \in \mathbb{R}^n$, $\mathbf{z}^\top X\mathbf{z} > 0$ and $\mathbf{z}'^\top X\mathbf{z}' < 0$).

There are many ways to characterise the definiteness of a matrix but the following relies on eigenvalues.

Proposition 65. *A symmetric matrix $X \in \mathbb{R}^{n\times n}$ is*

(i) positive definite if and only if all its eigenvalues are strictly positive;

(ii) negative definite if and only if all its eigenvalues are strictly negative;

(iii) positive semidefinite if and only if all its eigenvalues are nonnegative;

(iv) negative definite if and only if all its eigenvalues are nonpositive;

(v) indefinite if and only if it has both positive and negative eigenvalues.

Another useful characterisation is the follows.

Proposition 66 (LDL decomposition)**.** *A symmetric matrix $X \in \mathbb{R}^{n\times n}$ is positive definite if and only if there exists a diagonal matrix $D \in \mathbb{R}^{n\times n}$ with strictly positive entries on its diagonal and a lower triangle matrix $L \in \mathbb{R}^{n\times n}$ with all 1's on its diagonal such that $X = LDL^\top$.*

In the theorem above, if we define $P \coloneqq L\sqrt{D}$ (where $\sqrt{D} \coloneqq \operatorname{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right)$), then we have

$$X = L\sqrt{D}\sqrt{D}L^\top = PP^\top.$$

This is called the Cholesky decomposition.

Theorem 11 (Cholesky decomposition)**.** *A symmetric matrix $X \in \mathbb{R}^{n\times n}$ is positive definite if and only if there exists a lower triangle matrix P with strictly positive entries on its diagonal such that $X = PP^\top$.*

Remark 6. Denote the set of symmetric $n \times n$ matrices $S_n \subseteq \mathbb{R}^{n\times n}$. Define $\geq_{\text{psd}}$ such that

$$X \geq_{\text{psd}} Y \Leftrightarrow X - Y \geq_{\text{psd}} 0 \Leftrightarrow X - Y \text{ is positive semifinite.}$$

Then, $(S_n, \geq_{\text{psd}})$ is a partially order set.

4 Analysis

4.1 Sequences

Remark 7. We will go over the materials in this section in more depth in ECON 6701. We only briefly introduce the relevant notions here necessary to make sense of what is to follow.

A *sequence* in some nonempty set X is a function $x : \mathbb{N} \to X$. Instead of using the standard notation $x(n)$ for functions we use x_n. Some (equivalent) notations for a sequence x are:

$$(x_1, x_2, \ldots) \equiv (x_n)_{n=1}^{\infty} \equiv (x_n)_{n\in\mathbb{N}} \equiv (x_n)_n \equiv (x_n).$$

Let X^{∞} denote the set of all sequences in X. For brevity, we will generally adopt the notation $(x_n)_n$ if no confusion arise. The set of sequences. A *subsequence* of (x_n) is a sequence obtained by (only) deleting elements of (x_n). More formally, a subsequence of (x_n) is any sequence $(x_{n_k})_{k=1}^{\infty} \equiv (x_{n_k})$ where (n_k) is a strictly increasing sequence of non-negative integers.

Remark 8. You might see sequences denoted as $\{x_n\}_{n=1}^{\infty}$. Braces are exclusively for *sets*, which are unordered: $\{2, 3\}$ is the same set as $\{3, 2\}$, which are both the same as $\{2, 3, 2, 2, 2, 3\}$ (with some abuse of notation), etc.

Example 22. Consider the sequence of Real numbers $(1, -1, 1, -1, \ldots) = ((-1)^n)_{n=1}^{\infty}$. (Make sure you understand the notation on the right hand side of the equality.) Its *set of values* is $\{(-1)^n : n \in \mathbb{N}\} = \{1, -1\}$. Seen as a function, $\{1, -1\}$ is the range and $\mathbb{N}$ is the domain (like it is for all sequences). Examples of subsequence of $((-1)^n)_{n=1}^{\infty}$ is $(1, 1, \ldots)$, $(-1, -1, \ldots)$, and $(1, -1, 1, 1, -1, -1, \ldots)$.

4.2 Limits

Let (X, ρ) be a metric space. A sequence (x_n) *converges to* $x \in X$ if, for every $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ such that $n > N$ implies $\rho(x_n, x) < \epsilon$. The point x is called the *limit* of (x_n), and we write

$$\lim_{n\to\infty} x_n = x \text{ or } x_n \to x.$$

Observe that the limit must be in X and we call a sequence (x_n) that converges to a limit a *convergent sequence*. A sequence that is not convergent is a *divergent sequence*. Taking limits preserves order, addition and multiplications etc.

A sequence (x_n) is *Cauchy* if, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\rho(x_m - x_n) < \epsilon$ for all $m, n \in \mathbb{N}$ and $m, n > N$. Thus, elements of Cauchy becomes arbitrarily close.

A subset $S \subseteq X$ is *sequentially compact* if every sequence in S has a convergent subsequence. When (X, ρ) is a metric space, compactness is equivalent to sequential compactness.

Recall that if (X, ρ_X) and (Y, ρ_Y) are metric spaces, then $(X \times Y, \rho)$ is a metric space where ρ is product metric given by (2.3). More generally, if (X_i, ρ_i) is a metric space for each $i \in \{1, \ldots, d\}$, then (X, ρ) is a metric space, where $X \coloneqq \times_{i=1}^{d} X_i$ and

$$\rho(x, y) = \rho((x_1, \ldots, x_d), (y_1, \ldots, y_d)) = \left(\sum (\rho_i(x_i, y_i))^p\right)^{1/p}$$

for any $p \geq 1$. Call such (X, ρ) space an *product metric space*.

Proposition 67. *Let $(X \coloneqq \times_{i=1}^{d} X_i, \rho)$ be a product metric space. A sequence (x_n) in X converges to a limit x if and only if $x_{n,i} \to x_i$ for all $i \in \{1, \ldots, d\}$.*

Note that limits do not always exist; e.g., the sequence $((-1)^n)_n$ does not converge. Nevertheless, we can define alternative notions of limits that always exist. The *limit superior* of a sequence (x_n),

denoted $\limsup_{n\to\infty} x_n$, is defined as

$$\limsup_{n\to\infty} x_n := \lim_{m\to\infty} \sup\{x_n : n \geq m\}.$$

Similarly, the *limit inferior* of (x_n) is defined

$$\liminf_{n\to\infty} x_n := \lim_{m\to\infty} \inf\{x_n : n \geq m\}.$$

When a sequence (x_n) has a limit, then

$$\lim_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = \liminf_{n\to\infty} x_n.$$

We will study these more in ECON 6701.

4.3 Continuity

We now introduce some equivalent notions of continuous functions.

Proposition 68. *Suppose $f : X \to Y$, where (X, ρ_X) and (Y, ρ_Y) are two metric spaces. The following are equivalent.*

(i) f is continuous at $x \in X$.

(ii) (sequential characterisation) $\lim_{n\to\infty} f(x_n) = f(x)$ $\forall x_n \in X^\infty : x_n \to x$.

(iii) (ϵ-δ criterion) $\forall \epsilon > 0$, $\exists \delta_{\epsilon,x} > 0$, $\rho_Y(f(x), f(s)) < \epsilon$ $\forall s \in X\backslash\{x\} : \rho_X(x, s) < \delta_{\epsilon,x}$.[20]

(iv) (topological characterisation) For all open sets $V \subseteq Y$ such that $f(x) \in Y$, there is an open set $U \subseteq X$ such that $x \in U$ and $f(z) \in Y$ for all $z \in U$ (i.e., $f(U) \subseteq Y$).

Say $f : S \to Y$ is *continuous on* $S \subseteq X$ if f is continuous at all $x \in S$. If $S = X$, then we simply say that f is continuous. If f is not continuous at $x \in S$, f is *discontinuous at* x.

The sequential characterisation tells us that continuous functions preserves limits. The topological characterisation gives that a function is continuous if and only if the preimage of every open set is open; i.e., $f^{-1}(V)$ is open for any open subset V in Y.

We say that a function $f : X \to Y$ is *bounded* if $f(X)$ is bounded.

We will prove this in ECON 6701 but you should be aware of it and be familiar with it!

Theorem 12 (Intermediate Value Theorem)**.** *Suppose $f : [a, b] \to \mathbb{R}$ is continuous and $f(a) < 0 < f(b)$. Then, there exists $c \in (a, b)$ such that $f(c) = 0$.*

[20] Sometimes, this is written $\rho(x, s) < \delta_{\epsilon,x} \Rightarrow \rho(y, f(s)) < \epsilon$.

5 Differentiation

5.1 Univariate functions

Suppose X is a linear space. A function $f : X \to \mathbb{R}$ is *univariate* if $\dim X = 1$ and *multivariate* if $\dim X > 1$. A function $f : X \subseteq \mathbb{R} \to \mathbb{R}$ is *differentiable* at $x_0 \in \text{int}(X)$ if the following limit exists:

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

When the limit exists, it is called the *derivative* of f at x_0, and we denote it by $f'(x_0)$ or $Df(x_0)$. Moreover, the function f is *differentiable* on $\text{int}(S) \subseteq X$ if it is differentiable at all $x \in \text{int}(S)$, and f is *differentiable* if it is differentiable on $\text{int}(X)$.

Proposition 69. *Let $f : X \subseteq \mathbb{R} \to \mathbb{R}$ and suppose that f is differentiable at $x \in \text{int}(X)$. Then, f is continuous at x. The converse does not hold.*

Proof. It suffices to show that $\lim_{n\to\infty} f(x_n) - f(x) = 0$ for any $x_n \to x$. To that end, let (x_n) be any sequence that converges to x. Since f is differentiable

$$\begin{aligned}
\lim_{n\to\infty} f(x_n) - f(x) &= \lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x}(x_n - x) \\
&= \lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x} \lim_{n\to\infty} (x_n - x) \\
&= f'(x) \cdot 0 = 0.
\end{aligned}$$

To see that the converse does not hold, take any continuous function with a "kink" point; e.g.,

$$f(x) = \begin{cases} -x & \text{if } x < 0 \\ 2x & \text{if } x \geq 0 \end{cases}.$$

Observe that f is continuous but has a kink at $x = 0$. Then, the left- and right-limits of f at $x$will not coincide; i.e., f is not differentiable $x = 0$. ■

5.1.1 Rules of derivatives

Recognising that derivatives are just limits allows us to obtain the following,

Proposition 70. *Suppose $f, g : X \to \mathbb{R}$ are differentiable. Then,*

(i) $(f + g)' = f' + g'$;

(ii) $(\lambda \cdot f)' = \lambda \cdot f \ \forall \lambda \in \mathbb{R}$;

(iii) (product rule) $(f \cdot g)' = f' \cdot g + f \cdot g'$;

(iv) (quotient rule) $(f/g)' = (f' \cdot g - f \cdot g')/g^2$.

Proposition 71 (Chain rule)**.** *Suppose $f : X \subseteq \mathbb{R} \to \mathbb{R}$ is differentiable at $x \in \text{int}(X)$. Suppose $g : S \to \mathbb{R}$, where $f(I) \subseteq S$ and g is differentiable at $f(x)$. Then, $g \circ f$ is differentiable at x and its derivative is given by*

$$(g \circ f)' = (g' \circ f) f'.$$

Theorem 13 (L'Hôpital's rule)**.** *Let $-\infty \leq a < b \leq +\infty$ and $f : (a, b) \to \mathbb{R}$ and $g : (a, b) \to \mathbb{R}\backslash\{0\}$ are differentiable on (a, b). If $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ are both 0 or $\pm\infty$, and $\lim_{x\to a} f'(x)/g'(x)$*

has a finite value or is $\pm\infty$*, then*

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

The statement holds for $x\to b$*.*

L'Hôpital's rule can be applied many times over when functions are continuously differentiable.

Example 23. Consider the constant relative risk aversion (CRRA) utility function, $u:\mathbb{R}\to\mathbb{R}$,

$$u(x,\gamma)\coloneqq\frac{x^{1-\gamma}-1}{1-\gamma}.$$

Consider limit of $u(x,\gamma)$ as $\gamma\to 1$. Note that $u(x,1)=$ "0/0" and so we can immediately apply LHôpital's rule to obtain

$$\begin{aligned}\lim_{\gamma\to 1}\frac{x^{1-\gamma}-1}{1-\gamma}&=\lim\frac{e^{(1-\gamma)\ln x}-1}{1-\gamma}\\&=\lim\frac{-\ln(x)\,e^{(1-\gamma)\ln x}}{-1}=\ln x.\end{aligned}$$

That is, CRRA utility is a generalisation of log utility.

Proposition 72 (Derivative of inverse)**.** *Suppose* $f:(a,b)\subset\mathbb{R}\to\mathbb{R}$ *and* f *is strictly increasing and differentiable on* (a,b)*. Then,*

$$\left(f^{-1}\right)'(f(x))=\frac{1}{f'(x)}\ \forall x\in(a,b).$$

A heuristic "proof" is to differentiate the following identity while using chain rule:

$$\begin{aligned}f^{-1}(f(x))&\equiv x\\\Rightarrow\left(f^{-1}\right)'(f(x))\,f'(x)&=1\\\Rightarrow\left(f^{-1}\right)'(f(x))&=\frac{1}{f'(x)}.\end{aligned}$$

Theorem 14 (Mean Value Theorem)**.** *Suppose* $f:[a,b]\subseteq\mathbb{R}\to\mathbb{R}$ *is continuous and differentiable on* (a,b)*. Then, there exists* $x\in(a,b)$ *such that*

$$f(b)-f(a)=f'(x)(b-a).$$

Let $b=a+h$ then Mean Value Theorem tells us that

$$\exists x\in(a,a+h),\ f(a+h)=f(a)+f'(x)\,h.$$

Thus, it gives us a way of approximating f around a via an affine function $f(a)+f'(x)h$.

Proposition 73. *Suppose that* $f:I\to\mathbb{R}$ *is differentiable on* $\operatorname{int}(I)$*, where* I *is a closed and bounded interval in* $\mathbb{R}$*. Suppose also that* f' *is bounded. Then,* f *is Lipschitz continuous; i.e., there exists* $M>0$ *such that*

$$|f(x)-f(y)|\le M|x-y|\ \ \forall x,y\in\operatorname{int}(I).$$

Proof. Let $M\ge 0$ be a bound on f' so that $|f'(x)|<M$ for all $x\in\operatorname{int}(I)$. Since f is differentiable on $\operatorname{int}(I)$, f is continuous on $\operatorname{int}(I)$ by Proposition 69. Then, by the Mean Value Theorem (Theorem

14), for any $a, b \in \text{int}(I)$ with $b > a$, there exists $c \in (a, b)$

$$\begin{aligned} f(a) - f(b) &= f'(c)(b-a) \\ \Rightarrow |f(a) - f(b)| &= |f'(c)||b-a| \leq M|b-a|. \end{aligned}$$

Since above holds for any $a, b \in \text{int}(I)$, f is Lipschitz continuous on $\text{int}(I)$. ■

5.1.2 Taylor expansion

Suppose f has a derivative f' on an interval and that f' is itself differentiable. Then, the derivative of f' is denoted f'', $f^{(2)}$ or $D^2 f$. Suppose f is a real-valued, continuous function defined on interval $(a, b) \subseteq \mathbb{R}$). f is *continuously differentiable* if its derivative is a continuous function on (a, b). A function f is *twice continuously differentiable* if f' is continuously differentiable. More generally, for any $k \in \mathbb{N}$, a function f is *k-times continuously differentiable* if (i) f has up to and including kth derivative; and (ii) f and all its derivatives up to and including the kth derivative are all continuous on (a, b).

Theorem 15 (Taylor's Theorem in $\mathbb{R}$)**.** *Suppose f is n-times continuously differentible. Let α and β be distinct point in $[a, b]$ and define*

$$P_{n-1}(t) \coloneqq f(\alpha) + f'(\alpha)(t-\alpha) + \frac{1}{2}f''(\alpha)(t-\alpha)^2 + \cdots + \frac{f^{(n-1)}(\alpha)}{(n-1)!}(t-\alpha)^{n-1}.$$

Then, there exists $x \in (\alpha, \beta)$ such that

$$f(\beta) = P_{n-1}(\beta) + \frac{f^{(n)}(x)}{n!}(\beta-\alpha)^n. \tag{5.1}$$

If $n = 1$, then this says that $P_0(t) = f(\alpha)$ and so $f(\beta) = f(\alpha) + f'(x)(\beta - \alpha)$ for some $x \in (\alpha, \beta)$, which is just the Mean Value Theorem. Thus, we can think of Taylor's Theorem as an nth-order generalisation of the Mean Value Theorem. Recalling that the Mean Value Theorem can be thought of an affine approximation of f, (5.1) can be thought of an $n-1$-degree polynomial approximation of $f(\beta)$ (note that $P_{n-1}(t)$ is a polynomial of degree $n-1$), where the term $f^n(x)(\beta-\alpha)^n/n!$ can be thought of the error term.

Example 24. *[Maclaurin series]*An infinite Taylor expansion of functions around 0 are called Maclaurin series. Here are some useful ones to know:

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{5.2}$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}x^n}{n} \ \forall x \in (-1, 1]$$

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots = \sum_{n=1}^{\infty} -\frac{x^n}{n} \ \forall x \in (-1, 1].$$

5.2 Multivariate function

Let us now generalise what we found above to multivariate functions $f : \mathbb{R}^n \to \mathbb{R}$. We say that a function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ is *differentiable* at $\mathbf{x} \in \text{int}(X)$ if there exists an $1 \times n$ vector $\nabla f(\mathbf{x})$ such that

$$\lim_{\mathbf{h} \in \mathbb{R}^n : \mathbf{h} \to \mathbf{0}} \frac{|f(\mathbf{x}+\mathbf{h}) - (f(\mathbf{x}) + \nabla f(\mathbf{x})\mathbf{h})|}{\|\mathbf{h}\|_{\mathbb{R}^n}} = 0.$$

Given $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$, we define the *ith partial derivative* of f at $\mathbf{x} \in \text{int}(X)$ as

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) \coloneqq \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \ \forall i \in \{1, \ldots, n\},$$

where $\mathbf{e}_i \in \mathbb{R}^n$ is the ith canonical basis of $\mathbb{R}^n$. (It should be clear from this definition that $\partial f(\mathbf{x})/\partial x_i$ is *not* a fraction!) Observe that ith partial derivative of f considers how the value of f changes when $\mathbf{x}$ moves in the direction of the ith coordinate. $(\partial f/\partial x_i)(\mathbf{x})$ is equivalent to computing the derivative of the univariate function, $\tilde{f}_i(h) \coloneqq f(\mathbf{x} + h\mathbf{e}_i)$, and evaluating this at $h = 0$; i.e.,

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}h}\tilde{f}_i(0).$$

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x} \in \text{int}(X)$, then

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}_{1 \times n}.$$

The row vector of partial derivatives on the right-hand side is called *Jacobian matrix* at $\mathbf{x}$. Hence, another way to state the result is that when f is differentiable, then its total derivative is given by the Jacobian matrix.

We can define higher-order derivatives in the similar way to the univariate case. The second derivative at $\mathbf{x} \in \text{int}(X)$ of a real-valued function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ is called the *Hessian matrix* of f at $\mathbf{x}$:

$$H_f(\mathbf{x}) \coloneqq \begin{bmatrix} \left(\nabla \frac{\partial f}{\partial x_1}\right)(\mathbf{x}) \\ \vdots \\ \left(\nabla \frac{\partial f}{\partial x_n}\right)(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial\left(\frac{\partial f}{\partial x_1}\right)}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial\left(\frac{\partial f}{\partial x_1}\right)}{\partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial\left(\frac{\partial f}{\partial x_n}\right)}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial\left(\frac{\partial f}{\partial x_n}\right)}{\partial x_n}(\mathbf{x}) \end{bmatrix}_{n \times n},$$

where $(\nabla \frac{\partial f}{\partial x_i})(\mathbf{x})$ is the gradient of the function $\frac{\partial f}{\partial x_i}$ at $\mathbf{x}$ and $\frac{\partial(\frac{\partial f}{\partial x_i})}{\partial x_j}(\mathbf{x})$ is the jth partial derivative of the function $\frac{\partial f}{\partial x_i}$ at $\mathbf{x}$. We refer to the latter as a *cross partial* at $\mathbf{x}$ and write

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}) \coloneqq \frac{\partial\left(\frac{\partial f}{\partial x_i}\right)}{\partial x_j}(\mathbf{x}).$$

The following relates $\frac{\partial^2 f}{\partial x_j \partial x_i}$ with $\frac{\partial^2 f}{\partial x_i \partial x_j}$ when f is twice differentiable, which tells us that when f is twice-differentiable at $\mathbf{x}$, then the Hessian matrix at $\mathbf{x}$ is symmetric.

Theorem 16 (Young's Theorem)**.** *Suppose $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable at $\mathbf{x} \in \text{int}(X)$, then whenever the cross partials exist,*

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

5.3 Implicit function theorem

Suppose $\mathbf{x} \in \mathbb{R}^n$ and $(\mathbf{x}^*, y^*)$ is a solution to $F(\mathbf{x}, y) = 0$, where F is continuously differentiable in an open ball around $(\mathbf{x}^*, y^*)$ with $\frac{\partial F}{\partial y}(\mathbf{x}^*, y^*) \neq 0$. Then, the implicit function theorem tells us that there is a continuously differentiable function g defined on an open ball around $\mathbf{x}^*$ such that

$g(\mathbf{x}^*) = y^*$, $F(\mathbf{x}, g(\mathbf{x})) = 0$, and

$$\frac{\partial y}{\partial x_i}(\mathbf{x}, y) = -\frac{\frac{\partial F}{\partial x_i}(\mathbf{x}, y)}{\frac{\partial F}{\partial y}(\mathbf{x}, y)}.$$

Example 25. Suppose $F(x, y) \coloneqq 3x^2 - 2y$ and we implicit define y as a function of x via $F(x, y) = 0$. Then

$$\frac{\partial y(x)}{\partial x} = -\frac{6x}{(-2)} = 3x.$$

Example 26. Suppose $u : \mathbb{R}^2 \to \mathbb{R}$ is a utility function that is "well behaved.". The slope of the indifference curve $u(x, y) = c$ for some constant c is given by implicitly differentiating the equation:

$$\frac{\partial y}{\partial x} = -\frac{\frac{\partial u(x,y)}{\partial x}}{\frac{\partial u(x,y)}{\partial y}}.$$

5.4 Inverse function theorem

Note that an inverse function of $f : X \subseteq \mathbb{R} \to Y \subseteq \mathbb{R}$, f^{-1}, satisfies following equation:

$$y - f\left(f^{-1}(y)\right) \equiv 0.$$

Thus, we can think of $x = f^{-1}(0)$ as being implicitly defined via the expression above. Inverse function theorem tells us that, if f is continuously differentiable with nonzero derivative at an interior point x_0, then the derivative of the inverse function at $y = f(x_0)$ is the reciprocal of the derivative of f at x_0; i.e.,

$$\left(f^{-1}\right)'(y) = \frac{1}{f'(x_0)} = \frac{1}{f'\left(f^{-1}(y)\right)}.$$

5.5 Concavity and convexity

If a function is twice continuously differentiable (so that the Hessian f is symmetric), then concavity and convexity of functions can be characterised via the Hessian matrix (i.e., the second derivative).

Proposition 74. *Let $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable on $\operatorname{int}(X)$ and suppose $\operatorname{int}(X)$ is a convex set.*

(i) f is concave on $\operatorname{int}(X)$ if and only if $H_f(\mathbf{x})$ is negative semidefinite for all $\mathbf{x} \in \operatorname{int}(X)$.

(ii) f is convex on $\operatorname{int}(X)$ if and only if $H_f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in \operatorname{int}(X)$.

(iii) If $H_f(\mathbf{x})$ is negative definite for all $\mathbf{x} \in \operatorname{int}(X)$, then f is strictly concave on $\operatorname{int}(X)$.

(iv) If $H_f(\mathbf{x})$ is positive definite for all $\mathbf{x} \in \operatorname{int}(X)$, then f is strictly convex on $\operatorname{int}(X)$.

When $d = 1$, H_f is effectively a number and so the proposition says that $f : X \subseteq \mathbb{R} \to \mathbb{R}$ is concave (resp. convex) if and only if f'' is negative (resp. positive) for all $x \in \operatorname{int}(X)$.

Exercise 18. Consider $f : \mathbb{R} \to \mathbb{R}$ defined via $f(x) = -x^4$. Show that f is strictly concave and compute $f''(0)$. What does this tell you about condition (iii)?

6 Riemann integration

The idea behind the 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.

6.1 Construction

6.1.1 Upper and lower Riemann integrals

For any $n \in \mathbb{N}$, define the *dissection* of $[a,b]$ (with $a<b$) by

$$[a_0,\ldots,a_n] \coloneqq \{[a_0,a_1],[a_1,a_2],\ldots,[a_{m-1},a_m]\},$$

where $a = a_0 < \cdots < a_m = b$. Let $\mathcal{D}[a,b]$ denote the class of all dissections of $[a,b]$. Define $\mathcal{D}[a,b] \coloneqq \{\{a\}\}$ if $a=b$.

For any $\mathbf{a} \coloneqq [a_0,\ldots,a_n]$ and $\mathbf{b} \coloneqq [b_0,\ldots,b_m]$ in $\mathcal{D}[a,b]$, we write $\mathbf{a} \Cup \mathbf{b}$ for the dissection $[c_0,\ldots,c_\ell] \in \mathcal{D}[a,b]$, where

$$\{c_0,\ldots,c_\ell\} = \{a_0,\ldots,a_n\} \cup \{b_0,\ldots,b_m\}.$$

We say that $\mathbf{b}$ is finer than $\mathbf{a}$ if $\{a_0,\ldots,a_m\} \subseteq \{b_0,\ldots,b_m\}$. Note that $\mathbf{a} \Cup \mathbf{b} = \mathbf{b}$ if and only if $\mathbf{b}$ is finer than $\mathbf{a}$.

Let $f:[a,b]\to\mathbb{R}$ be any bounded function. For any $\mathbf{a} \coloneqq [a_0,\ldots,a_n] \in \mathcal{D}[a,b]$, define, for each $i \in \{1,\ldots,n\}$,

$$K_{f,\mathbf{a}}(i) \coloneqq \sup\{f(x) : a_{i-1} \le x \le a_i\},$$
$$k_{f,\mathbf{a}}(i) \coloneqq \inf\{f(x) : a_{i-1} \le x \le a_i\}.$$

By the $\mathbf{a}$-*upper Riemann sum of* f, we mean the number

$$R_{\mathbf{a}}(f) \coloneqq \sum_{i=1}^{n} K_{f,\mathbf{a}}(i)(a_i - a_{i-1})$$

and by the $\mathbf{a}$-*lower Riemann sum of* f, we mean the number

$$r_{\mathbf{a}}(f) \coloneqq \sum_{i=1}^{n} k_{f,\mathbf{a}}(i)(a_i - a_{i-1}).$$

We define the *upper Riemann integral of* f by

$$R(f) \coloneqq \inf\{R_{\mathbf{a}}(f) : \mathbf{a} \in \mathcal{D}[a,b]\}$$

and the *lower Riemann integral of* f by

$$r(f) \coloneqq \sup\{r_{\mathbf{a}}(f) : \mathbf{a} \in \mathcal{D}[a,b]\}.$$

Proposition 75. *$R_{\mathbf{a}}(f)$ decreases and $r_{\mathbf{a}}(f)$ increases as $\mathbf{a}$ becomes finer.*

The following tells us that all lower sums are less than all upper sums, not just the lower and upper sums associated with the same partition.

Proposition 76. *$R(f) \ge r(f)$.*

6.1.2 Riemann integrability and Riemann integral

A bounded function $f : [a, b] \to \mathbb{R}$ is *Riemann integrable* if $R(f) = r(f)$ and the number

$$\int_a^b f \equiv \int_a^b f(x)\,\mathrm{d}x \coloneqq R(f)$$

is called the *Riemann integral of* f.[21] In this case, we also define

$$\int_b^a f \equiv \int_b^a f(x)\,\mathrm{d}x \coloneqq -R(f).$$

If $g : [a, \infty) \to \mathbb{R}$ is a bounded function, we define the *improper Riemann integral of* g as

$$\int_a^\infty g \equiv \int_a^\infty g(x)\,\mathrm{d}x \coloneqq \lim_{b\to\infty} R\left(g|_{[a,b]}\right)$$

provided that $g|_{[a,b]}$ is Riemann integrable for each $b > a$ and the limit exists (in $\overline{\mathbb{R}}$). Analogously, if $g : (-\infty, b] \to \mathbb{R}$ is a bounded function, we define

$$\int_{-\infty}^b g \equiv \int_{-\infty}^b g(x)\,\mathrm{d}x \coloneqq \lim_{a\to-\infty} R\left(g|_{[a,b]}\right)$$

provided that $g|_{[a,b]}$ is Riemann integrable for each $a < b$ and the limit exists (in $\overline{\mathbb{R}}$).

Remark 9 (Riemann integrals over hyperrectangles). A *hyperrectangle* is Cartesian product of intervals. One can also define what it means to integrate functions such as $f : [0, 1]^2 \to \mathbb{R}$ over the hyperrectangle $[0, 1]^2$. For any hyperrectangle $C \coloneqq [a^1, b^1] \times \cdots \times [a^n, b^n] \subseteq \mathbb{R}^n$, define the *d-dimensional volume* of R by

$$V(C) \coloneqq \prod_{i=1}^d \left(b^i - a^i\right).$$

Define the set of all dissections of hyperrectangle R by

$$\mathcal{D}\left(\left[a^i, b^i\right]\right)_{i=1}^d \coloneqq \times_{i=1}^n \mathcal{D}\left[a^i, b^i\right].$$

Then, $A = \{A_1, \ldots, A_N\} \in \mathcal{D}([a^i, b^i])_{i=1}^d$, where the ordering is arbitrary. To simply, consider the case in which $d = 2$. Let $\mathbf{a}^i = [a_0^i, \ldots, a_{n_i}^i] \in \mathcal{D}[a^i, b^i]$ for each $i \in \{1, 2\}$. If $n_1 = 2$ and $n_2 = 2$, we have

$$\begin{aligned}
\mathbf{a}^i &= \left\{\left[a_0^i, a_1^i\right], \left[a_1^i, a_2^i\right]\right\} = \left\{\left[a^i, a_1^i\right], \left[a_1^i, b^i\right]\right\} \\
\Rightarrow A &= \mathbf{a}^1 \times \mathbf{a}^2 \\
&= \{C_1, C_2, C_3, C_4\} \\
&= \left\{\left[a^1, a_1^1\right] \times \left[a^2, a_1^2\right], \left[a^1, a_1^1\right] \times \left[a_1^2, b^2\right], \left[a_1^1, b^1\right] \times \left[a_1^2, b^2\right], \left[a_1^1, b^1\right] \times \left[a_1^2, b^2\right]\right\}.
\end{aligned}$$

As before, we define

$$R_A(f) \coloneqq \sum_{i=1}^N K_{f,A}(i)\, V(A_i), \;\; R_A(f) \coloneqq \sum_{i=1}^N k_{f,A}(i)\, V(A_i),$$

where

$$K_{f,A}(i) \coloneqq \sup\{f(x) : x \in A_i\}, \;\; k_{f,A}(i) \coloneqq \inf\{f(x) : x \in A_i\}$$

[21] A function $f : X \to Y$ is bounded if $f(X)$ is bounded.

and

$$R(f) \coloneqq \inf\left\{R_A(f) : A \in \mathcal{D}\left(\left[a^i, b^i\right]\right)_{i=1}^d\right\},$$
$$r(f) \coloneqq \sup\left\{r_A(f) : A \in \mathcal{D}\left(\left[a^i, b^i\right]\right)_{i=1}^d\right\}.$$

Then a bounded function $f : \times_{i=1}^d [a^i, b^i] \to \mathbb{R}$ is Riemann integrable if $R(f) = r(f)$ and

$$\int_{\times_{i=1}^d [a^i, b^i]} f \equiv \int_{\mathbf{z} \in \times_{i=1}^d [a^i, b^i]} f(\mathbf{z})\, \mathrm{d}\mathbf{z} \coloneqq R(f).$$

Recalling the definition of R and r, observe that if f is unbounded above, then $R(f) = \infty$ and if f is unbounded below, then $r(f) = -\infty$. Hence, unbounded functions are not integrable. The following is simply the contrapositive of these observations.

6.1.3 Riemann integrable functions

Example 27 (Constant function)**.** Define $f : [0,1] \to \mathbb{R}$ by $f(x) \coloneqq 1$. Let us show that

$$\int_0^1 1 \mathrm{d}x = 1.$$

Since f is a constant function,

$$K_{f,\mathbf{a}}(i) = 1 = k_{f,\mathbf{a}}(i) \;\; \forall \mathbf{a} \in \mathcal{D}[0,1].$$

Therefore,

$$R_{\mathbf{a}}(f) = 1 = r_{\mathbf{a}}(f) \;\; \forall \mathbf{a} \in \mathcal{D}[0,1],$$

which, in turn, implies that $R(f) = r(f) = 1$. Similar argument gives us that for any $f : [a,b] \to \mathbb{R}$ defined as $f(x) \coloneqq c$, we have

$$\int_a^b c \mathrm{d}x = c(b-a).$$

Example 28 (Integrable discontinuous function)**.** Define $f : [0,1] \to \mathbb{R}$ by $f(x) \coloneqq 1 - \mathbb{1}_{\{x=0\}}$. We will show that

$$\int_0^1 \left(1 - \mathbb{1}_{\{x=0\}}\right) = 0.$$

For any $\mathbf{a} = [0, a_1, \ldots, a_{n-1}, 1] \in \mathcal{D}[a,b]$, the first partition contains the end of zero, and so

$$K_{f,\mathbf{a}}(1) = 1, \;\; k_{f,\mathbf{a}}(1) = 0,$$
$$K_{f,\mathbf{a}}(i) = 0, \;\; k_{f,\mathbf{a}}(i) = 0 \;\forall i \in \{2, \ldots, n\}.$$

Thus,

$$R_{\mathbf{a}}(f) = a_1, \;\; r_{\mathbf{a}}(f) = 0 \;\forall \mathbf{a} \in \mathcal{D}[a,b].$$

and, in turn,

$$R(f) \equiv \inf\{R_{\mathbf{a}}(f) : a \in \mathcal{D}[a,b]\} = 0 = \sup\{r_{\mathbf{a}}(f) : a \in \mathcal{D}[a,b]\} \equiv r(f).$$

Hence, $f(x)$ is Riemann integrable, and its Riemann integral is zero. Similar argument tells us that the Riemann integrable of a function that is zero except at finitely many points in $[a,b]$ is Riemann integrable with Riemann integrable of zero.

Let us give an example of a bounded function that is not Riemann integrable.

Example 29 (Dirichlet's function)**.** Let $f : [0,1] \to [0,1]$ be defined as

$$f(x) \coloneqq \mathbb{1}_{\{x \in \mathbb{Q}\}} = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \notin \mathbb{Q} \end{cases}.$$

This is a function that is discontinuous everywhere (recall Proposition 2). Let $\mathbf{a} \in \mathcal{D}[0,1]$. Note that $\mathbb{Q}$ is dense in $\mathbb{R}$; i.e. between any two real numbers, there exists a rational number, which implies that in any interval, there exists some rational $q \in \mathbb{Q}$. Hence,

$$r_{\mathbf{a}}(f) = 0 \neq 1 = R_{\mathbf{a}}(f).$$

Thus, f is not Riemann integrable.

Changing the values of a function at finitely many points does not change its integrability of the value of its integral. Consequently, the value of the function at the end points does not affect the integral and, as such, integration on (a,b), $[a,b]$, $[a,b)$, $(a,b]$ all result in the same value of the integral.

Proposition 77. *Suppose that $f, g : [a,b] \to \mathbb{R}$ are bounded and $f(x) = g(x)$ except at finitely many points $x \in [a,b]$. Then, f is Riemann integrable if and only if g is integrable, and*

$$\int_a^b f = \int_a^b g.$$

If f and g differ at countably infinite number of points, it is not the case that f is Riemann integrable if and only if g is Riemann integrable.

Example 30. Consider $f : [0,1] \to \mathbb{R}$ defined as

$$f(x) \coloneqq \begin{cases} n & \text{if } x = \frac{1}{n} \text{ for some } n \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}.$$

Then, f is equal to the zero function except on countably infinite set $\{1/n : n \in \mathbb{N}\}$. However, f is unbounded and therefore not Riemann integrable. Another example is the Dirichlet function, which differs from the zero function at countably infinite set of rationals but is not Riemann integrable.

The example above notwithstanding, even if f and g differ at countably infinitely many points but are both Riemann integrable, then the value of their integrals are equal.

Proposition 78. *Any bounded continuous function on $[a,b]$ is Riemann integrable.*

6.2 Properties of Riemann integration

Let us collect the properties of Riemann integration below.

Proposition 79. *Let $\alpha \in \mathbb{R}$ and $f, g : [a,b] \to \mathbb{R}$ be bounded functions that are Riemann integrable.*

(i) (Linearity) $\alpha f + g$ is Riemann integrable and

$$\int_a^b (\alpha f + g) = \alpha \int_a^b f + \int_a^b g.$$

(ii) (Multiplication) $f \cdot g$ is Riemann integrable.

(iii) (Division) if $g \neq 0$ and $1/g$ is bounded, then f/g is Riemann integrable.

(iv) (Monotonicity) if $f \geq g$, then

$$\int_a^b f \geq \int_a^b g.$$

(v) (Additivity) if $a < c < b$, then

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

(vi) (Equivalence) If $f = g$ in all but countably many points, then

$$\int_a^b f = \int_a^b g.$$

(vii) (Converse of equivalence) If $f \geq 0$, then

$$\int_a^b f = 0 \Rightarrow f = 0.$$

(viii) (Absolute value)

$$\left| \int_a^b f \right| \leq \int_a^b |f| .$$

Remark 10. Notice how linearity, monotonicity and additivity are analogous to the corresponding properties of sums:

$$\sum_{i=1}^n (\alpha x_i + y_i) = \alpha \sum_{i=1}^n x_i + \sum_{i=1}^n y_i,$$
$$x_i \geq y_i \ \forall i \in \{1, \ldots, n\} \Rightarrow \sum_{i=1}^n x_i \geq \sum_{i=1}^n y_i,$$
$$\sum_{i=1}^n x_i = \sum_{i=1}^m x_i + \sum_{i=m+1}^n x_i.$$

The following is a result that you will use quite a lot in Empirical Analysis I—it allows you to split integral of f into positive and negative regions.

Proposition 80. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function and define $f^+, f^- : [a, b] \to \mathbb{R}$ by*

$$f^+(x) \coloneqq \max\{f(x), 0\}, \ f^-(t) \coloneqq \max\{-f(x), 0\}.$$

Then,

$$f = f^+ - f^- \text{ and } |f| = f^+ + f^-.$$

Proof. Fix any $x \in [a, b]$. There are three cases:

(i) $f(x) > 0$. Then, $f^+(x) = f(x)$ and $f^-(x) = 0$. Thus, $(f^+ - f^-)(x) = f(x)$ and $(f^+ + f^-)(x) = f(x) = |f|(x)$.

(ii) $f(x) = 0$. Then, $f^+(x) = f(x) = f^-(x) = 0$. Thus, $(f^+ - f^-)(x) = 0 = f(x)$ and $(f^+ + f^-)(x) = 0 = |f|(x)$.

(iii) $f(x) < 0$. Then, $f^+(x) = 0$ and $f^-(x) = -f(x) > 0$. Thus, $(f^+ - f^-)(x) = f(x)$ and $(f^+ + f^-)(x) = -f(x) = |f|(x)$.

It follows then that $f = f^+ - f^-$ and $|f| = f^+ + f^-$ and the difference and the sums are well defined since f is bounded. ■

Lemma 4. *Suppose $f, g : [a, b] \to \mathbb{R}$ are Riemann integrable. Then, $\max\{f(x), g(x)\}$ is Riemann integrable.*

Proof. Define $h(x) \coloneqq \max\{f(x), g(x)\}$. Take any $\mathbf{a} = [a_0, \ldots, a_n] \in \mathcal{D}[a, b]$. Observe that

$$K_{h,\mathbf{a}}(i) = \max\{K_{f,\mathbf{a}}(i), K_{g,\mathbf{a}}(i)\}, \; k_{h,\mathbf{a}}(i) = \max\{k_{f,\mathbf{a}}(i), k_{g,\mathbf{a}}(i)\}.$$

Hence,

$$\begin{aligned} K_{h,\mathbf{a}}(i) - k_{h,\mathbf{a}}(i) &= \max\{K_{f,\mathbf{a}}(i), K_{g,\mathbf{a}}(i)\} - \max\{k_{f,\mathbf{a}}(i), k_{g,\mathbf{a}}(i)\} \\ &\leq \max\{K_{f,\mathbf{a}}(i) - k_{f,\mathbf{a}}(i), K_{g,\mathbf{a}}(i) - k_{g,\mathbf{a}}(i)\} \\ &\leq K_{f,\mathbf{a}}(i) - k_{f,\mathbf{a}}(i) + K_{g,\mathbf{a}}(i) - k_{g,\mathbf{a}}(i). \end{aligned}$$

It follows then that

$$R_{\mathbf{a}}(h) - r_{\mathbf{a}}(h) \leq R_{\mathbf{a}}(f) - r_{\mathbf{a}}(f) + R_{\mathbf{a}}(g) - r_{\mathbf{a}}(g).$$

Since f and g are both Riemann integrable, by the Cauchy criterion for Riemann integrability, for any $\epsilon/2 > 0$, there exists $\mathbf{a}, \mathbf{b} \in \mathcal{D}[a, b]$ such that

$$R_{\mathbf{a}}(f) - r_{\mathbf{a}}(f) < \frac{\epsilon}{2} \text{ and } a_{\mathbf{b}}(g) - r_{\mathbf{b}}(g) < \frac{\epsilon}{2}.$$

Therefore, by Proposition 75,

$$R_{\mathbf{a}\cup\mathbf{b}}(h) - r_{\mathbf{a}\cup\mathbf{b}}(h) \leq R_{\mathbf{a}\cup\mathbf{b}}(f) - r_{\mathbf{a}\cup\mathbf{b}}(f) + R_{\mathbf{a}\cup\mathbf{b}}(g) - r_{\mathbf{a}\cup\mathbf{b}}(g) < \epsilon.$$

Thus, h satisfies the Cauchy criterion for Riemann integrability; i.e., h is Riemann integrable. ■

Proposition 81. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then so are f^+, f^- and $|f|$.*

Proof. If f is Riemann integrable, then $-f$ is Riemann integrable. Since the zero function is Riemann integrable, the Riemann integrability of f^+ and f^- follows from the previous lemma. Finally, to prove that $|f|$ is Riemann integrable, set $g \coloneqq -f$ in the previous lemma. ■

The following tells us gives us a bound on the Riemann integral of a Riemann integrable function.

Proposition 82. *If $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then*

$$\left| \int_a^b f \right| \leq (b - a) \sup\{|f(t)| : a \leq t \leq b\}.$$

Proof. Since f is bounded, $\sup|f(t) : a \leq t \leq b|$ is finite. By definition of $R(f)$ and the fact that $\{[a, b]\} \in \mathcal{D}[a, b]$,

$$\begin{aligned} \int_a^b f = R(f) &= \inf\{R_{\mathbf{a}}(f) : \mathbf{a} \in \mathcal{D}[a, b]\} \\ &\leq R_{\{[a,b]\}}(f) = K_{f,\{[a,b]\}}(b - a) = \sup\{f(t) : a \leq t \leq b\} \\ &\leq \sup\{|f(t)| : a \leq t \leq b\}. \end{aligned}$$

Similarly,

$$\begin{aligned}\int_a^b f = R(f) = r(f) &= \sup\{r_{\mathbf{a}}(f) : \mathbf{a} \in \mathcal{D}[a,b]\} \\ &\geq r_{\{[a,b]\}}(f) = k_{f,\{[a,b]\}}(b-a) = \inf\{f(t) : a \leq t \leq b\}(b-a) \\ \Rightarrow -\int_a^b f &\leq -\inf\{f(t) : a \leq t \leq b\}(b-a) \\ &= \sup\{-f(t) : a \leq t \leq b\}(b-a) \\ &\leq \sup\{|f(t)| : a \leq t \leq b\}(b-a).\end{aligned}$$

■

6.3 Approximation by step functions

Suppose $f : [a,b] \to \mathbb{R}$ is Riemann integrable. We will show that f can be approximate by a collection of *step functions*. Formally, a function $\varphi : [a,b] \to \mathbb{R}$ is a *step function* if there exists $\mathbf{a} = [a_0, \ldots, a_n] \in \mathcal{D}[a,b]$ of $[a,b]$ and values $\{c_i\}_{i=1}^{n+1} \subseteq \mathbb{R}$ such that

$$\varphi(x) = \sum_{i=1}^{n} c_i \mathbb{1}_{\{x \in (x_{i-1}, x_i)\}},$$

where we ignore the equality at the boundaries of the intervals.

Proposition 83. *Step functions are Riemann integrable.*

Proof. Let $\varphi : [a,b] \to \mathbb{R}$ be a step function and denote

Given a step function $\varphi(x)$, observe that

$$\begin{aligned}K_{\varphi,\mathbf{a}}(i) &= c_i \sup\left\{\mathbb{1}_{\{x \in (x_{i-1}, x_i]\}} : a_{i-1} \leq a \leq a_i\right\} = c_i, \\ k_{\varphi,\mathbf{a}}(i) &= c_i \inf\left\{\mathbb{1}_{\{x \in (x_{i-1}, x_i]\}} : a_{i-1} \leq a \leq a_i\right\} = c_i.\end{aligned}$$

Therefore,

$$r_{\varphi,\mathbf{a}} = \sum_{i=0}^{n} c_i (x_i - x_{i-1}) = R_{\varphi,\mathbf{a}}.$$

Since this holds for any φ and corresponding $\mathbf{a}$, it follows that φ is Riemann integrable, and its Riemann integral is given by

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

■

Proposition 84. *Let $f : [a,b] \to \mathbb{R}$ be a Riemann integrable function. Then, there exists a sequence of step functions (φ_n) such that*

$$\lim_{n\to\infty} \int_a^b \varphi_n = \int_a^b f.$$

Thus, this gives us a way of construction the Riemann integral. As we will see later, we will construct Lebesgue integral through similar approximations.

6.4 Fundamental theorem of calculus

The fundamental theorem of calculus connects derivative of a function with its integral (or *antiderivative*). In particular, it allows us to think about integration as the reverse operation of taking a derivative and gives us a practical way of computing integral of functions.

Theorem 17 (Fundamental theorem of (Riemann integral) calculus)**.**

(i) If $F[a,b] \to \mathbb{R}$ is continuous on $[a,b]$ and differentiable in (a,b) with $F' = f$, where $f : [a,b] \to \mathbb{R}$ is Riemann integrable, then

$$\int_a^b f(x)\,\mathrm{d}x \equiv \int_a^b F'(x)\,\mathrm{d}x = F(b) - F(a).$$

(ii) If $f : [a,b] \to \mathbb{R}$ is Riemann integrable and $F : [a,b] \to \mathbb{R}$ is defined by

$$F(x) \coloneqq \int_a^x f(t)\,\mathrm{d}t,$$

then F is continuous on $[a,b]$. Moreover, if f is continuous at $c \in [a,b]$, then F is differentiable at c and

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\,\mathrm{d}t\bigg|_{x=c} \equiv F'(c) = f(c).$$

.

Remark 11 (Discrete analogs). (i) is a continuous analog of the telescoping sum:

$$A_n - A_0 = \sum_{i=1}^{n} (A_i - A_{i-1}).$$

(ii) is a continuous analog of the differences of sums:

$$a_n = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n-1} a_i.$$

Here is another formulation with slightly stronger assumptions.

Corollary 9. *For any $f \in \mathbf{C}[a,b]$ and $F : [a,b] \to \mathbb{R}$,*

$$F(x) = F(a) + \int_a^x f(t)\,\mathrm{d}t \ \forall x \in \mathbb{R} : a \le x \le b$$

if and only if $F \in \mathbf{C}^1[a,b]$ and $F' = f$.

6.4.1 Existence of antiderivative

One consequence of the fundamental theorem of calculus is that if $f : [a,b] \to \mathbb{R}$ is continuous, then f has an antiderivative, even if it cannot be expressed explicitly (i.e., without integrals). One example is the cumulative distribution function of the normal distribution (we will see more on this later). Here, we introduce another example called the *gamma function* which was discovered by Euler as an answer to the question: is there a "smooth" function on $(0,\infty)$ that maps each $n \in \mathbb{N}$ to the value $n!$. The gamma function is defined by

$$\Gamma(x) \coloneqq \int_0^\infty e^{-t} t^{x-1} \mathrm{d}t, \ x > 0.$$

Try showing that Γ is well defined, $\Gamma(1) = 1$ and that $\Gamma(x+1) = x\Gamma(x)$ (hint: integration by parts). The last expression is called the *functional equation* for the gamma function. Indeed, this tell us that

$$\Gamma(n) = (n-1)!.$$

The gamma function is difficult to evaluate if $x \notin \mathbb{N}$. For $x = 1/2$, you can use the Poisson integral formula,

$$\int_0^\infty e^{-t^2} \mathrm{d}t = \frac{1}{2}\sqrt{\pi}$$

to obtain that $\Gamma(1/2) = \sqrt{\pi}$.

6.4.2 Integration by parts

This is one of the most commonly used result. Be sure that you're familiar with this.

Theorem 18 (Integration by parts)**.** *Suppose that $f, g : [a, b] \to \mathbb{R}$ are continuous on $[a, b]$, differentiable on (a, b) and f', g' are Riemann integrable on $[a, b]$. Then,*

$$\int_a^b fg' = f(b)\, g(b) - f(a)\, g(a) - \int_a^b f'g.$$

Proof. That fg is continuous on $[a, b]$ follows from continuity of f and g. Since f and g are differentiable, by the product rule (Proposition 70), $(fg)' = f'g + fg'$. Since f, g, f' and g' are Riemann integrable, Proposition 79 implies that fg' , $f'g$ and $(fg)'$ are Riemann integrable. Define $h \coloneqq fg$. Then, by the fundamental theorem of calculus (Theorem 17 part (i)), we obtain

$$\begin{aligned} h(b) &= h(a) + \int_a^b h' \\ \Leftrightarrow f(b)\, g(b) - f(b)\, g(b) &= \int_a^b (f'g + g'f) = \int_a^b f'g + \int_a^b fg' \\ \Leftrightarrow \int_a^b fg' &= f(b)\, g(b) - f(b)\, g(b) - \int_a^b f'g. \end{aligned}$$

■

Remark 12. I was taught this using u and v instead of f and g. I always thought I get confused which way around but it turns out it doesn't matter. Anyway, a good way to heuristically derive integration by parts equation to remember $uv = (uv)' = u'v + v'u$. Taking integral then gives $uv = \int u'v + \int v'u$.

6.4.3 Change of variable

Theorem 19 (Change of variable)**.** *Suppose that $g : (a, b) \to \mathbb{R}$ is differentiable on (a, b) and g' is Riemann integrable on (a, b). Define $J \coloneqq g((a, b))$. If $f : J \to \mathbb{R}$ is continuous, then, for any $x, y \in (a, b)$,*

$$\int_a^b f(g(x))\, g'(x)\, \mathrm{d}x = \int_{g(a)}^{g(b)} f(t)\, \mathrm{d}t.$$

Proof. Define

$$F(x) \coloneqq \int_a^x f(t)\, \mathrm{d}t.$$

Since f is continuous, the fundamental theorem of calculus (Theorem 17 part (ii)) tells us that F is differentiable on J with $F' = f$. Chain rule gives us that $F \circ g : (a, b) \to \mathbb{R}$ is differentiable on (a, b) with

$$(F \circ g)'(x) = f(g(x))\, g'(x).$$

The derivative is Riemann integrable on $[a, b]$ since $f \circ g \in \mathbf{C}[a, b]$ (Proposition 78) so that it is Riemann integrable and g' is Riemann integrable by assumption. By the fundamental theorem of

calculus (Theorem 17 part (i)),

$$\begin{aligned}\int_a^b f(g(x))\,g'(x)\,\mathrm{d}x &= \int_a^b (F\circ g)'(x)\,\mathrm{d}x \\ &= (F\circ g)(b) - (F\circ g)(a) \\ &= F(g(b)) - F(g(a)) \\ &= \int_{g(a)}^{g(b)} F'(t)\,\mathrm{d}t = \int_{g(a)}^{g(b)} f(t)\,\mathrm{d}t.\end{aligned}$$

■

6.5 Riemann integral and limits

6.5.1 Convergence of sequences of functions

Suppose we have a converging sequence of functions, $f_n \to f$. We want to know whether this implies $\int f_n \to \int f$. We will introduce two notions of convergence of functions: pointwise and uniform convergence.

A sequence (f_n) in $\mathbb{R}^X$ *converges pointwise to* $f \in \mathbb{R}^X$ if $f_n(x) \to f(x)$ for every $x \in X$. Equivalently, for every $x \in X$, for any $\epsilon > 0$, there exists $N_{x,\epsilon} \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for all $n > N_{x,\epsilon}$.

A sequence (f_n) in $\mathbb{R}^X$ *converges uniformly to* $f \in \mathbb{R}^X$ if, for any $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ such that

$$\begin{aligned}&|f_n(x) - f(x)| < \epsilon \ \forall x \in X \ \forall n > N_\epsilon \\ \Leftrightarrow\ &\|f_n - f\|_\infty \equiv \sup\{|f_n(x) - f(x)| : x \in X\} < \epsilon \ \forall n > N_\epsilon.\end{aligned}$$

As can be seen from the dependence of N on x and ϵ, uniform convergence implies pointwise convergence, but not conversely. For example, consider the sequence (f_n) in $\mathbb{R}^{\mathbb{R}}$ defined by $f_n(x) \coloneqq x^2/n$ for each $n \in \mathbb{N}$. Fixing x,

$$\lim_{n\to\infty} f_n(x) = 0 =: f(x).$$

Hence, (f_n) converges pointwise to the zero function. However, (f_n) is not uniformly convergent. To see this, note first that

$$f_n(x) - f(x) = f_n(x) = \frac{x^2}{n}.$$

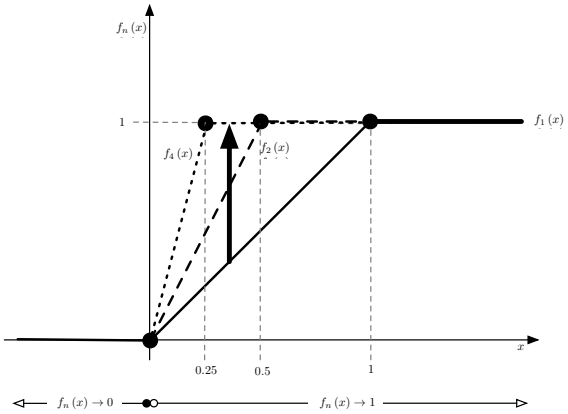
For uniform convergence, we must have that, for any $\epsilon > 0$, $|f_n(x)| < \epsilon$ for all $x \in \mathbb{R}$ for all $n > N$ for some N. However, since $|f_n(x)| \to \infty$ as $x \to -\infty, +\infty$, we cannot find such an N. Therefore, (f_n) is not uniformly convergent.

Note that pointwise limit of continuous functions may not be continuous. For example, consider

$$f_n(x) = \begin{cases} 0 & x \le 0 \\ nx & 0 < x < \frac{1}{n} \\ 1 & x \ge \frac{1}{n} \end{cases},$$

which is shown in the figure below.

Figure 6.1: Pointwise limit of a continuous function.



First, note that this function is continuous. And, as can be seen from the figure:

- if we fix any $x \leq 0$, then $f_n(x) = 0$ for all $n \in \mathbb{N}$ and so $f_n(x \leq 0) \to 0$;
- if we fix any $x > 0$, then $f_n(x) = 1$ for all $n \in \mathbb{N}$ and so $f_n(x > 0) \to 1$.

Hence, the pointwise limit of $f_n(x)$ is not continuous at $x = 0$. However, the following shows that uniform limit of continuous function is continuous.

Proposition 85. *Let (f_n) be a sequence of functions in $\mathbb{R}^X$ that are continuous and suppose that $f_n \to f \in \mathbb{R}^X$ uniformly. Then, f is continuous.*

Proof. Fix some $x \in X$ and $\epsilon > 0$. We wish to show that there exists $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon \ \forall y \in X : |x - y| < \delta.$$

Since $f_n \to f$ uniformly, there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < \frac{\epsilon}{3} \ \forall x \in X \ \forall n > N.$$

Since f_n are continuous, we can find a $\delta > 0$ such that

$$|f_{N+1}(x) - f_{N+1}(y)| < \frac{\epsilon}{3} \ \forall |x - y| < \delta.$$

Then, by the Triangle inequality, for any $y \in X$ such that $|x - y| < \delta$,

$$\begin{aligned} |f(x) - f(y)| &= |f(x) + f_{N+1}(x) - f_{N+1}(x) - f(y) + f_{N+1}(y) - f_{N+1}(y)| \\ &\leq |f_{N+1}(x) - f(x)| + |f_{N+1}(x) - f_{N+1}(y)| + |f_{N+1}(y) - f(y)| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

Hence, f is continuous. ■

Proposition 86. *Let (f_n) be a sequence of functions in $\mathbb{R}^X$ that are bounded and suppose that $f_n \to f \in \mathbb{R}^X$ uniformly. Then, f is bounded.*

Proof. Since $f_n \to f$ uniformly, setting $\epsilon = 1$, there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < 1$ for all $n > N$. Choose some $n > N$. Then, since f_n is bounded, there exists $M \geq 0$ such that

$|f_n(x)| \leq M$ for all $x \in X$. It follows that

$$|f(x)| = |f(x) + f_n(x) - f_n(x)| \leq |f(x) - f_n(x)| + |f_n(x)| < 1 + M \ \forall x \in X.$$

Hence, f is bounded. ■

A sequence (f_n) in $\mathbb{R}^X$ is *uniformly bounded* if there exists $M \in \mathbb{R}$ such that $|f_n(x)| \leq M$ for all $x \in X$ and all $n \in \mathbb{N}$.

Proposition 87. *Let (f_n) be a sequence of functions in $\mathbb{R}^X$ that are bounded and suppose that $f_n \to f \in \mathbb{R}^X$ uniformly. Then, f is uniformly bounded.*

Proof. Suppose $f_n \to f$. We wish to show that f is bounded; i.e., there exists $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $x \in X$. For each $n \in \mathbb{N}$, since f_n is bounded, there exists $M_n \in \mathbb{R}$ such that $|f_n(x)| \leq M_n$ for all $x \in X$. Since $f_n \to f$ uniformly, setting $\epsilon = 1$, there exists $N \in \mathbb{N}$ such that

$$|f_n(x) - f(x)| < 1 \ \forall x \in X \ \forall n > N.$$

Then, for $n > N$ and any $x \in X$,

$$\begin{aligned} |f_n(x)| &= |f_n(x) - f(x) - f(x) + f_{N+1}(x) - f_{N+1}(x)| \\ &\leq |f_n(x) - f(x)| + |f_{N+1}(x) - f(x)| + |f_{N+1}(x)| \\ &\leq 2 + M_{N+1}. \end{aligned}$$

Define $M \coloneqq \max\{M_1, \ldots, M_N, 2 + M_{N+1}\}$. Then, for any $x \in X$,

$$|f_n(x)| \leq M \ \forall n \in \mathbb{N}.$$

Hence, (f_n) is uniformly bounded. ■

6.5.2 A convergence theorem

Proposition 88. *Suppose that $f_n : [a,b] \to \mathbb{R}$ is Riemann integrable for each $n \in \mathbb{N}$ and $f_n \to f$ uniformly on $[a,b]$. Then, $f : [a,b] \to \mathbb{R}$ is Riemann integrable on $[a,b]$ and*

$$\int_a^b f \equiv \int_a^b \lim_{n\to\infty} f = \lim_{n\to\infty} \int_a^b f_n.$$

Proof. By Proposition 86, uniform limit of bounded function is bounded so f is bounded. Fix $\epsilon > 0$. Since $f_n \to f$ uniformly, there exists $N \in \mathbb{N}$ such that, for any $n > N$,

$$\begin{aligned} |f_n(x) - f(x)| &< \frac{\epsilon}{b-a} \ \forall a \leq x \leq b \\ \Leftrightarrow -\frac{\epsilon}{b-a} &< f_n(x) - f(x) < \frac{\epsilon}{b-a} \\ \Rightarrow f_n(x) - \frac{\epsilon}{b-a} &< f(x) < f_n(x) + \frac{\epsilon}{b-a}. \end{aligned} \tag{6.1}$$

This implies that

$$r\left(f_n - \frac{\epsilon}{b-a}\right) \leq r(f), \ R(f) \leq R\left(f_n + \frac{\epsilon}{b-a}\right).$$

Since f_n is integrable and $R(f) \geq r(f)$ by Proposition 76, above implies that

$$\left(\int_a^b f_n\right) - \epsilon = \int_a^b \left(f_n - \frac{\epsilon}{b-a}\right) \leq r\left(f\right)$$
$$\leq R\left(f\right) \leq \int_a^b \left(f_n + \frac{\epsilon}{b-a}\right) \leq \left(\int_a^b f_n\right) + \epsilon.$$

This, in turn, implies that

$$0 \leq R\left(f\right) - r\left(f\right) \leq \left(\int_a^b f_n\right) + \epsilon - \left(\left(\int_a^b f_n\right) - \epsilon\right) = 2\epsilon.$$

Since $\epsilon > 0$ was chosen arbitrary, above implies that $R(f) = r(f)$ (why?). Thus, f is integrable.

Moreover, since f_n for any $n > N$, we also have

$$\left|\int_a^b f_n - \int_a^b f\right| \leq \epsilon \ \forall \epsilon > 0.$$

Therefore, $\int_a^b f_n \to \int_a^b f$. ■

Pointwise convergence of functions is not sufficient to imply convergence of their integrals. For example, consider $f_n[0,1] \to \mathbb{R}$ defined by

$$f_n\left(x\right) \coloneqq n\mathbb{1}_{\left\{x \in \left(0, \frac{1}{n}\right)\right\}} = \begin{cases} n & x \in \left(0, \frac{1}{n}\right) \\ 0 & \text{otherwise} \end{cases}.$$

Then, $f_n \to f \equiv 0$ pointwise on $[0,1]$ so that $\int_0^1 f = 0$; however, $\int_0^1 f_n = 1$ for every $n \in \mathbb{N}$. Observe that f_n does not converge to f uniformly since, for each $n \in \mathbb{N}$,

$$\left\|f_n - f\right\|_\infty = \sup\left\{\left|n\mathbb{1}_{\left\{x \in \left(0, \frac{1}{n}\right)\right\}}\right| : x \in [0,1]\right\} = n.$$

Observe that fact that f_n's have a bounded domain is important. Consider the following

$$f_n\left(x\right) \coloneqq \frac{1}{n}\mathbb{1}_{\left\{x \in (0,n)\right\}} = \begin{cases} \frac{1}{n} & x \in (0,n) \\ 0 & \text{otherwise} \end{cases}.$$

We again have that $f_n \to f \equiv 0$ pointwise on $[0,1]$ but $\int_0^1 f_n = 1 \neq 0 = \int_0^1 f$. However,

$$\left\|f_n - f\right\|_\infty = \sup\left\{\left|\frac{1}{n}\mathbb{1}_{\left\{x \in (0,n)\right\}}\right| : x \in [0,1]\right\} = \frac{1}{n}$$

so that the convergence is, in fact, uniform. The problem here is that f_n's effectively have an unbounded domain of $\mathbb{R}$ so that we cannot infer (6.1) from the fact that f_n converges uniformly to f.

A more problematic feature of the Riemann integral is that the pointwise limit of integrable functions need not be integrable (even if bounded). For example, let $\{q_k : k \in \mathbb{N}\}$ be an enumeration of the rationals in $[0,1]$ and define $g_n[0,1] \to \mathbb{R}$ as

$$f_n\left(x\right) \coloneqq \begin{cases} 1 & \text{if } x = q_k \text{ for } k \leq n \\ 0 & \text{otherwise} \end{cases}.$$

Then, each f_n is Riemann integrable since it differs from zero function at finitely many points. However, f_n converges pointwise on $[0,1]$ to the Dirichlet function, which is not Riemann integrable. As we will see later, Lebesgue integral has better properties than Riemann integrable.

We have just showed that uniformly convergent sequence of functions that are Riemann integrable is a Riemann integrable at the limit. Is the limiting function also differentiable? In general, the answer is no. Here, we provided a sufficient condition for $f_n \to f$ to imply $f_n' \to f$. We do so by using the results about the convergence of integrals (together with the fundamental theorem of calculus) rather than about convergence of derivatives directly.

Proposition 89. *Let $f_n : (a,b) \to \mathbb{R}$ be a sequence of differentiable functions whose derivatives $f_n' : (a,b) \to \mathbb{R}$ are Riemann integrable on (a,b). Suppose that $f_n \to f$ pointwise and $f_n' \to g$ uniformly on (a,b), where $g : (a,b) \to \mathbb{R}$ is continuous. Then, $f : (a,b) \to \mathbb{R}$ is continuously differentiable on (a,b) and $f' = g$; i.e.,*

$$f(x) = \int f'$$

Proof. Fix $c \in (a,b)$. Since f_n' is Riemann integrable, by the fundamental theorem of calculus (Theorem 17),

$$f_n(x) = f_n(c) + \int_c^x f_n' \ \forall x \in (a,b).$$

Since $f_n \to f$ pointwise, and $f_n' \to g$ uniformly on $[a,x]$, by Proposition 88, we have

$$f(x) = f(c) + \int_c^x g.$$

Since g is continuous, by the fundamental theorem of calculus (Theorem 17) again, we have that f is differentiable in (a,b) and $f' = g$. ■

6.5.3 Differentiating under the integral and the Leibniz' rule

Suppose that $f : [a,b] \times [c,d] \to \mathbb{R}$ and that $f(x,t)$ is integrable with respect to t for each $x \in [a,b]$. Define $F : [a,b] \to \mathbb{R}$ as

$$F(x) \coloneqq \int_a^b f(x,t)\,\mathrm{d}t.$$

We would expect the the derivative of F to be

$$\frac{\mathrm{d}F}{\mathrm{d}x}(x) = \int_a^b \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t,$$

which, by definition of a derivative, is equivalent to

$$\begin{aligned}\frac{\mathrm{d}F}{\mathrm{d}x}(x) &\equiv \lim_{h\to 0}\frac{F(x+h)-F(x)}{h} \\ &= \lim_{h\to 0}\int_a^b \frac{f(x+h,t)-f(x,t)}{h}\mathrm{d}t = \int_a^b \lim_{h\to 0}\frac{f(x+h,t)-f(x)}{h}\mathrm{d}t \equiv \int_a^b \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t.\end{aligned}$$

Thus, the question we are asking is is equivalent to whether we can move the limit inside the integral. By Proposition 88, we can move the limit inside integral as above if we know that $g_n \to \frac{\partial f}{\partial x}$ uniformly, where

$$g_n(x,t) \coloneqq \frac{f(x+h_n,t)-f(x,t)}{h_n},$$

for any sequence (h_n) (in $[a,b]$ such that $x + h_n \in [a,b]$) that converges to 0.

Moreover, provided that the fundamental theorem of calculus (Theorem 17) holds, since

$$\int_a^b f(x,t)\,\mathrm{d}t = F_x(b) - F_x(a),$$

where $F_x(t) \coloneqq \int_a^t f(x,t)\mathrm{d}t$, we have

$$\frac{\mathrm{d}}{\mathrm{d}b}\int_a^b f(x,t)\,\mathrm{d}t = \frac{\mathrm{d}}{\mathrm{d}b}\left(F_x(b) - F_x(a)\right) = F_x'(b) = f(x,b).$$

Similarly,

$$\frac{\mathrm{d}}{\mathrm{d}a}\int_a^b f(x,t)\,\mathrm{d}t = -f(x,a).$$

Thus, if the limits a and b are, in fact, functions of x; say $u, v : [a,b] \to \mathbb{R}$, by the chain rule,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}F(x) &\equiv \frac{\mathrm{d}}{\mathrm{d}x}\left(\int_{u(x)}^{v(x)} f(x,t)\,\mathrm{d}t\right)\\ &= \frac{\partial}{\partial b}\left(\int_a^b f(x,t)\,\mathrm{d}t\right)\frac{\mathrm{d}v}{\mathrm{d}x}(x) + \frac{\partial}{\partial a}\left(\int_a^b f(x,t)\,\mathrm{d}t\right)\frac{\mathrm{d}u}{\mathrm{d}x}(x) + \frac{\mathrm{d}}{\mathrm{d}x}\left(\int_a^b f(x,t)\,\mathrm{d}t\right)\Bigg|_{a=u(x),b=v(x)}\\ &= f(x,b)\,v'(x) - f(x,a)\,u'(x) + \int_a^b \frac{\partial f}{\partial x}(x,t)\,\mathrm{d}t.\end{aligned}$$

Above is called the Leibniz integral rule, which we make explicit below.

Theorem 20 (Leibniz integral rule)**.** *Let $f(x,t)$ and $D_x f(x,t)$ be continuous on $[a,b]\times[c,d] \subseteq \mathbb{R}^2$. Suppose that $u, v \in \mathbf{C}^1[a,b]$ and $u([a,b]), v([a,b]) \subseteq [c,d]$. Define*

$$F(x) \coloneqq \int_{u(x)}^{v(x)} f(x,t)\,\mathrm{d}t.$$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x,v(x))\frac{\mathrm{d}v}{\mathrm{d}x}(x) - f(x,u(x))\frac{\mathrm{d}u}{\mathrm{d}x}(x) + \int_{u(x)}^{v(x)} \frac{\partial f(x,t)}{\partial x}\mathrm{d}t.$$

6.5.4 Exchanging Riemann integrals

Suppose that $f : [0,1]^2 \to \mathbb{R}$. What are the relationships between the following expressions (assuming they are well defined)?

$$\int_{(x,y)\in[0,1]^2} f(x,y)\,\mathrm{d}(x,y), \quad \int_0^1\left(\int_0^1 f(x,y)\,\mathrm{d}x\right)\mathrm{d}y, \text{ and } \int_0^1\left(\int_0^1 f(x,y)\,\mathrm{d}y\right)\mathrm{d}x. \tag{6.2}$$

You might be tempted to assume that they are all equal. Consider the following example. Define $f : [0,1]^2 \to \mathbb{R}$ by

$$f(x,y) \coloneqq \mathbb{1}_{\{x=1,\ y\in\mathbb{Q}\}}.$$

Fix any $y \in [0,1]$, then $f(x,y)$ is a function that is zero except at a single point $x = 1$. As per Example 28 above, such a function is Riemann integrable and its Riemann integral is 0. Thus, it follows that $\int_0^1\int_0^1 f(x,y)\mathrm{d}x\mathrm{d}y = 0$. However, for any fixed $x \in [0,1]$, $f(x,y)$ is a Dirichlet's function, which we know is not Riemann integrable (Example 29). Hence, $\int_0^1\int_0^1 f(x,y)\mathrm{d}y\mathrm{d}x$ does

not exist. (It can be shown that $\int_{\mathbf{z}\in[0,1]^2} f(\mathbf{z})\mathrm{d}\mathbf{z}$ is Riemann integrable and its Riemann integral is zero.)

The following gives us a sufficient condition to ensure that all the expressions in (6.2) are equal. The result is useful for at least two reasons. First, it tells us that we can write Riemann integrals over hyperrectangles as composition of Riemann integrals over each of the associated intervals. Thus, it allows us to apply the fundamental theorem of calculus (Theorem 17) to Riemann integrals over hyperrectangles. Second, the result tells us when we can exchange the order of integration.

Theorem 21 (Fubini's Theorem for Riemann integral)**.** *Let $f : [a,b]\times[c,d] \to \mathbb{R}$ and f be Riemann integrable. If $\int_c^d f(x,y)\mathrm{d}y$ and $\int_a^b \int_c^d f(x,y)\mathrm{d}y\mathrm{d}x$ exist, then*

$$\int_{(x,y)\in[a,b]\times[c,d]} f(x,y)\,\mathrm{d}(x,y) = \int_a^b \int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x.$$

If $\int_a^b f(x,y)\mathrm{d}x$ and $\int_c^d \int_a^b f(x,y)\mathrm{d}x\mathrm{d}y$ exist, then

$$\int_{(x,y)\in[a,b]\times[c,d]} f(x,y)\,\mathrm{d}(x,y) = \int_c^d \int_a^b f(x,y)\,\mathrm{d}x\mathrm{d}y.$$

Proof. Take any $C \coloneqq (C_{ij}) \in \mathcal{D}([a,b],[c,d])$, where $C = \mathbf{a}\times\mathbf{c} = [a_0,\ldots,a_n]\times[c_0,\ldots,c_m]$ for $\mathbf{a}\in\mathcal{D}[a,b]$ and $\mathbf{c}\in\mathcal{D}[c,d]$, and $C_{ij} = [a_{i-1},a_i]\times[c_{j-1},c_j]$. Since, for any $(x,y)\in C_{ij}$,

$$\inf_{C_{ij}} f \le f(x,y) \le \sup_{C_{ij}} f,$$

Integrating with respect to y over $[c_{j-1},c_j]$ yields

$$\left(\inf_{C_{ij}} f\right)(c_j - c_{j-1}) \le \int_{c_{j-1}}^{c_j} f(x,y)\,\mathrm{d}y \le \left(\sup_{C_{ij}} f\right)(c_j - c_{j-1}),$$

where we used the fact that $f(x,y)$ is integrable over y and Proposition 79. Summing across j, and by Proposition 79 again, we obtain

$$\sum_{j=1}^m \left(\inf_{C_{ij}} f\right)(c_j - c_{j-1}) \le \sum_{j=1}^m \int_{c_{j-1}}^{c_j} f(x,y)\,\mathrm{d}y = \int_c^d f(x,y)\,\mathrm{d}y \le \sum_{j=1}^m \left(\sup_{C_{ij}} f\right)(c_j - c_{j-1}).$$

Since this holds for each $x\in[a_{i-1},a_i]$, integrating with respect to x over $[a_{i-1},a_i]$ yields (which we can do since $\int_c^d f(x,y)\,\mathrm{d}y$ exists and thus $f(x,y)$ is integrable over x),

$$\sum_{j=1}^m \left(\inf_{C_{ij}} f\right)(c_j - c_{j-1})(a_i - a_{i-1}) \le \int_{a_{i-1}}^{a_i}\int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x \le \sum_{j=1}^m \left(\sup_{C_{ij}} f\right)(c_j - c_{j-1})(a_i - a_{i-1}).$$

Summing over i gives

$$\sum_{i=1}^n\sum_{j=1}^m \left(\inf_{C_{ij}} f\right)(c_j - c_{j-1})(a_i - a_{i-1}) \le \int_a^b\int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x \le \sum_{i=1}^n\sum_{j=1}^m \left(\sup_{C_{ij}} f\right)(c_j - c_{j-1})(a_i - a_{i-1}).$$

That is,

$$r_C(f) \le \int_a^b\int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x \le R_C(f).$$

Since this holds for any $C \in \mathcal{D}([a,b],[c,d])$, it follows that

$$r(f) \leq \int_a^b \int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x \leq R(f).$$

By assumption that f is Riemann integrable and thus

$$\int_a^b \int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x = R(f) = \int_a^b \int_c^d f(x,y)\,\mathrm{d}y\mathrm{d}x. \qquad \blacksquare$$

Remark 13. To extend the previous result to the case when the domain of f is a d-dimensional hyperrectangle with $d > 2$, we need f to be continuous.

When the limit depends on the variable of integration, we have to be careful when exchanging the order of integration. Consider exchanging the order of integral of the following double integral:

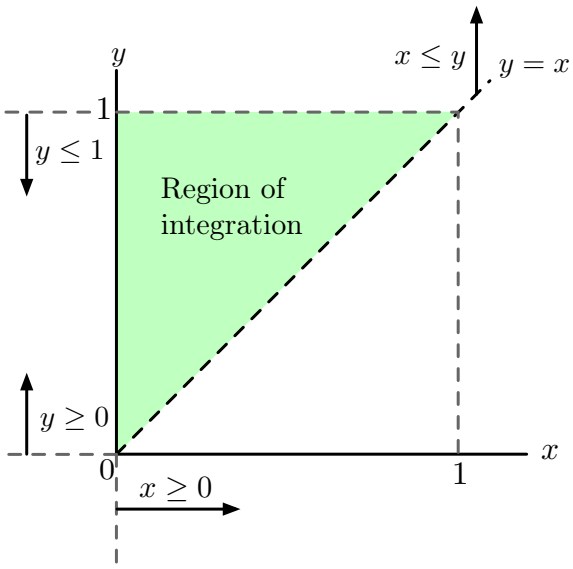
$$\int_0^1 \int_0^y f(x,y)\,\mathrm{d}x\mathrm{d}y,$$

where the region of integration for x depends on y. The inner integration is with respect to x from zero to y; i.e., $0 \leq x \leq y$. The outer integration is with respect to y from 0 to 1 and so $0 \leq y \leq 1$. Together, we therefore have

$$0 \leq x \leq y \leq 1.$$

We can plot this region in (x,y) space as below.

Figure 6.2: Region of integration.



In the figure, for each $y \in [0,1]$, the inner integration takes the horizontal distance up to from $x = 0$ to $x = y$, and the outer integration sums the horizontal distances for all different values of $y \in [0,1]$. We wish now to integrate with respect to y first, and then x. In other words, we wish to integrate vertically for each x, and then sum the vertical distances for all different values of x Since the region of integration lies to the left of $y = x$; then it must be that $y \geq x$. Of course, upper

bound for y is one. On the other hand, we would integrate over the whole domain of x. So,

$$\int_0^1 \int_0^y f(x,y)\,\mathrm{d}x\mathrm{d}y = \int_0^1 \int_x^1 f(x,y)\,\mathrm{d}y\mathrm{d}x.$$

Observe that, on the right-hand side, we are (still) integrating over $0 \leq x \leq y \leq 1$.

Part II
Optimisation

7 Static optimisation

Our goal is to be able to *solve* the following type of problem, which we will call the *primal problem*.

$$\begin{aligned} \sup_{\mathbf{x}\in\mathbb{R}^n} \; & f(\mathbf{x}) \\ \text{s.t. } & h_k(\mathbf{x}) = 0 \; \forall k \in \{1,\dots,K\} \\ & g_j(\mathbf{x}) \geq 0 \; \forall j \in \{1,\dots,J\}, \end{aligned}$$

where $f, h_k, g_j : \mathbb{R}^n \to \mathbb{R}$ for all $k \in \{1,\dots,K\}$ and $j \in \{1,\dots J\}$. By "solve", we want to obtain the set of maximisers, i.e., $\mathbf{x}$ that satisfy the constraints and maximises the objective function. This, in turn, allows us to compute the maximised objective function. We will also think about comparative statistics; i.e., how the maximisers and thus the objective function change as we "vary" the optimisation problem.

Since $\sup -f = -\inf f$, once we know how to solve maximisation problems, we also know how to solve minimisation problems. It is pedagogical fact that in economics, we focus on maximisation problems, whereas in mathematics/computer science, the focus is on minimisation problems.

Let us note some help properties of optimisation problems. First, relaxing constraints cannot lead to lower value of maximised objective function. Second, if we transform the objective function with a strictly increasing function, then the set of maximisers do not change. For example, you will often use this to take maximise $\ln f$ instead of f.

7.1 Maxima and minima

Given a function $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$, an element $\mathbf{x} \in X$ is a *global maximum* of f if

$$f(\mathbf{x}) \geq f(\mathbf{y}) \;\; \forall \mathbf{y} \in X \backslash \{\mathbf{x}\}.$$

The element $\mathbf{x} \in X$ is a *strict global maximum* if the inequality above holds strictly (and if it exists, it must be unique). An element $\mathbf{x} \in X$ is a *local maximum* of f if, for some $\epsilon > 0$,

$$f(\mathbf{x}) \geq f(\mathbf{y}) \;\; \forall \mathbf{y} \in B_\epsilon(\mathbf{x}) \setminus \{\mathbf{x}\}.$$

The element $\mathbf{x} \in X$ is a *strict local maximum* if the inequality above holds strictly. Local minima are defined analogously with the inequalities reversed.

Proposition 90. *Let $X \subseteq \mathbb{R}^d$ be convex, If $f : X \to \mathbb{R}$ is strictly quasi-concave, any local maximum f is a global maximum of f on X and the set of maximisers contains at most one element.*

Proof. Now suppose f is strictly quasiconcave and the $\mathbf{x} \in X$ is a local maximum of f on X. Thus, there exists $\epsilon > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{y})$ for all $\mathbf{y} \in B_\epsilon(\mathbf{x})$. Toward a contradiction, suppose that $\mathbf{x}$ is not a global maximum; i.e., there exists $\mathbf{z} \in X$ such that $f(\mathbf{z}) > f(\mathbf{x})$. But then since X is convex, for any $\alpha \in [0,1]$, $\alpha\mathbf{x} + (1-\alpha)\mathbf{z} \in X$ and by the quasiconcavity of f, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{z}) > \min\{f(\mathbf{x}), f(\mathbf{z})\} = f(\mathbf{x}) \;\; \forall \alpha \in (0,1).$$

In particular, for sufficiently large $\alpha > 0$, $\alpha\mathbf{x} + (1-\alpha)\mathbf{z} \in B_r(\mathbf{x})$ and hence we must have $f(\mathbf{x}) \geq f(\alpha\mathbf{x} + (1-\alpha)\mathbf{z})$; a contradiction.

Suppose that there are two global maximisers, $\mathbf{x}, \mathbf{y} \in X$. Pick any $\alpha \in (0,1)$ and define $\mathbf{z} \coloneqq \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in X$. By strict quasiconcavity,

$$f(\mathbf{z}) > \min\{f(\mathbf{x}), f(\mathbf{y})\} = f(\mathbf{x}) = f(\mathbf{y}).$$

But this contradicts the fact that $\mathbf{x}$ and $\mathbf{y}$ were global maximisers. ■

7.2 First-order approach

Let us assume that objective function and the constraints are differentiable.

7.2.1 Unconstrained optima

Consider the problem of maximising a function without any constraints.

Proposition 91. *Suppose $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ has a local maximum or a local minimum at $x^* \in \operatorname{int}(X)$ and that f is differentiable at x^*. Then, $\mathbf{x}^*$ satisfies the first-order condition; i.e.,*

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \Leftrightarrow \frac{\partial f}{\partial x_i}(\mathbf{x}^*) = 0 \ \forall i \in \{1, \ldots, n\}.$$

Any $x \in \operatorname{int}(X)$ that satisfies the first-order condition is called a *critical point of f*. Although first-order condition is necessary for a point to be a local maximum or a local minimum (assuming differentiability), it is not sufficient (e.g., $f(x) = x^3$). This leads us to the idea of second-order conditions that helps us distinguish between local maxima and minima.

Proposition 92. *Suppose f is twice continuously differentiable on $\operatorname{int}(X) \subseteq \mathbb{R}^n$ and $\mathbf{x} \in \operatorname{int}(X)$.*

(i) If f has a local maximum at $\mathbf{x}$, then $H_f(\mathbf{x})$ is negative semidefinite.

(ii) If f has a local minimum at $\mathbf{x}$, then $H_f(\mathbf{x})$ is positive semidefinite.

These results has nothing to say about the case in which $\mathbf{x} \in \operatorname{int}(X)$ is such that $\mathbf{x}$ satisfies the first-order condition (i.e., $\nabla f(\mathbf{x}) = 0$)) and $H_f(\mathbf{x})$ is negative (resp. positive) semidefinite but not negative (resp. positive) definite. Such a point can be a local maximum, a local minimum or neither. In other words, first- and second-order conditions are all that useful in knowing whether a point is a local maximum (or minimum). We need to impose stronger conditions such as concavity or quasiconcavity of f.

7.3 Constrainted optimisation: Theorem of Karush-Kuhn-Tucker

The theorem of KKT gives us a set of necessary conditions for constrained optimisation problems. Throughout we assume that both the objective and the constraint functions are continuously differentiable.

7.3.1 A "cook-book" approach

Let us focus on the following problem with inequality and nonnegativity constraints.

$$\max_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \text{ s.t. } h_k(\mathbf{x}) = 0 \ \forall k \in \{1, \ldots, K\},$$
$$g_j(\mathbf{x}) \geq 0 \ \forall j \in \{1, \ldots, J\}$$

The following cookbook approach "usually" works.

Step 1 Define the *Lagrangian* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^K \times \mathbb{R}^J \to \mathbb{R}$ as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \coloneqq f(\mathbf{x}) + \sum_{k=1}^{K} \mu_k h_k(\mathbf{x}) + \sum_{j=1}^{J} \lambda_j g_j(\mathbf{x})$$

Step 2 Find all solutions $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ to the following set of equations and inequalities

(i) First-order conditions:

$$\frac{\partial \mathcal{L}}{\partial x_i}(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \equiv \frac{\partial f}{\partial x_i}(\mathbf{x}^*) + \sum_{j=1}^{K} \mu_k \frac{\partial h_k}{\partial x_i}(\mathbf{x}) + \sum_{j=1}^{J} \lambda_j^* \frac{\partial g_j}{\partial x_i}(\mathbf{x}^*) = 0 \ \forall i \in \{1, \ldots, n\}.$$

(ii) Nonnegativity:

$$\lambda_j^* \geq 0 \ \forall j \in \{1, \ldots J\}.$$

(iii) Complementary slackness (for inequality constraints):

$$\lambda_j^* \frac{\partial \mathcal{L}}{\partial \lambda_j}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \lambda_j^* g_j(\mathbf{x}^*) = 0 \ \forall j \in \{1, \ldots J\}.$$

(iv) Constraints:

$$h_k(\mathbf{x}^*) = 0 \ \forall k \in \{1, \ldots, K\},$$
$$g_j(\mathbf{x}^*) \geq 0 \ \forall j \in \{1, \ldots, J\}.$$

Step 3 Compute the value of f at each solution found in the previous step, and choose the one that maximises f as the solution to the problem.

Example 31. Consider the following problem:

$$\max_{x,y \in \mathbb{R}} \ x - y^2 \text{ s.t. } x^2 + y^2 = 4, \quad x, y \geq 0.$$

The Lagrangian is

$$\mathcal{L}(x, y, \mu, \lambda_1, \lambda_2) = x - y^2 + \mu\left(x^2 + y^2 - 4\right) + \lambda_1 x + \lambda_2 y.$$

First-order conditions

$$\frac{\partial \mathcal{L}}{\partial x} = 1 + 2\mu^* x^* + \lambda_1^* = 0,$$
$$\frac{\partial \mathcal{L}}{\partial y} = -2y^* + 2\mu^* y^* + \lambda_2^* = 0.$$

Nonnegativity conditions:

$$\lambda_1^* \geq 0, \ \lambda_2^* \geq 0.$$

Complementary slackness conditions:

$$\lambda_1^* x_1^* = \lambda_2^* y^* = 0.$$

Constraints:

$$(x^*)^2 + (y^*)^2 = 4, \ x^* \geq 0, \ y^* \geq 0.$$

Rearranging the first-order condition with respect to x gives

$$1 + \lambda_1^* = -2\mu^* x^*.$$

If $x^* = 0$, then $\lambda_1^* < 0$ to satisfy the equation above but $\lambda_1^* \geq 0$ by nonnegativity. Thus, we must have $x^* \neq 0$. Since $x^* \geq 0$, we realise that $x^* > 0$. Furthermore, since $1 + \lambda_1^* > 0$, the equation above also implies that $\mu^* < 0$. Complementary slackness means that $\lambda_1^* = 0$ because $x^* = 0$. Rearranging the first-order condition with respect to y gives

$$2y^* (1 - \mu^*) = \lambda_2^*.$$

Since $1 - \mu^* > 0$, either (i) $y^* = \lambda_2^* = 0$ or (ii) $y^*, \lambda_2^* > 0$. But the latter would contradict complementary slackness condition $\lambda_2^* y^* = 0$. Hence, we must have $y^* = \lambda_2^* = 0$. The equality constraint and the nonnegativity constraint gives that $x^* = 2$. Finally, the first-order condition with respect to x gives $\mu^* = -\frac{1}{4}$. That is, the solution is

$$(x^*, y^*, \mu^*, \lambda_1^*, \lambda_2^*) = \left(2, 0, -\frac{1}{4}, 0, 0\right).$$

Remark 14. If we were solving a minimisation problem, then define

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \coloneqq f(\mathbf{x}) - \sum_{j=1}^{J} \lambda_j g_j(\mathbf{x})$$

in step 1 and, among the solutions find in step 2, choose the solution that minimises f as the solution to the problem.

Remark 15. It is often easy to forget which sign we should have the constraint part in $\mathcal{L}$. One way to is to remember that the Lagrange multiplier are always positive and think of the Lagrangian multipliers as the *punishment* to the objective function. So suppose we are maximising f and we have a constraint that takes the form $g_j(\mathbf{x}) \geq 0$. Then, we must punish times in which $g_j(\mathbf{x}) < 0$ and so it must be the case that we add $\lambda_j g_j(\mathbf{x})$ to f in writing the Lagrangian so that negative $g_j(\mathbf{x})$ would lower (and thus punish) the value of the objective. Similarly, if $g_j(\mathbf{x}) \leq 0$, then we must punish the objective whenever $g_j(\mathbf{x}) > 0$ and so we add $-\lambda_j g_j(\mathbf{x})$ to f in forming the Lagrangian.

Remark 16. Since

$$\begin{aligned} h_k(\mathbf{x}) = 0 &\Leftrightarrow (h_k(\mathbf{x}) \leq 0) \wedge (h_k(\mathbf{x}) \geq 0) \\ &\Leftrightarrow (-h_k(\mathbf{x}) \geq 0) \wedge (h_k(\mathbf{x}) \geq 0), \end{aligned}$$

any equality constraints can be expressed as two inequality constraints. Thus, the optimisation problem can be expressed as a problem consisting only of inequality constraints.

7.3.2 When does it work?

So when does this approach work? First, there must be a solution to a problem. As we will cover in ECON 6701, Weirstrass Theorem usually gives us this. The other condition is that a condition called *constraint qualification* must be satisfied at solution $\mathbf{x}^*$. This is a requirement that Jacobian matrix of the equality constraints and the binding constraints at $\mathbf{x}^*$ has full rank. We will go over this latter condition in more detail also in ECON 6701. In the meantime, you should just make sure that the following condition holds instead.

(i) The objective f as well as the constraints $g_1, g_2 \ldots, g_J$ are all concave, continuously differentiable functions.

(ii) Slater's condition

$$\exists \mathbf{x} \in \operatorname{int}(X),\ g_i(\mathbf{x}) > 0\ \forall i \in \{1,\ldots,J\}$$

holds.

We will see in ECON 6701 that it is possible to relax the requirement of concavity of the objective and constraints to quasiconcavity; however, this comes at the cost of a more difficult condition to verify than the Slater's condition above.

7.3.3 Nonnegativity constraints

Suppose the problem involves only inequality constraints but there are nonnegativity constraints:

$$\begin{aligned}\max_{\mathbf{x}\in\mathbb{R}^n}\ f(\mathbf{x})\ \text{s.t.}\ & g_j(\mathbf{x}) \geq 0\ \forall j \in \{1,\ldots,J\},\\ & x_i \geq 0\ \forall i \in \{1,\ldots,n\}.\end{aligned}$$

An alternative way to approach this problem is to consider the Lagrangian that ignores the nonnegativity constraints (called Kuhn-Tucker Lagrangian):

$$\tilde{\mathcal{L}}(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{j=1}^{J} \lambda_j g_j(\mathbf{x}).$$

Then, compute $(\mathbf{x}^*,\boldsymbol{\lambda}^*)$ that satisfies the following conditions: $\lambda_j^* \geq 0$ for all $j \in \{1,\ldots,J\}$ and

$$\begin{aligned}\frac{\partial \tilde{\mathcal{L}}}{\partial x_i} \leq 0\ \forall i \in \{1,\ldots,n\},\quad \frac{\partial \tilde{\mathcal{L}}}{\partial \lambda_j} \leq 0\ \forall j \in \{1,\ldots,J\},\\ x_i^* \frac{\partial \tilde{\mathcal{L}}}{\partial x_i} = 0\ \forall i \in \{1,\ldots,n\}, \lambda_j^* \frac{\partial \tilde{\mathcal{L}}}{\partial \lambda_j} = 0\ \forall j \in \{1,\ldots,J\}.\end{aligned}$$

Example 32. Consider again the following problem:

$$\begin{aligned}\max_{x,y\in\mathbb{R}}\ x - y^2\ \text{s.t.}\ & x^2 + y^2 = 4,\\ & x, y \geq 0.\end{aligned}$$

Treating the equality constraints as two inequality constraints, we have

$$\begin{aligned}\tilde{\mathcal{L}}(x,y,\lambda_1,\lambda_2) &= x - y^2 + \lambda_1\left(x^2+y^2-4\right) - \lambda_2\left(x^2+y^2-4\right)\\ &= x - y^2 + (\lambda_1-\lambda_2)\left(x^2+y^2-4\right).\end{aligned}$$

The conditions are that $\lambda_1^*, \lambda_2^* \geq 0$,

$$\begin{aligned}1 + 2(\lambda_1^* - \lambda_2^*)x^* &\leq 0,\\ -2y^* + 2(\lambda_1^* - \lambda_2^*)y^* &\leq 0,\\ (x^*)^2 + (y^*)^2 - 4 &= 0,\\ x^*\left(1 + 2(\lambda_1^* - \lambda_2^*)x^*\right) &= 0,\\ y^*\left(-2y^* + 2(\lambda_1^* - \lambda_2^*)y^*\right) &= 0.\end{aligned}$$

Note that $\lambda_1^* \neq \lambda_2^*$ and $x^* > 0$; otherwise the first line would imply $1 \leq 0$. But the first inequality

implies that $\lambda_1^* - \lambda_2^* < 0$. Letting $\lambda^* \coloneqq \lambda_2^* - \lambda_1^*$The penultimate equation also implies

$$1 + 2\lambda^* x^* = 0 \Leftrightarrow x^* = -\frac{1}{2\lambda^*} > 0.$$

The equality constraint then gives that

$$(y^*)^2 = 4 - \frac{1}{4}\frac{1}{(\lambda^*)^2}.$$

Suppose $y^* = 0$. Since $\lambda^* < 0$ and

$$4 - \frac{1}{4}\frac{1}{\lambda^2} = 0 \Leftrightarrow \lambda^2 = \frac{1}{16} \Leftrightarrow \lambda = \frac{1}{4} \text{ or } -\frac{1}{4},$$

we must have $\lambda^* = -\frac{1}{4}$ and $x^* = 2$ and so we have $x^* = 2$ and $y^* = 0$ as we had before!

7.4 Comparative statics on optimisation problems

Consider the following problem:

$$f^*(\theta) \coloneqq \max_{x \in \Gamma(\theta)} f(x, \theta),$$

where we think of x as the choice variables (e.g., consumption), θ as the parameter of the problem (e.g., prices), and $\Gamma(\theta)$ as the constraint set that can also depend on the parameters. Let $\Gamma^*(\theta)$ denote the set of maximisers; i.e., $x^* \in \Gamma^*(\theta)$ if and only if $f^*(\theta) = f(x^*, \theta)$ and $x^* \in \Gamma(\theta)$. The goal of a comparative static exercise is to see how exactly $f^*(\theta)$ or $\Gamma^*(\theta)$ changes with the parameter θ. Our goal here is to provide such comparative statistic results.

7.4.1 Implicit function theorem

We are often interested in how the maximisers vary with parameters. We can use implicit function theorem to obtain results for this type of comparative static.

To illustrate how it can be used, suppose that $f : X \subseteq \mathbb{R} \to \mathbb{R}$ and we know that the first-order condition is necessary and sufficient to characterise a unique maximiser for a given parameter value θ,

$$\frac{\partial}{\partial x} f(x^*, \theta) = 0.$$

By the implicit function theorem, we can think of x^* as being a function of θ around θ so that we may write

$$\frac{\partial}{\partial x} f(x^*(\theta), \theta) = 0.$$

Since this holds in the neighbourhood of θ, we may differentiate both sides with respect to θ, and using chain rule yields that

$$\frac{\partial^2}{\partial x \partial x} f(x^*(\theta), \theta) \frac{\mathrm{d}x^*}{\mathrm{d}\theta}(\theta) + \frac{\partial^2}{\partial x \partial \theta} f(x^*(\theta), \theta) = 0.$$

Rearranging gives that

$$\frac{\mathrm{d}x^*}{\mathrm{d}\theta}(\theta) = -\frac{\frac{\partial^2}{\partial x \partial \theta} f(x^*(\theta), \theta)}{\frac{\partial^2}{\partial x \partial x} f(x^*(\theta), \theta)}.$$

Observe that for x^* to be a maximum, the denominator must be negative. Hence, the sign of the derivative depends on the sign of the cross derivative in the numerator. Of course, if we can compute functional forms of the second derivatives, then we can compute the $\mathrm{d}x^*/\mathrm{d}\theta$ exactly.

7.4.2 Envelope theorem

Envelope Theorems helps us study how the maximised objective changes with parameters (note there are many versions!).

Suppose that $f : X \subseteq \mathbb{R} \to \mathbb{R}$ and we know that the first-order condition is necessary and sufficient to characterise a unique maximiser, x^*, for a given parameter value θ,

$$\frac{\partial}{\partial x} f\left(x^*, \theta\right) = 0.$$

By the implicit function theorem, we can think of x^* as being a function of θ around θ so that we may write the maximised objective function as

$$f^*\left(\theta\right) = f\left(x^*\left(\theta\right), \theta\right).$$

Since this holds in the neighbourhood of θ, we may differentiate both sides with respect to θ, and using chain rule yields that

$$\frac{\mathrm{d}f^*}{\mathrm{d}\theta}\left(\theta\right) = \frac{\partial f}{\partial x}\left(x^*\left(\theta\right), \theta\right) \frac{\mathrm{d}x^*}{\mathrm{d}\theta}\left(\theta\right) + \frac{\partial f}{\partial \theta}\left(x^*\left(\theta\right), \theta\right).$$

The second term captures the direct effect of a change in θ on the objective function while the first term captures the indirect effect of a change in θ through its effect on the maximiser. However, due to to the first-order condition, the indirect effect is, in fact, zero; i.e.,

$$\frac{\mathrm{d}f^*}{\mathrm{d}\theta}\left(\theta\right) = \frac{\partial f}{\partial \theta}\left(x^*\left(\theta\right), \theta\right).$$

Therefore, the effect of change in θ on the value of the maximised objective function comes solely from the direct effect that θ has on f. All envelope theorems have this flavour (with differing assumptions).

When we study constrained optimisation problems, we can think of taking the derivative of the Lagrangian, rather than (just) the objective function.

8 Dynamic optimisation

A dynamic problem is one which actions today have consequences for future periods. We will study both *discrete time* (i.e., $t \in \{0\} \cup \mathbb{N}$} and *continuous time* (i.e., $t \in [0, \infty)$). The goal is to obtain a path of relevant variables that maximises a discounted sum of payoffs over time. In this section, we show the conditions under which Euler Equations and Transversality Conditions are necessary and sufficient for a path to be optimal. We then introduce the Maximum Principle and its relation to the classical variational approach used to derive the Euler Equations in continuous time. Finally, we will show how to obtain the continuous time version of the neoclassical growth model from its discrete time version counterpart.[22] We will restrict ourselves to studying deterministic (as opposed to stochastic) problems.

8.1 Discrete time

A *state formulation* of a dynamic problem is a tuple (X, Γ, F, β), where

- X is the set of states. We typically use x to denote the current state and y to denote the next-period state.
- $\Gamma : X \to X$ is the correspondence describing the feasibility constraints. That is, for each $x \in X$, $\Gamma(x)$ gives the set of feasible values for the state variable next period if the current state is x.
- $F : \text{gr}(\Gamma) \to \mathbb{R}$ is the period-return function, where $\text{gr}(\Gamma) = \{(x, y) : X \times X : y \in \Gamma(x)\}$ is the graph of Γ.
- $\beta \in (0, 1)$ is the discount factor.

The *sequence problem* is then defined as follows:

$$\begin{aligned} V^*(x_0) \coloneqq \max_{(x_{t+1})_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}) \\ \text{s.t.} \quad & x_{t+1} \in \Gamma(x_t), \ \forall t \geq 0, \\ & x_0 \text{ given.} \end{aligned}$$

The sequence $(x_{t+1})_{t=0}^{\infty}$ is called a *path* and our goal is to find a path that maximises the objective function F subject to the path satisfying constraints expressed by Γ. Note that we have already embedded the assumption that x_{t+1} depends directly only on x_t and not on, e.g., x_{t-1}, x_{t-2}, ...

A related notation, called *control-state formulation*, distinguishes between controls, u_t, and states, x_t. In this notation, the sequence problem is described by a tuple (X, U, h, g, β), where

- U is the set of feasible controls
- h is the period-return function,
- g is the law of motion of the state.

[22] The material in the remainder of this part on dynamic optimisation is almost entirely from Professor Fernando Alvarez's class, Theory of Income I.

The sequence problem is then defined as

$$\begin{aligned} V^*(x_0) := \max_{(u_t)_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t h(x_t, u_t) \\ \text{s.t.} \quad & x_{t+1} = g(x_t, u_t), \\ & u_t \in U, \\ & x_0 \text{ given.} \end{aligned}$$

In this set up, rather than choosing the state directly, the agent chooses the control variable u_t, which, together with the current state (x_t), determines the next period's state (x_{t+1}).

Exercise 19. Show that the control-state and state formulations are equivalent.

8.1.1 Euler equations and Transversality condition

Assume that $X \in \mathbb{R}^m$, F is continuously differentiable and $\beta \in (0,1)$. Say that a path $(x_{t+1})_{t=0}^{\infty}$ satisfies

- the *Euler equations* (EE) if

$$F_y(x_t, x_{t+1}) + \beta F_x(x_{t+1}, x_{t+2}) = 0, \ \forall t \geq 0,$$

  where F_x and F_y are the partial derivative of F with respect to x and y respectively. When x is a vector (i.e., $m > 1$), then the Euler equations must hold for each dimension of x.

- the Transversality Condition (TC) if

$$\lim_{t \to \infty} \beta^t F_x(x_t, x_{t+1}) \cdot x_t = 0.$$

  where $\cdot$ is the inner product.

Remark 17. To remember where the Euler equations comes from, note that the first-order condition of the sequence problem with respect to x_{t+1} is

$$\beta^t F_y(x_t, x_{t+1}) + \beta^{t+1} F_x(x_{t+1}, x_{t+2}) = 0.$$

Dividing through by β^t gives the Euler equation.

8.1.2 Optimal path

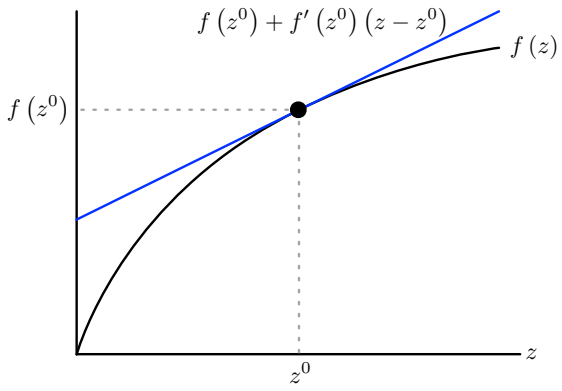
We now show that EE and TC are sufficient for optimality if the problem is convex.

Proposition 93. *Assume that F is concave in (x, y), $F_x(x_t^*, x_{t+1}^*) \geq 0$ and $X = \mathbb{R}_+^m$. Then, if $(x_{t+1}^*)_{t=0}^{\infty}$ satisfies EE and TC, the path $(x_{t+1}^*)_{t=0}^{\infty}$ is optimal.*

Proof. We use the fact that, if f is concave the tangency line at any given point always lies above f; i.e.

$$f(z) \leq f(z^0) + f'(z^0)(z - z^0), \ \forall z. \tag{8.1}$$

See also the figure below.



Now, take an arbitrary path that has the same initial condition as the optimal path: i.e. take $(x_{t+1})_{t=0}^{\infty}$ with $x_0 = x_0^*$. By the assumption that $X = \mathbb{R}_+^m$, $x_{t+1} \geq 0$ for all t. We wish to show that $F(x_t^*, x_{t+1}^*)$ is greater than $F(x_t, x_{t+1})$ across all periods; i.e. we want to show that:

$$\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[F\left(x_t, x_{t+1}\right) - F\left(x_t^*, x_{t+1}^*\right)\right] \leq 0.$$

Note that we can rearrange (8.1),

$$f\left(z\right) - f\left(z^0\right) \leq f'\left(z^0\right)\left(z - z^0\right).$$

Then, letting $z^0 = x^*$,

$$F\left(x_t, x_{t+1}\right) - F\left(x_t^*, x_{t+1}^*\right) \leq F_x\left(x_t^*, x_{t+1}^*\right)\left(x_t - x_t^*\right) + F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right).$$

That is,

$$\begin{aligned}
&\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[F\left(x_t, x_{t+1}\right) - F\left(x_t^*, x_{t+1}^*\right)\right] \\
&\leq \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\left(x_t - x_t^*\right) + F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right)\right].
\end{aligned}$$

Expanding the summation, we have that

$$\begin{aligned}
&\sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\left(x_t - x_t^*\right) + F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right)\right] \\
&= F_x\left(x_0^*, x_1^*\right)\left(x_0 - x_0^*\right) + F_y\left(x_0^*, x_1^*\right)\left(x_1 - x_1^*\right) \\
&\quad + \beta\left(F_x\left(x_1^*, x_2^*\right)\left(x_1 - x_1^*\right) + F_y\left(x_1^*, x_2^*\right)\left(x_2 - x_2^*\right)\right) \\
&\quad + \beta^2\left(F_x\left(x_2^*, x_3^*\right)\left(x_2 - x_2^*\right) + F_y\left(x_2^*, x_3^*\right)\left(x_3 - x_3^*\right)\right) \\
&\quad + \cdots \\
&\quad + \beta^t\left(F_x\left(x_t^*, x_{t+1}^*\right)\left(x_t - x_t^*\right) + F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right)\right) \\
&\quad + \beta^{t+1}\left(F_x\left(x_{t+1}^*, x_{t+2}^*\right)\left(x_{t+1} - x_{t+1}^*\right) + F_y\left(x_{t+1}^*, x_{t+2}^*\right)\left(x_{t+2} - x_{t+2}^*\right)\right) \\
&\quad + \cdots \\
&\quad + \beta^T\left(F_x\left(x_T^*, x_{T+1}^*\right)\left(x_T - x_T^*\right) + F_y\left(x_T^*, x_{T+1}^*\right)\left(x_{T+1} - x_{T+1}^*\right)\right).
\end{aligned}$$

By assumption, $x_0 = x_0^*$ so that the first term is zero. Rewriting the expression gives

$$\begin{aligned}
&\sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\left(x_t - x_t^*\right) + F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right)\right] \\
&= \left[F_y\left(x_0^*, x_1^*\right)\left(x_1 - x_1^*\right) + \beta\left(F_x\left(x_1^*, x_2^*\right)\left(x_1 - x_1^*\right)\right)\right] \\
&\quad + \beta\left[F_y\left(x_1^*, x_2^*\right)\left(x_2 - x_2^*\right) + \beta F_x\left(x_2^*, x_3^*\right)\left(x_2 - x_2^*\right)\right] \\
&\quad + \cdots \\
&\quad + \beta^t\left[F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right) + \beta F_x\left(x_{t+1}^*, x_{t+2}^*\right)\left(x_{t+1} - x_{t+1}^*\right)\right] \\
&\quad + \cdots \\
&\quad + \beta^T F_y\left(x_T^*, x_{T+1}^*\right)\left(x_{T+1} - x_{T+1}^*\right),
\end{aligned}$$

where we realise that the terms inside the square brackets are all zero by the EE. Thus, the expression above simplifies to

$$\begin{aligned}
&\sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\left(x_t - x_t^*\right) + F_y\left(x_t^*, x_{t+1}^*\right)\left(x_{t+1} - x_{t+1}^*\right)\right] \\
&= \beta^T F_y\left(x_T^*, x_{T+1}^*\right)\left(x_{T+1} - x_{T+1}^*\right).
\end{aligned}$$

From EE, we know that $F_y\left(x_t, x_{t+1}\right) = -\beta F_x\left(x_{t+1}, x_{t+2}\right)$, hence we can write

$$\begin{aligned}
&\beta^T F_y\left(x_T^*, x_{T+1}^*\right)\left(x_{T+1} - x_{T+1}^*\right) \\
&= -\beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right)\left(x_{T+1} - x_{T+1}^*\right).
\end{aligned}$$

By assumption, we know that $x_{T+1} \geq 0$ and $F_x\left(x_{T+1}^*, x_{T+2}^*\right) \geq 0$ so that

$$\begin{aligned}
&\beta^T F_y\left(x_T^*, x_{T+1}^*\right)\left(x_{T+1} - x_{T+1}^*\right) \\
=& -\beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right)\left(x_{T+1} - x_{T+1}^*\right) \\
=& \underbrace{-\beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right) x_{T+1}}_{\leq 0} + \beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right) x_{T+1}^* \\
\leq& \beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right) x_{T+1}^*.
\end{aligned}$$

By assumption, TC holds so that

$$\begin{aligned}
&\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[F\left(x_t, x_{t+1}\right) - F\left(x_t^*, x_{t+1}^*\right)\right] \\
&\leq \lim_{T\to\infty} \beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right) x_{T+1}^* = 0.
\end{aligned}$$

■

We now prove that EE and TC are necessary conditions for the path to be optimal

Proposition 94. *Assume that F is continuously differentiable and $(x_{t+1}^*)_{t=0}^{\infty}$ is optimal. Then, $(x_{t+1}^*)_{t=0}^{\infty}$ satisfies EE and TC.*

Proof. We will consider adding perturbations to the optimal path $(x_t^*)_{t=1}^{\infty}$, denoted by ε. Let

$$x_t(\alpha, \varepsilon) = x_t^* + \alpha\varepsilon_t, \ \forall t \geq 0$$

for $\alpha \in \mathbb{R}$ and $\varepsilon = (\varepsilon_t)_{t=0}^{\infty}$ with $\varepsilon_t \in \mathbb{R}^m$ and $\varepsilon_0 = 0$ (again, we must start from the same point).

Since $(x_{t+1}^*)_{t=0}^{\infty}$ is optimal, it must be that

$$\begin{aligned} V^*(x_0) = v(0) &\coloneqq \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t F\left(x_t(0,\varepsilon), x_{t+1}(0,\varepsilon)\right) \\ &\geq v(\alpha) \coloneqq \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t F\left(x_t(\alpha,\varepsilon), x_{t+1}(\alpha,\varepsilon)\right) \end{aligned}$$

for any α, ε such that $x_{t+1}(\alpha,\varepsilon) \in \Gamma(x_t(\alpha,\varepsilon)),\ \forall t \geq 0$ (i.e. perturbed path must still be feasible).

Since $\alpha = 0$ maximises v, if v is differentiable, it must be that

$$\frac{\partial v(0)}{\partial \alpha} = 0.$$

Assuming that the limits involved in the derivative (with respect to α) and in the summation (with respect to T) can be exchanged, we obtain that (since $\partial x_t(\alpha,\varepsilon)/\partial\alpha = \varepsilon_t$),

$$\frac{\partial v(0)}{\partial \alpha} = \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\varepsilon_t + F_y\left(x_t^*, x_{t+1}^*\right)\varepsilon_{t+1}\right].$$

Consider the summation separately,

$$\begin{aligned} &\sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\varepsilon_t + F_y\left(x_t^*, x_{t+1}^*\right)\varepsilon_{t+1}\right] \\ &= F_x(x_0^*, x_1^*)\varepsilon_0 + F_y(x_0^*, x_1^*)\varepsilon_1 \\ &\quad + \beta\left[F_x(x_1^*, x_2^*)\varepsilon_1 + F_y(x_1^*, x_2^*)\varepsilon_2\right] \\ &\quad + \beta^2\left[F_x(x_2^*, x_3^*)\varepsilon_2 + F_y(x_2^*, x_3^*)\varepsilon_3\right] \\ &\quad + \cdots \\ &\quad + \beta^t\left[F_x\left(x_t^*, x_{t+1}^*\right)\varepsilon_t + F_y\left(x_t^*, x_{t+1}^*\right)\varepsilon_{t+1}\right] \\ &\quad + \beta^{t+1}\left[F_x\left(x_{t+1}^*, x_{t+2}^*\right)\varepsilon_{t+1} + F_y\left(x_{t+1}^*, x_{t+2}^*\right)\varepsilon_{t+2}\right] \\ &\quad + \cdots \\ &\quad + \beta^{T-1}\left[F_x\left(x_{T-1}^*, x_T^*\right)\varepsilon_{T-1} + F_y\left(x_{T-1}^*, x_T^*\right)\varepsilon_T\right] \\ &\quad + \beta^T\left[F_x\left(x_T^*, x_{T+1}^*\right)\varepsilon_T + F_y\left(x_T^*, x_{T+1}^*\right)\varepsilon_{T+1}\right]. \end{aligned}$$

By the assumption that $\varepsilon_0 = 0$, the first term is zero. Then, we can write

$$\begin{aligned} \frac{\partial v(0)}{\partial \alpha} &= \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t \left[F_x\left(x_t^*, x_{t+1}^*\right)\varepsilon_t + F_y\left(x_t^*, x_{t+1}^*\right)\varepsilon_{t+1}\right] \\ &= \lim_{T\to\infty} \sum_{t=0}^{T-1} \beta^t \left(F_y\left(x_t^*, x_{t+1}^*\right) + \beta F_x\left(x_{t+1}^*, x_{t+2}^*\right)\right)\varepsilon_{t+1} + \lim_{T\to\infty} \beta^T F_y\left(x_T^*, x_{T+1}^*\right)\varepsilon_{T+1}. \end{aligned}$$

Consider the case where $\varepsilon_s = 0$ for all s except at time $t+1$. In this case, $x_{t+1}(\alpha,\varepsilon)$ will be feasible if $(x_{t+1}^*, x_t^*) \in \text{int}(\text{Gr}(\Gamma))$ (if it was on the boundary, then perturbing would lead to $x_{t+1}(\alpha,\varepsilon)$ that is not feasible). Also, assume that v is differentiable and the limits can be interchanged. Then, it must be that

$$\frac{\partial v(0)}{\partial \alpha} = \left[F_y\left(x_t^*, x_{t+1}^*\right) + \beta F_x\left(x_{t+1}^*, x_{t+2}^*\right)\right]\varepsilon_{t+1} = 0.$$

Since this must hold for any ε_{t+1} in the neighbourhood of 0 such that $x_{t+1}(\alpha,\varepsilon)$ is feasible, it must

be that the term inside the square brackets sum to zero; i.e.

$$F_y\left(x_t^*, x_{t+1}^*\right) + \beta F_x\left(x_{t+1}^*, x_{t+2}^*\right) = 0.$$

Since this must be true for any $t+1$, we obtain the EE and that

$$0 = \frac{\partial v\left(0\right)}{\partial \alpha} = \lim_{T\to\infty} \beta^T F_y\left(x_T^*, x_{T+1}^*\right) \varepsilon_{T+1}.$$

Using the EE that we've already established, we know $F_y\left(x_T^*, x_{T+1}^*\right) = -\beta F_x\left(x_{T+1}^*, x_{T+2}^*\right)$ so that

$$0 = \frac{\partial v\left(0\right)}{\partial \alpha} = -\lim_{T\to\infty} \beta^{T+1} F_x\left(x_{T+1}^*, x_{T+2}^*\right) \varepsilon_{T+1}.$$

Finally, if $\varepsilon_{T+1} = -x_{T+1}^*$ is feasible then

$$0 = \frac{\partial v\left(0\right)}{\partial \alpha} = \lim_{T\to\infty} \beta^T F_x\left(x_T^*, x_{T+1}^*\right) x_T^*.$$

That is, TC must hold. ■

8.1.3 Steady state

We say that $\bar{x} \in X$ is a *steady state* if it is a solution to

$$F_y\left(\bar{x}, \bar{x}\right) + \beta F_x\left(\bar{x}, \bar{x}\right) = 0.$$

Concavity of F does not imply that steady states are unique. Concavity ensures unique solution *for any given* initial condition. However, there may be many steady states depending on where you start.

Exercise 20. For what kind of problems is $x_{t+1} = \bar{x}$ for $t \geq 0$ optimal if $x_0 = \bar{x}$?

8.1.4 EE as a second-order difference equation

EE is, in fact, a second-order difference equation; define $x_{t+2} = \psi(x_{t+1}, x_t)$ so that EE can be written as

$$F_y\left(x_t, x_{t+1}\right) + \beta F_x\left(x_{t+1}, \psi\left(x_{t+1}, x_t\right)\right) = 0.$$

To be able to solve for ψ, we need F_x to change with ψ (so that the problem is dynamic). For example, if $f(x,y) = h(x) + g(x)$, then the initial condition on y does not affect x so that the problem is not dynamic.

Exercise 21. Assume that F is C^2. What condition will suffice to uniquely define ψ?

Remark 18. *[Shooting algorithm]*Given initial condition $x_0 \in X$, select x_1 arbitrary. Generate a sequence $(x_t)_t$ using $x_{t+1} = \psi(x_{t+1}, x_t)$ for all $t \geq 2$. Compute if the limit of this sequence satisfies TC for the arbitrary choice of x_1. If not, try a different one. If the dynamic system has a unique solution and the system is convergent, this method will give us the optimal path (since starting from any initial condition, we will eventually converge to the steady state if we are on the saddle path.)

8.2 Continuous Time

The problem now is to choose the derivative of the state with respect to time, $\dot{x}(t)$, for each period. The elements of a dynamic programming problem is described by a tuple (X, Γ, F, β), where

- X is the set of states;
- $\Gamma : X \to X$ is the correspondence describing the feasibility constraints, which gives the set of feasible values for the state variable next period if the current state is x.
- $F(x, y) : \text{gr}(\Gamma) \to \mathbb{R}$ is the period-return function;
- $\rho \in (0, 1)$ is the discount rate.

The sequence problem can then be defined as follows:

$$\begin{aligned} V^*(x_0) \coloneqq \max_{\dot{x}(t)} \quad & \lim_{T\to\infty} \int_0^T e^{-\rho t} F(x(t), \dot{x}(t))\, \mathrm{d}t \\ \text{s.t.} \quad & \dot{x}(t) \in \Gamma(x(t)), \forall t \geq 0, \\ & x(0) \text{ given.} \end{aligned}$$

8.2.1 Euler equations and Transversality condition

A path $(\dot{x}(t))_{t=0}^{\infty}$ satisfies

- the Euler equations (EE) if

$$\begin{aligned} & F_x(x(t), \dot{x}(t)) + \rho F_y(x(t), \dot{x}(t)) \\ & = F_{yx}(x(t), \dot{x}(t))\, \dot{x}(t) + F_{yy}(x(t), \dot{x}(t))\, \ddot{x}(t), \ \forall t \geq 0. \end{aligned} \tag{8.2}$$

  Loosely speaking, the continuous time counterparts to x_t, x_{t+1}, x_{t+1} are $x(t)$, $\dot{x}(t)$ and $\ddot{x}(t)$.

- the Transversality Condition (TC) if

$$\lim_{T\to\infty} e^{-\rho T} F_y(x(T), \dot{x}(T)) \cdot x(T) = 0.$$

8.2.2 Optimal path

Proposition 95. *Assume that $F(x, \dot{x})$ is concave in $(x, \dot{x})$, $F_{\dot{x}} \leq 0$, $\text{Gr}(\Gamma)$ is convex,*[23] *that the optimal path $x^*(t)$ is interior, and $X = \mathbb{R}^m_+$. Then, if $(\dot{x}^*(t))_{t=0}^{\infty}$ satisfies EE and TC, the path $(\dot{x}^*(t))_{t=0}^{\infty}$ is optimal.*

Proof. As in the discrete case, we wish to show that

$$\lim_{T\to\infty} \int_0^T e^{-\rho t} \left(F(x(t), \dot{x}(t)) - F(x^*(t), \dot{x}^*(t))\right) \mathrm{d}t \leq 0.$$

As before, we may write

$$\begin{aligned} F(x(t), \dot{x}(t)) - F(x^*(t), \dot{x}^*(t)) \leq\ & F_x(x^*(t), \dot{x}^*(t))\,(x(t) - x^*(t)) \\ & + F_y(x^*(t), \dot{x}^*(t))\,(\dot{x}(t) - \dot{x}^*(t)). \end{aligned}$$

[23] Convexity is required to ensure existence of a solution.

Hence,

$$\begin{aligned}
&\lim_{T\to\infty}\int_0^T e^{-\rho t}\left(F\left(x\left(t\right),\dot{x}\left(t\right)\right)-F\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\right)\mathrm{d}t\\
&\leq \lim_{T\to\infty}\int_0^T e^{-\rho t}F_x\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\left(x\left(t\right)-x^*\left(t\right)\right)\mathrm{d}t\\
&\qquad+\lim_{T\to\infty}\int_0^T e^{-\rho t}F_y\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\left(\dot{x}\left(t\right)-\dot{x}^*\left(t\right)\right)\mathrm{d}t.
\end{aligned}$$

Using integration by parts on the second integral:

$$\begin{aligned}
&\int_0^T e^{-\rho t}F_y\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\left(\dot{x}\left(t\right)-\dot{x}^*\left(t\right)\right)\\
&=\left[e^{-\rho t}F_y\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\left(x\left(t\right)-x^*\left(t\right)\right)\right]_0^T-\int_0^T\mathcal{X}\left(x\left(t\right)-x^*\left(t\right)\right)\mathrm{d}t,\\
&=e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)\left(x\left(T\right)-x^*\left(T\right)\right)-\int_0^T\mathcal{X}\left(x\left(t\right)-x^*\left(t\right)\right)\mathrm{d}t,
\end{aligned}$$

where we use the fact that $x\left(0\right)=x^*\left(0\right)$ and

$$\begin{aligned}
\mathcal{X}&=\frac{\mathrm{d}}{\mathrm{d}t}\left[e^{-\rho t}F_y\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\right]\\
&=-\rho e^{-\rho t}F_y\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\\
&\qquad+e^{-\rho t}F_{yx}\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\dot{x}^*\left(t\right)+e^{-\rho t}F_{yy}\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\ddot{x}^*\left(t\right)\\
&=e^{-\rho t}\left[-\rho F_y\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)+F_{yx}\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\dot{x}^*\left(t\right)+F_{yy}\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\ddot{x}^*\left(t\right)\right]
\end{aligned}$$

Suppressing some arguments,

$$\begin{aligned}
&\lim_{T\to\infty}\int_0^T e^{-\rho t}\left(F_x^*\left(x-x^*\right)+F_y^*\left(\dot{x}-\dot{x}^*\right)\right)\mathrm{d}t\\
&\leq\int_0^\infty e^{-\rho t}F_x^*\left(x-x^*\right)\mathrm{d}t-\int_0^\infty e^{-\rho t}\left[-\rho F_y^*+F_{yx}\dot{x}^*+F_{yy}\ddot{x}^*\right]\left(x-x^*\right)\mathrm{d}t\\
&\qquad+\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)\left(x\left(T\right)-x^*\left(T\right)\right)\\
&=\int_0^\infty e^{-\rho t}\left[\underbrace{F_x^*+\rho F_y^*-F_{yx}\dot{x}^*-F_{yy}\ddot{x}^*}_{=0\because EE}\right]\left(x-x^*\right)\mathrm{d}t+\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)\left(x\left(T\right)-x^*\left(T\right)\right)\\
&=\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)\left(x\left(T\right)-x^*\left(T\right)\right)\\
&=\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)x\left(T\right)-\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)x^*\left(T\right).
\end{aligned}$$

Since $F_y\leq 0$ and $x\left(t\right)\geq 0$, then

$$\begin{aligned}
&\lim_{T\to\infty}\int_0^T e^{-\rho t}\left(F_x^*\left(x-x^*\right)+F_y^*\left(\dot{x}-\dot{x}^*\right)\right)\mathrm{d}t\\
&\leq\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)x\left(T\right)-\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)x^*\left(T\right)\\
&\leq-\lim_{T\to\infty}e^{-\rho T}F_y\left(x^*\left(T\right),\dot{x}^*\left(T\right)\right)x^*\left(T\right)=0,
\end{aligned}$$

where the last equality follows from the Transversality Condition. Thus,

$$\lim_{T\to\infty}\int_0^T e^{-\rho t}\left(F\left(x\left(t\right),\dot{x}\left(t\right)\right)-F\left(x^*\left(t\right),\dot{x}^*\left(t\right)\right)\right)\mathrm{d}t\leq 0. \qquad \blacksquare$$

As in the discrete time case, EE and TC are necessary conditions for optimality.

Proposition 96. *Assume that F is continuously differentiable and that the optimal path is interior; i.e.*

$$\left(x\left(t\right),\dot{x}\left(t\right)\right)\in\operatorname{int}\left(\operatorname{gr}\left(\Gamma\right)\right),\forall t\geq 0.$$

Then, the optimal path $(\dot{x}(t))_{t=0}^{\infty}$ satisfies EE and TC.

Proof. Consider the variation path

$$x\left(\alpha,\varepsilon\right)\left(t\right)=x\left(t\right)+\alpha\varepsilon\left(t\right),$$

where $\alpha\in\mathbb{R}$ and $\varepsilon(t)$ is a differentiable function from $\mathbb{R}_+$ to $\mathbb{R}^m$ with $\varepsilon(0)=0$. Define the value of the variational path as:

$$\begin{aligned}v\left(\alpha\right)&\coloneqq\lim_{T\to\infty}\int_0^T e^{-\rho t}F\left(x\left(\alpha,\varepsilon\right)\left(t\right),\dot{x}\left(\alpha,\varepsilon\right)\left(t\right)\right)\mathrm{d}t\\&=\lim_{T\to\infty}\int_0^T e^{-\rho t}F\left(x\left(t\right)+\alpha\varepsilon\left(t\right),\dot{x}\left(t\right)+\alpha\dot{\varepsilon}\left(t\right)\right)\mathrm{d}t.\end{aligned}$$

If the variation path is feasible—i.e. $\dot{x}\left(\alpha,\varepsilon\right)\left(t\right)\in\Gamma\left(x\left(\alpha,\varepsilon\right)\left(t\right)\right)$ for all $t\geq 0$, since $\left(x\left(t\right),\dot{x}\left(t\right)\right)$ is the optimal path, it must be that

$$v\left(0\right)\geq v\left(\alpha\right).$$

Assuming that v is differentiable and that we can interchange derivative of the integral as the integral of the derivative:

$$\begin{aligned}\frac{\partial v\left(0\right)}{\partial\alpha}=&\lim_{T\to\infty}\int_0^T e^{-\rho t}F_x\left(x\left(t\right),\dot{x}\left(t\right)\right)\varepsilon\left(t\right)\mathrm{d}t\\&+\lim_{T\to\infty}\int_0^T e^{-\rho t}F_y\left(x\left(t\right),\dot{x}\left(t\right)\right)\dot{\varepsilon}\left(t\right)\mathrm{d}t.\end{aligned}$$

To find the impact of the variational path ε for each t, we use integration by parts on the second term:

$$\begin{aligned}&\int_0^T e^{-\rho t}F_y\left(x\left(t\right),\dot{x}\left(t\right)\right)\dot{\varepsilon}\left(t\right)\mathrm{d}t\\=&\left[e^{-\rho t}F_y\left(x\left(t\right),\dot{x}\left(t\right)\right)\varepsilon\left(t\right)\right]_0^T-\int_0^T\mathcal{X}\varepsilon\left(t\right)\mathrm{d}t\\=&e^{-\rho T}F_y\left(x\left(T\right),\dot{x}\left(T\right)\right)\varepsilon\left(T\right)-\int_0^T\mathcal{X}\varepsilon\left(t\right)\mathrm{d}t\end{aligned}$$

where we use the fact that $\varepsilon\left(0\right)=0$ and (assuming $x\left(t\right)$ is C^2),

$$\begin{aligned}\mathcal{X}=&-\rho e^{-\rho t}F_y\left(x\left(t\right),\dot{x}\left(t\right)\right)\\&+e^{-\rho t}F_{yx}\left(x\left(t\right),\dot{x}\left(t\right)\right)\dot{x}\left(t\right)+e^{-\rho t}F_{yy}\left(x\left(t\right),\dot{x}\left(t\right)\right)\ddot{x}\left(t\right).\end{aligned}$$

Omitting some arguments, we therefore have that

$$0 = \frac{\partial v(0)}{\partial \alpha} = \int_0^\infty e^{-\rho t} F_x \varepsilon dt - \int_0^T e^{-\rho t} \left[-\rho F_y + F_{yx}\dot{x} + F_{yy}\ddot{x}\right] \varepsilon \mathrm{d}t + \lim_{T\to\infty} e^{-\rho T} F_y\left(x(T), \dot{x}(T)\right) \varepsilon(T).$$

As in the discrete time case, take ε such that $\varepsilon(t) \neq 0$ and zero otherwise. Then, the expression above becomes

$$0 = e^{-\rho t}\left[F_x + \rho F_y - F_{yx}\dot{x} - F_{yy}\ddot{x}\right]\varepsilon.$$

Since this has to hold for any feasible values of $\varepsilon(t)$, it must be that the term inside the brackets is zero. Rearranging and writing the arguments in full, we obtain the EE:

$$F_x\left(x(t), \dot{x}(t)\right) + \rho F_y\left(x(t), \dot{x}(t)\right) = F_{yx}\left(x(t), \dot{x}(t)\right)\dot{x}(t) - F_{yy}\left(x(t), \dot{x}(t)\right)\ddot{x}(t).$$

Note that the proof is heuristic since we assumed that ε is differentiable but, for the Euler equations, we use a function ε that was discontinuous. A rigorous proof needs to approximate this discontinuous case using a smooth function.

Take a sequence x that satisfies EE then,

$$0 = \frac{\partial v(0)}{\partial \alpha} = \lim_{T\to\infty} e^{-\rho T} F_y\left(x(T), \dot{x}(T)\right)\varepsilon(T).$$

If $\varepsilon(T) = -x(T)$ is feasible, we obtain the TC:

$$0 = \lim_{T\to\infty} e^{-\rho T} F_y\left(x(T), \dot{x}(T)\right) x(T). \qquad \blacksquare$$

8.2.3 Steady state

In the steady state $\dot{x}(t) = \ddot{x}(t) = 0$. Substituting this into the EE characterises the steady state $\bar{x}$:

$$\begin{aligned} F_x(\bar{x}, 0) + \rho F_y(\bar{x}, 0) &= F_{yx}(\bar{x}, 0)\, 0 + F_{yy}(\bar{x}, 0)\, 0 \\ \Rightarrow F_x(\bar{x}, 0) &= -\rho F_y(\bar{x}, 0). \end{aligned}$$

8.2.4 The Maximum Principle: Hamiltonian

We use the control-state formulation. For this, we have the instantaneous return function h that depends on the state vector $x \in X \subseteq \mathbb{R}^m$ and a control vector $u \in U \subseteq \mathbb{R}^n$ (m need not equal n). The problem is

$$\begin{aligned} V^*(x_0) \coloneqq \max_{u(t)} \quad & \lim_{T\to\infty} \int_0^T e^{-\rho t} h\left(x(t), u(t)\right) \mathrm{d}t \\ \text{s.t.} \quad & \dot{x}(t) = g\left(x(t), u(t)\right), \ \forall t \geq 0 \\ & u(t) \in U, \ \forall t \geq 0, \\ & x(0) \text{ given.} \end{aligned}$$

We will study a procedure to obtain necessary and sufficient conditions for an optimum. This requires some regularity conditions.

Let $\boldsymbol{\lambda}$ be a vector on $\mathbb{R}^m$ of *co-state* variable and H be the *current-value Hamiltonian* function

defined as

$$H(x,u,\lambda) \coloneqq h(x,u) + \lambda g(x,u).$$

The following conditions are necessary (under regularity assumptions) and sufficient (under regularity and convexity assumptions) for a path of x and u to be optimal:

$$H_u(x(t), u(t), \lambda(t)) = 0, \tag{8.3}$$

$$\rho\lambda(t) - H_x(x(t), u(t), \lambda(t)) = \dot{\lambda}(t), \tag{8.4}$$

$$g(x(t), u(t)) = \dot{x}(t), \tag{8.5}$$

for all $t \geq 0$.

The state variable(s), x, has an initial value of x_0 and the co-state variable(s), $\lambda(t)$, have a boundary condition—the Transversality Condition—given by

$$\lim_{T\to\infty} e^{-\rho T}\lambda(T)\, x(T) = 0.$$

The initial value of the co-state variable, $\lambda(0)$, is not predetermined and it has to be solved as part of the system.

The interpretation of co-state is that $e^{-\rho t}\lambda(t)$ is the marginal value at time zero of an infinitesimal increase in the state x at time t.

The first condition (8.3) says that the derivative of the Hamiltonian with respect to the control is zero—this gives the optimal choice of u (one for each control variable). To interpret the second condition (8.4), recall that

$$rP = D + \dot{P},$$

can be interpreted as saying that P is the present value of D with interest rate r (D is the dividends, $\dot{P}$ is capital gain). Here, we see that $\lambda(t)$ is the present value of H_x given discount factor ρ. The condition gives the (shadow) marginal value of a unit of x. Finally, the last condition (8.5) is the feasibility condition.

Heuristic proof Let us form the Lagrangian, using $e^{-\rho t}\lambda(t)$ for the multiplier of $\dot{x}(t) = g(x(t), u(t))$:

$$\mathcal{L}(x,u,\lambda) = \lim_{T\to\infty}\left(\int_0^T e^{-\rho t}h(x(t), u(t))\,\mathrm{d}t + \int_0^T e^{-\rho t}\lambda(t)\left[g(x(t), u(t)) - \dot{x}(t)\right]\mathrm{d}t\right).$$

Consider the problem of maximising $\mathcal{L}$ with respect to x and u, and minimise with respect to λ.[24] First, consider the following term, and use integration by parts to obtain:

$$\int_0^T e^{-\rho t}\lambda(t)\,\dot{x}(t)\,\mathrm{d}t = \left[e^{-\rho t}\lambda(t)\,x(t)\right]_0^T - \int_0^T \left[-\rho e^{-\rho t}\lambda(t) + e^{-\rho t}\dot{\lambda}(t)\right]x(t)\,\mathrm{d}t.$$

We therefore have

$$\begin{aligned}\mathcal{L}(x,u,\lambda) = &\lim_{T\to\infty}\int_0^T e^{-\rho t}h(x(t), u(t))\,\mathrm{d}t \\ &+ \lim_{T\to\infty}\int_0^T e^{-\rho t}\left[\lambda(t)\,g(x(t), u(t)) - \rho\lambda(t)\,x(t) + \dot{\lambda}(t)\,x(t)\right]\mathrm{d}t \\ &- \lim_{T\to\infty}\left[e^{-\rho t}\lambda(t)\,x(t)\right]_0^T.\end{aligned}$$

[24] In ECON6701, we will see that the Lagrangian method is a max-min problem.

Since $\mathcal{L}(x,u,\lambda)$ has to be maximised by x, the first-order conditions with respect to $x(t)$ gives

$$\frac{\mathrm{d}\mathcal{L}\left(x,u,\lambda\right)}{\mathrm{d}x\left(t\right)}=e^{-\rho t}\left[h_{x}\left(x\left(t\right),u\left(t\right)\right)+\lambda\left(t\right)g_{x}\left(x\left(t\right),u\left(t\right)\right)-\rho\lambda\left(t\right)+\dot{\lambda}\left(t\right)\right]=0.$$

Since $e^{-\rho t}>0$, it follows that

$$\dot{\lambda}\left(t\right)=\rho\lambda\left(t\right)-h_{x}\left(x\left(t\right),u\left(t\right)\right)-\lambda\left(t\right)g_{x}\left(x\left(t\right),u\left(t\right)\right).$$

Note that

$$\begin{aligned}\frac{\mathrm{d}H\left(x,u,\lambda\right)}{\mathrm{d}x\left(t\right)} &= H_{x}\left(x\left(t\right),u\left(t\right),\lambda\left(t\right)\right)\\ &= h_{x}\left(x\left(t\right),u\left(t\right)\right)+\lambda\left(t\right)g_{x}\left(x\left(t\right),u\left(t\right)\right)\end{aligned}$$

so that we can rewrite $\dot{\lambda}(t)$ as

$$\dot{\lambda}\left(t\right)=\rho\lambda\left(t\right)-H_{x}\left(x\left(t\right),u\left(t\right),\lambda\left(t\right)\right). \tag{8.6}$$

The first-order condition with respect to $u\left(t\right)$ is

$$\frac{\mathrm{d}\mathcal{L}\left(x,u,\lambda\right)}{\mathrm{d}u\left(t\right)}=e^{-\rho t}\left[h_{u}\left(x\left(t\right),u\left(t\right)\right)+\lambda\left(t\right)g_{u}\left(x\left(t\right),u\left(t\right)\right)\right]=0.$$

Since

$$\begin{aligned}\frac{\mathrm{d}H\left(x,u,\lambda\right)}{\mathrm{d}u\left(t\right)} &= H_{u}\left(x\left(t\right),u\left(t\right),\lambda\left(t\right)\right)\\ &= h_{u}\left(x\left(t\right),u\left(t\right)\right)+\lambda\left(t\right)g_{u}\left(x\left(t\right),u\left(t\right)\right),\end{aligned}$$

we therefore have that

$$H_{u}\left(x\left(t\right),u\left(t\right),\lambda\left(t\right)\right)=0. \tag{8.7}$$

Relation to the EE analysis To see the relation with the classical EE analysis, consider the special case:

$$\begin{aligned}u &= \dot{x},\\ F\left(x,u\right) &= h\left(x,u\right),\\ g\left(x,u\right) &= u.\end{aligned}$$

In this case,

$$H\left(x,u,\lambda\right)=F\left(x,\dot{x}\right)+\lambda\dot{x}$$

so that (8.7) becomes

$$\begin{aligned}H_{u}\left(x,u,\lambda\right)=0 &\Rightarrow F_{y}\left(x\left(t\right),\dot{x}\left(t\right)\right)+\lambda\left(t\right)=0\\ &\Rightarrow \lambda\left(t\right)=-F_{y}\left(x\left(t\right),\dot{x}\left(t\right)\right)\end{aligned} \tag{8.8}$$

and

$$H_{x}\left(x,u,\lambda\right)=F_{x}\left(x\left(t\right),\dot{x}\left(t\right)\right).$$

Therefore, (8.6) becomes

$$\dot{\lambda}\left(t\right)=\rho\lambda\left(t\right)-F_{x}\left(x\left(t\right),\dot{x}\left(t\right)\right).$$

Combining the two we get that

$$\dot{\lambda}(t) = -\rho F_y(x(t), \dot{x}(t)) - F_x(x(t), \dot{x}(t)).$$

Thus, $-\dot{\lambda}(t)$ is the right-hand side of the EE in (8.2). To confirm, differentiating (8.8) with respect to t gives

$$\dot{\lambda}(t) = -\frac{\mathrm{d}}{\mathrm{d}t} F_y(x(t), \dot{x}(t)) = -F_{yx}(x(t), \dot{x}(t))\dot{x}(t) - F_{yy}(x(t), \dot{x}(t))\ddot{x}(t).$$

Thus, we have the EE as in (8.2):

$$\begin{aligned} &F_x(x(t), \dot{x}(t)) + \rho F_y(x(t), \dot{x}(t)) \\ &= F_{yx}(x(t), \dot{x}(t))\dot{x}(t) + F_{yy}(x(t), \dot{x}(t))\ddot{x}(t). \end{aligned}$$

8.3 Neoclassical growl model

8.3.1 Discrete time

In the discrete-time neoclassical growth model, output can be used for consumption, C_t, and investment, I_t:

$$C_t + I_t = G(k_t, 1),$$

where $G(\cdot, 1)$ is a production function that is strictly increasing, strictly concave and satisfies we assume that $\lim_{k\to 0} G'(k, 1) = \infty$ and that $\lim_{k\to\infty} G'(k, 1) = 0$. The next-period capital is given by

$$k_{t+1} = k_t(1 - \delta) + I_t,$$

where δ is the depreciation rate. Combining the two gives the *law of motion*:

$$\begin{aligned} C_t &= G(k_t, 1) + k_t(1 - \delta) - k_{t+1} \\ &= f(k_t) - k_{t+1}, \end{aligned}$$

where we defined $f(k) \coloneqq G(k, 1) + (1 - \delta)k$. The consumer then chooses an infinite sequence of k_{t+1} to maximise utility (in any given period t, k_t is given). Thus, the neoclassical growth model takes the following form:

$$\begin{aligned} V^*(k_0) \coloneqq \max_{(k_{t+1})_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t U(f(k_t) - k_{t+1}) \\ \text{s.t.} \quad & 0 \le k_{t+1} \le f(k_t), \\ & k_0 \text{ given.} \end{aligned}$$

We assume that U is strictly increasing and strictly concave. Moreover, we assume that $\lim_{c\to 0} U'(c) = \infty$ and that $\lim_{c\to\infty} U'(c) = 0$. In setting the lower bound for k_{t+1} to be zero, we are implicitly assuming that capital can be dismantled at no cost and consumed. If capital is irreversible, then the lower bound for k_{t+1} should be set to k_t.

We can fit the problem above into our general notation:

$$F(x, y) = U(f(x) - y), \ \Gamma(x) = [0, f(x)].$$

EE and TC Since

$$F_x(x,y) = U'(f(x)-y)f'(x),$$
$$F_y(x,y) = -U'(f(x)-y),$$

the Euler Equations for the neoclassical growth model is given by:

$$-U'(f(k_t)-k_{t+1}) + \beta U'(f(k_{t+1})-k_{t+2})f'(k_{t+1}) = 0 \ \forall t \geq 0.$$

The Transversality Condition is given by

$$\lim_{T\to\infty} \beta^T U'(f(k_T)-k_{T+1})f'(k_T)k_T = 0.$$

Steady state The steady state solves

$$\begin{aligned}\beta U'\left(f\left(\bar{k}\right)-\bar{k}\right)f'\left(\bar{k}\right) &= U'\left(f\left(\bar{k}\right)-\bar{k}\right)\\ \Leftrightarrow f'\left(\bar{k}\right) &= \frac{1}{\beta}.\end{aligned}$$

8.3.2 Continuous time

In the continuous-time neoclassical growth model, law of motion for capital is given by

$$\dot{k}(t) = f(k(t)) - c(t) - \delta k(t),$$

where we implicitly assume unit inelastic supply of labour as in the discrete time case. Unlike in the discrete-time case, above f is the production function gross of depreciation. The consumer then chooses an infinite path of $\dot{k}(t)$ to maximise his utility. Thus, the neoclassical growth model now takes the following form:

$$\begin{aligned}V^*(k(0)) \coloneqq \max_{\dot{k}(t)} \quad & \int_0^\infty e^{-\rho t} U\left(f(k) - \delta k(t) - \dot{k}(t)\right)\\ s.t. \quad & \dot{k}(t) \in \mathbb{R} \ \forall t \geq 0,\\ & k(0) \text{ given.}\end{aligned}$$

We can fit the problem above into our general notation:

$$F\left(k,\dot{k}\right) = U\left(f(k) - \delta k - \dot{k}\right), \ \Gamma(k) = \mathbb{R}.$$

EE and TC Since

$$\begin{aligned}F_x &= \left(f'(k)-\delta\right)U'\left(f(k)-\delta k-\dot{k}\right),\\ F_y &= -U'\left(f(k)-\delta k-\dot{k}\right),\\ F_{yx} &= -\left(f'(k)-\delta\right)U''\left(f(k)-\delta k-\dot{k}\right),\\ F_{yy} &= U''\left(f(k)-\delta k-\dot{k}\right),\end{aligned}$$

the EE are

$$\begin{aligned}
&\left(f'\left(k\right)-\delta\right)U'-\rho U'=-\left(f'\left(k\right)-\delta\right)U''\dot{k}+U''\ddot{k}\\
\Leftrightarrow&\left(f'\left(k\right)-\delta-\rho\right)U'=-\left(\left(f'\left(k\right)-\delta\right)\dot{k}-\ddot{k}\right)U''.
\end{aligned}\tag{8.9}$$

The TC is given by

$$\lim_{T\to\infty}-e^{-\rho T}U'\left(f\left(k\left(T\right)\right)-\delta k\left(T\right)-\dot{k}\left(T\right)\right)k\left(T\right)=0.$$

Steady state The steady state is given by

$$\left(f'\left(k\right)-\delta-\rho\right)U'=0\Rightarrow f'\left(k\right)=\rho+\delta.$$

8.3.3 Hamiltonian

We can also analyse the neoclassical model using the Hamiltonian. The period-return function is $U(c)$ and the law of motion is given by $\dot{k}=f(k)-\delta k-c$ (note that $k(0)$ is given). That is,

$$\begin{aligned}
h\left(k,c\right)&=U\left(c\right),\\
g\left(k,c\right)&=f\left(k\right)-\delta k-c,\\
\Rightarrow H\left(k,c,\lambda\right)&=U\left(c\right)+\lambda\left(f\left(k\right)-\delta k-c\right).
\end{aligned}$$

Then,

$$\begin{aligned}
H_u=0\Rightarrow U'\left(c\right)&=\lambda,\\
\dot{\lambda}=\rho\lambda-H_x\Rightarrow\dot{\lambda}&=\lambda\left(\rho-\left(f'\left(k\right)-\delta\right)\right),\\
\dot{x}=g\Rightarrow\dot{k}&=f\left(k\right)-\delta k-c.
\end{aligned}$$

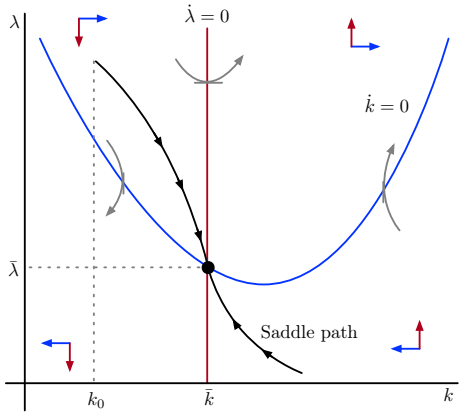
We therefore have the following dynamic equations:

$$\begin{aligned}
\dot{\lambda}&=\lambda\left(\rho-\left(f'\left(k\right)-\delta\right)\right),\\
\dot{k}&=f\left(k\right)-\delta k-c.
\end{aligned}$$

Phase diagram in (k,λ) space To draw the phase diagram in (k,λ) space, note that

- $\dot{\lambda}=0$: $f'(\bar{k})-\delta=\rho$. In (k,λ) space, this is a vertical line. Dynamics: From $\dot{\lambda}=0$ if $k>\bar{k}$, then $f'(k)$ is lower and $-f'(k)$ is higher so that $\dot{\lambda}>0$. And if $k<\bar{k}$ then $\dot{\lambda}<0$.
- $\dot{k}=0$: $c=f(k)-\delta k$. Since $\lambda=U'(c)$, we have that $\lambda=U'(f(k)-\delta k)$. Note that $U'>0$ and $U''<0$ so that U' is a strictly decreasing transformation. Thus, when $f(k)-\delta k$ achieves its maximum (at $\hat{k}$ such that $f'(\hat{k})=\delta$), $U'(f(k)-\delta)$ is at its minimum. As $k\to 0$, $f(k)\to 0$ and $U'(f(k)-\delta)\to\infty$. As $k>\hat{k}$, $f(\hat{k})$ falls so that λ increases. These observations imply that $\dot{k}=0$ locus is U-shaped in (k,λ) space. Dynamics: If $c>\bar{c}$, then $\dot{k}<0$ and if $c<\bar{c}$, then $\dot{k}>0$. Note that $c>\bar{c}$ represents points below the $\dot{k}=0$ locus.

This gives the following phase diagram.



We see that low value of k corresponds to a higher value of λ. Does it make sense? Recall that λ is the marginal value of a unit of k and, at the optimal, marginal benefit from c is equated to λ, the marginal value of capital. In a similar way that higher c implies lower marginal benefit, a higher k implies lower marginal value of capital; i.e. it represents diminishing returns. Note also that a higher k implies higher production, and since c is a normal good (this is due to the fact that we have separable utility function), higher output implies higher income and c increases in every period.

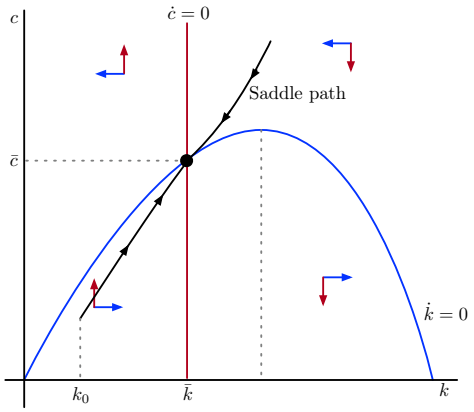
Phase diagram in (k, c) space Since there is a one-to-one mapping between c and λ, we can also draw the phase diagram in (k, c) space. To do so, we can differentiate $H_u = 0$ condition with respect to time to obtain $\dot{\lambda}$ in terms of c:

$$U''(c)\,\dot{c} = \dot{\lambda}.$$

Then we can rewrite the second first-order condition as

$$\begin{aligned} U''(c)\,\dot{c} &= U'(c)\,(\rho - (f'(k) - \delta)) \\ \Rightarrow \frac{\dot{c}}{c} &= \frac{1}{-\frac{U''(c)c}{U'(c)}}\,(f'(k) - \delta - \rho), \end{aligned}$$

where $-\frac{U''(c)}{U'(c)}c$ is the elasticity of intertemporal substitution, and $\frac{\dot{c}}{c}$ is the percentage change in c (i.e. elasticity). Thus, we see that given $f'(k)$ and ρ, consumption level c is determined by preferences. Notice also that, if $f'(k) - \delta = \rho$, then c is constant, and if $f'(k) \neq \rho$, then as u becomes linear (i.e. $U'' \to 0$), $\dot{c}$ is larger so that you converge more quickly to the steady state.



8.3.4 Deriving the continuous time EE from the discrete time version

Let Δ denote the length of time between periods when the state is determined. Decisions are taken whenever the state is determined at times 0, Δ, 2Δ, ... The sequence of state is then

$$\left(x_{\Delta(s+1)}\right)_{s=0}^{\infty} = \{x_0, x_\Delta, x_{2\Delta}, \ldots\},$$

where x_0, as usual, is given. We define the discount factor β for an interval Δ using (the instantaneous) discount rate ρ as

$$\beta = \frac{1}{1+\Delta\rho}.$$

We let U denote the instantaneous utility from consuming c_t amount of consumption. So, during an interval of length Δ in which consumption is given by $c_{s\Delta}$, the total utility is given by

$$\Delta U\left(c_{s\Delta}\right).$$

Market clearing condition must hold at all times (including during the interval Δ). Instantaneous market clearing condition is thus

$$c_s + i_s = f\left(k_s\right),$$

where i_s denotes investment. Market clearing condition for the interval of length Δ from period $t\Delta$ is thus

$$\Delta c_{s\Delta} + \Delta i_{s\Delta} = \Delta f\left(k_{s\Delta}\right).$$

Law of motion for capital is

$$k_{s\Delta+\Delta} = k_{(s+1)\Delta} = \Delta i_{s\Delta} + k_{s\Delta}\left(1-\Delta\delta\right),$$

where δ is the instantaneous depreciation rate. Thus, we can write the neoclassical growth model: as

$$\begin{aligned}
\max_{(c_t,i_t)_{t=0}^{\infty}} \quad & \sum_{s=0}^{\infty}\left(\frac{1}{1+\Delta\rho}\right)^s \Delta U\left(c_{s\Delta}\right), \\
s.t. \quad & \Delta c_t + \Delta i_t = \Delta f\left(k_t\right),\ \forall t \geq 0, \\
& k_{t+\Delta} = \Delta i_t + k_t\left(1-\Delta\delta\right),\ \forall t \geq 0, \\
& k_0 \text{ given},
\end{aligned}$$

where $t = s\Delta$ for some integer s. In other words, we may write the problem in an equivalent manner as

$$\begin{aligned}
\max_{(c_{i\Delta},i_{i\Delta})_{i=0}^{\infty}} \quad & \sum_{s=0}^{\infty}\left(\frac{1}{1+\Delta\rho}\right)^s \Delta U\left(c_{s\Delta}\right), \\
s.t. \quad & \Delta c_{s\Delta} + \Delta i_{s\Delta} = \Delta f\left(k_{s\Delta}\right),\ \forall s \geq 0, \\
& k_{(s+1)\Delta} = \Delta i_{s\Delta} + k_{s\Delta}\left(1-\Delta\delta\right),\ \forall s \geq 0, \\
& k_0 \text{ given},
\end{aligned}$$

Letting $\Delta = 1$, the problem reduces to the standard one.

Continuous-time version of the law of motion for capital Fix Δ. First, we eliminate Δi_t from the constraint to obtain the law of motion for capital:

$$k_{t+\Delta} = \Delta f(k_t) + k_t(1 - \Delta\delta) - \Delta c_t. \tag{8.10}$$

Rearranging above and dividing through by Δ:

$$\frac{k_{t+\Delta} - k_t}{\Delta} = f(k_t) - \delta k_t - c_t. \tag{8.11}$$

Taking limits as $\Delta \downarrow 0$:

$$\lim_{\Delta \downarrow 0} \frac{k_{t+\Delta} - k_t}{\Delta} \coloneqq \dot{k}(t) = f(k(t)) - \delta k(t) - c(t),$$

where we change notation following the convention (capital at time t is written k_t in discrete time and $k(t)$ in continuous time). This gives the continuous time version of the law of motion for capital.

Continuous-time version of EE Define

$$\dot{k}_t \coloneqq \frac{k_{t+\Delta} - k_t}{\Delta} \Rightarrow k_{t+\Delta} = \dot{k}_t \Delta + k_t.$$

We can then rewrite (8.11), while noting that $t = s\Delta$, as

$$\begin{aligned} c_t = c_{s\Delta} &= f(k_t) - \delta k_t - \frac{k_{t+\Delta} - k_t}{\Delta} \\ &= f(k_t) - \delta k_t - \dot{k}_t \end{aligned}$$

Using (8.10), we can then write the problem as

$$\max_{\{\dot{k}_t\}_{t=0}^{\infty}} \quad \sum_{t=0}^{\infty} \left(\frac{1}{1+\Delta\rho}\right)^t \Delta U\left(f(k_t) - \delta k_t - \dot{k}_t\right).$$

Writing the sequence as

$$\begin{aligned} &\cdots + \left(\frac{1}{1+\Delta\rho}\right)^t \Delta U\left(f(k_t) - \delta k_t - \dot{k}_t\right) + \left(\frac{1}{1+\Delta\rho}\right)^{t+1} \Delta U\left(f(k_{t+\Delta}) - \delta k_{t+\Delta} - \dot{k}_{t+\Delta}\right) \cdots \\ =&\cdots + \left(\frac{1}{1+\Delta\rho}\right)^t \Delta U\left(f(k_t) - \delta k_t - \dot{k}_t\right) + \left(\frac{1}{1+\Delta\rho}\right)^{t+1} \Delta U\left(f\left(\dot{k}_t\Delta + k_t\right) - \delta\left(\dot{k}_t\Delta + k_t\right) - \frac{k_{t+2\Delta} - k_{t+\Delta}}{\Delta}\right) \cdots \\ =&\cdots + \left(\frac{1}{1+\Delta\rho}\right)^t \Delta U\left(f(k_t) - \delta k_t - \dot{k}_t\right) + \left(\frac{1}{1+\Delta\rho}\right)^{t+1} \Delta U\left(f\left(\dot{k}_t\Delta + k_t\right) - \delta\left(\dot{k}_t\Delta + k_t\right) - \frac{k_{t+2\Delta} - \left(\dot{k}_t\Delta + k_t\right)}{\Delta}\right) \\ =&\cdots + \left(\frac{1}{1+\Delta\rho}\right)^t \Delta U\left(f(k_t) - \delta k_t - \dot{k}_t\right) + \left(\frac{1}{1+\Delta\rho}\right)^{t+1} \Delta U\left(f\left(\dot{k}_t\Delta + k_t\right) - \delta\left(\dot{k}_t\Delta + k_t\right) - \frac{k_{t+2\Delta} - k_t}{\Delta} + \dot{k}_t\right) \cdots \end{aligned}$$

The first-order condition with respect to $\dot{k}_t$ is

$$-\left(\frac{1}{1+\Delta\rho}\right)^t \Delta U'(c_t) + \left(\frac{1}{1+\Delta\rho}\right)^{t+1} \Delta U'(c_{t+\Delta})\left(\Delta f'(k_{t+\Delta}) - \Delta\delta + 1\right) = 0.$$

Rearranging this gives the EE:

$$U'(c_t) = \left(\frac{1}{1+\Delta\rho}\right) U'(c_{t+\Delta}) \left(\Delta f'(k_{t+\Delta}) + (1-\Delta\delta)\right).$$

Taylor expansion of $U'(c_{t+\Delta})$ around c_t gives

$$U'(c_{t+\Delta}) = U'(c_t) + U''(c_t)(c_{t+\Delta} - c_t) + R(c_{t+\Delta} - c_t), \tag{8.12}$$

where $R(c_{t+\Delta} - c_t) = o(\Delta)$ as $\Delta \to 0$; i.e.

$$\lim_{\Delta \downarrow 0} \frac{R(c_{t+\Delta} - c_t)}{\Delta} = 0.$$

Substituting (8.12) into the EE:

$$(1+\Delta\rho) U'(c_t) = \left(U'(c_t) + U''(c_t)(c_{t+\Delta} - c_t) + R(c_{t+\Delta} - c_t)\right)\left(\Delta f'(k_{t+\Delta}) + (1-\Delta\delta)\right).$$

Collecting $U'(c_t)$ together

$$\begin{aligned}
&\left[(1+\Delta\rho) - \Delta f'(k_{t+\Delta}) - (1-\Delta\delta)\right] U'(c_t) \\
&= \left[U''(c_t)(c_{t+\Delta} - c_t) + R(c_{t+\Delta} - c_t)\right]\left[\Delta f'(k_{t+\Delta}) + (1-\Delta\delta)\right].
\end{aligned}$$

Dividing through by Δ:

$$\begin{aligned}
&\left[\rho - f'(k_{t+\Delta}) + \delta\right] U'(c_t) \\
&= \left(U''(c_t)\frac{(c_{t+\Delta} - c_t)}{\Delta} + \frac{R(c_{t+\Delta} - c_t)}{\Delta}\right)\left(\Delta f'(k_{t+\Delta}) + (1-\Delta\delta)\right).
\end{aligned}$$

Taking limits as $\Delta \downarrow 0$ of each side

$$\lim_{\Delta \downarrow 0} \left[\rho - f'(k_{t+\Delta}) + \delta\right] U'(c_t) = \left[\rho - f'(k(t)) + \delta\right] U'(c(t)),$$

and

$$\lim_{\Delta \downarrow 0} \left(U''(c_t)\frac{(c_{t+\Delta} - c_t)}{\Delta} + \frac{R(c_{t+\Delta} - c_t)}{\Delta}\right)\left(\Delta f'(k_{t+\Delta}) + (1-\Delta\delta)\right) = U''(c(t))\,\dot{c}(t),$$

where

$$\dot{c}(t) \coloneqq \lim_{\Delta \downarrow 0} \frac{(c_{t+\Delta} - c_t)}{\Delta}.$$

Combining the two and rearranging, we obtain

$$\left(f'(k(t)) - \delta - \rho\right) U'(c(t)) = -U''(c(t))\,\dot{c}_t,$$

where

$$\begin{aligned}
c(t) &= f(k(t)) - \delta k(t) - \dot{k}(t) \\
\Rightarrow \dot{c}(t) &= \left(f'(k(t)) - \delta\right)\dot{k}(t) - \ddot{k}(t).
\end{aligned}$$

Therefore, we have

$$\left(f'(k(t)) - \delta - \rho\right) U'(c(t)) = -U''(c(t))\left[\left(f'(k(t)) - \delta\right)\dot{k}(t) - \ddot{k}(t)\right].$$

as we had in (8.9).

8.4 Summary

8.4.1 Discrete time

State formation

$$\max_{(x_{t+1})_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1})$$
$$s.t. \quad x_{t+1} \in \Gamma(x_t), \ \forall t \geq 0,$$
$$x_0 \text{ given.}$$

Necessary and sufficient conditions[25]
Euler equation:

$$F_y(x_t, x_{t+1}) + \beta F_x(x_t, x_{t+1}) = 0, \ \forall t \geq 0.$$

Transversality condition:

$$\lim_{T\to\infty} \beta^T F_x(x_T, x_{T+1}) x_T = 0.$$

Control-state formation

$$\max_{(u_t)_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t h(x_t, u_t)$$
$$s.t. \quad x_{t+1} = g(x_t, u_t), \ \forall t \geq 0,$$
$$u_t \in U$$
$$x_0 \text{ given.}$$

We can return to the state formation by letting

$$F(x, y) = \max\{h(x, u) : u \in U, \ y = g(x, u)\},$$
$$\Gamma(x) = \{y : \exists u \in U, \ y = g(x, u)\}.$$

8.4.2 Continuous time

State formation

$$\max_{\{\dot{x}(t)\}_{t=0}^{\infty}} \int_0^{\infty} e^{-\rho t} F(x(t), \dot{x}(t)) \, dt$$
$$s.t. \quad \dot{x}(t) \in \Gamma(x(t)), \ \forall t \geq 0,$$
$$x_0 \text{ given.}$$

Necessary and sufficient conditions
Euler equation:

$$F_x + \rho F_{\dot{x}} = F_{\dot{x}x}\dot{x} + F_{\dot{x}\dot{x}}\ddot{x}, \ \forall t \geq 0$$

Transversality condition:

$$\lim_{T\to\infty} e^{-\rho T} F_{\dot{x}}(x(T), \dot{x}(T)) x(T) = 0.$$

Control-state formation

$$\max_{\{u(t)\}_{t=0}^{\infty}} \int_0^{\infty} e^{-\rho t} h(x(t), u(t))$$
$$s.t. \quad \dot{x}(t) = g(x(t), u(t)), \ \forall t \geq 0,$$
$$u(t) \in U$$
$$x_0 \text{ given.}$$

Necessary and sufficient conditions
First-order conditions:

$$H(x, u, \lambda) = h(x, u) + \lambda g(x, u),$$
$$H_u(x, u, \lambda) = 0,$$
$$\dot{\lambda} = \rho\lambda - H_x(x, u, \lambda),$$
$$\dot{x} = g(x, u).$$

Transversality condition:

$$\lim_{T\to\infty} e^{-\rho T} \lambda(T) x(T) = 0.$$

[25] F concave and C^1.

9 Local stability of optimal paths and speed of convergence

Given an initial state x_0, the solution to a dynamic programming problem completely determines the evolution of the state through time. The purpose of this section is to study the local dynamics and stability of the optimal decision rules in discrete and continuous time models.

9.1 Stability of discrete-time linear dynamic systems of one dimension

9.1.1 Linearisation around the steady state

Let $x_{t+1} = g(x_t)$ be the optimal decision rule (i.e., the equation that describes the saddle path). Using first-order Taylor approximation around $x_t = \overline{x}$ ($\overline{x}$ denotes the steady-state value),

$$x_{t+1} = g(x) \simeq g(\overline{x}) + g'(\overline{x})(x_t - \overline{x}).$$

Since $g(\overline{x}) = \overline{x}$,

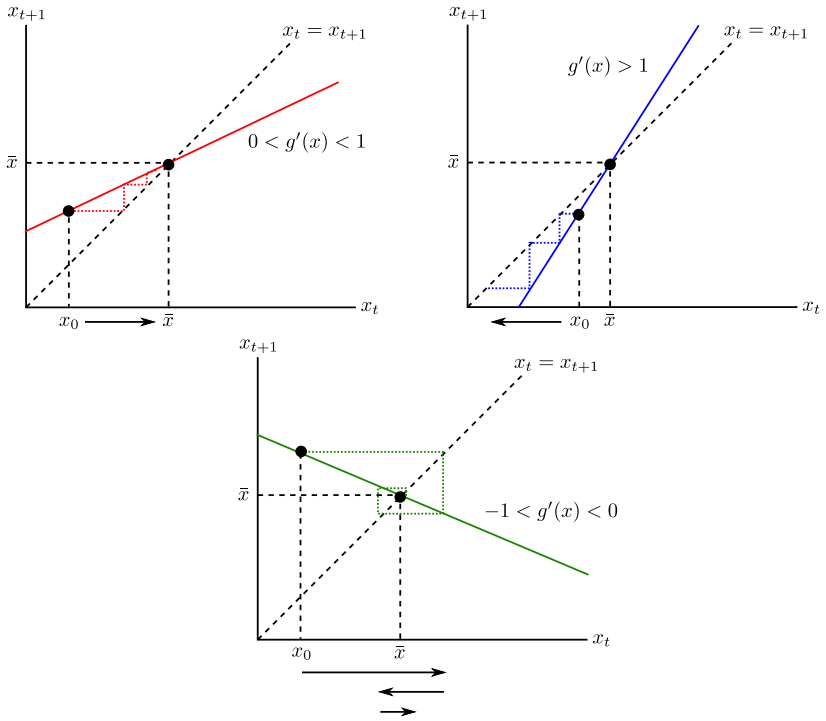
$$x_{t+1} - \overline{x} \simeq g'(\overline{x})(x_t - \overline{x}),$$

9.1.2 Condition for convergence

Suppose that $x_0 < \bar{x}$.

- If $|g'(\overline{x})| > 1$, then $x_{t+1} - \overline{x} > x_t - \overline{x}$; i.e. x_t diverges.
- If $|g'(\overline{x})| < 1$, then $x_{t+1} - \overline{x} < x_t - \overline{x}$; i.e. x_t converges.

Hence, for local stability, we require $|g'(\overline{x})| < 1$. See also the figure below.



Remark 19. If $-1 < g'(\overline{x}) < 0$, then we still converge to the steady state while oscillating above and below $\overline{x}$. This cannot happen in continuous time case since the movements are infinitesimally small.

9.1.3 Obtaining $g'(x)$

Recall the Euler equation (which must hold for any interior solutions):

$$0 = F_y\left(x, g\left(x\right)\right) + \beta F_x\left(g\left(x\right), g\left(g\left(x\right)\right)\right).$$

We want to find out $g'(\overline{x})$, where $\overline{x}$ is the steady state; i.e. $\overline{x}$ solves

$$F_y\left(\overline{x}, \overline{x}\right) + \beta F_x\left(\overline{x}, \overline{x}\right) = 0.$$

This would allow us to analyse the dynamics of x_t close to a steady state.

We follow the following steps to obtain $g'(x)$ close to a steady state.

(i) Differentiate the Euler equation with respect to x. This yields a quadratic equation in $g'(x)$.

(ii) Evaluate the resulting quadratic equation at the steady state, $\overline{x}$.

Differentiating the Euler equation with respect to x yields

$$\begin{aligned} 0 = {} & F_{yx}\left(x, g\left(x\right)\right) + F_{yy}\left(x, g\left(x\right)\right) g'\left(x\right) \\ & + \beta\left[F_{xx}\left(g\left(x\right), g\left(g\left(x\right)\right)\right) g'\left(x\right) + F_{xy}\left(g\left(x\right), g\left(g\left(x\right)\right)\right) g'\left(g\left(x\right)\right) g'\left(x\right)\right]. \end{aligned}$$

Evaluating this derivative at the steady state $\overline{x}$ and using the fact that $g(\overline{x}) = \overline{x}$,

$$\begin{aligned} 0 = {} & F_{yx}\left(\overline{x}, \overline{x}\right) + F_{yy}\left(\overline{x}, \overline{x}\right) g'\left(\overline{x}\right) \\ & + \beta\left[F_{xx}\left(\overline{x}, \overline{x}\right) g'\left(\overline{x}\right) + F_{xy}\left(\overline{x}, \overline{x}\right)\left(g'\left(\overline{x}\right)\right)^2\right] \\ = {} & \beta F_{xy}\left(\overline{x}, \overline{x}\right)\left(g'\left(\overline{x}\right)\right)^2 + \left(F_{yy}\left(\overline{x}, \overline{x}\right) + \beta F_{xx}\left(\overline{x}, \overline{x}\right)\right) g'\left(\overline{x}\right) + F_{yx}\left(\overline{x}, \overline{x}\right) \end{aligned} \tag{9.1}$$

Thus, we obtain a quadratic equation in $g'(\overline{x})$ so that there may potentially be two candidate values for $g'(\overline{x})$.

Let $\lambda = g'(\overline{x})$, we can define the quadratic equation:

$$\tilde{Q}\left(\lambda\right) \coloneqq \beta F_{xy}\lambda^2 + \left(F_{yy} + \beta F_{xx}\right)\lambda + F_{yx}. \tag{9.2}$$

The following proposition shows that the roots of $\tilde{Q}(\lambda)$ come in *almost reciprocal pairs.*

Proposition 97. *If λ_1 solves $\tilde{Q}(\lambda_1) = 0$, then so does $\lambda_2 = 1/\lambda_1\beta$.*

Proof. Suppose $\tilde{Q}(\lambda_1) = 0$; i.e.

$$\tilde{Q}\left(\lambda_1\right) = \beta F_{xy}\lambda_1^2 + \left(F_{yy} + \beta F_{xx}\right)\lambda_1 + F_{yx} = 0.$$

Consider $\tilde{Q}(1/\lambda_1\beta)$:

$$\begin{aligned} \tilde{Q}\left(\frac{1}{\lambda_1\beta}\right) &= \beta F_{xy}\left(\frac{1}{\lambda_1\beta}\right)^2 + \left(F_{yy} + \beta F_{xx}\right)\left(\frac{1}{\lambda_1\beta}\right) + F_{yx} \\ &= F_{xy}\frac{1}{\lambda_1^2\beta} + \left(F_{yy} + \beta F_{xx}\right)\left(\frac{1}{\lambda_1\beta}\right) + F_{yx} \\ &= \frac{1}{\lambda_1^2\beta}\left[F_{xy} + \left(F_{yy} + \beta F_{xx}\right)\lambda_1 + \beta F_{yx}\lambda_1^2\right] \\ [F_{xy} = F_{yx}] &= \frac{1}{\lambda_1^2\beta}\left[\beta F_{xy}\lambda_1^2 + \left(F_{yy} + \beta F_{xx}\right)\lambda_1 + F_{yx}\right] \\ &= \frac{1}{\lambda_1^2\beta}\tilde{Q}\left(\lambda_1\right) = 0. \end{aligned}$$

■

Therefore, if one root is $|\lambda_1| < 1$, then the other root, λ_2, must be larger than one in absolute value. i.e.

$$|\lambda_2| = \left|\frac{1}{\beta\lambda_1}\right| = \frac{1}{\beta\,|\lambda_1|} > 1.$$

As an aside, note that $\lambda_1\lambda_2 = 1/\beta$.

Let x_0 be close to the steady state $\overline{x}$ so that a linear approximation of g is appropriate. Assume we found that the smaller root has absolute value less than one. Now, consider the following sequence of $(x_{t+1})_t$:

$$x_{t+1} = \overline{x} + g'(\overline{x})(x_t - \overline{x}) \;\; \forall t \geq 0. \tag{9.3}$$

The sequence, by construction, satisfies the Euler Equations. Since $|g'(\overline{x})| < 1$, it converges to the steady state $\overline{x}$ and hence it also satisfies the transversality condition. Thus, if the problem is convex, we have found a solution. If, on the other hand, both roots are bigger than one in absolute value, then we do not know which one describes $g'(\overline{x})$, but we do know that the steady state is not locally stable (local since we are approximating).[26]

We now analyse under what circumstances local stability (i.e. $|g'(\overline{x})| < 1$) can be obtained in a one-dimensional discrete-time setting. Recall from (9.2) that

$$\begin{aligned}\tilde{Q}(\lambda) &\coloneqq \beta F_{xy}\lambda^2 + (F_{yy} + \beta F_{xx})\lambda + F_{yx},\\ &= F_{xy}\left[\beta\lambda^2 + \underbrace{\left(\frac{F_{yy} + \beta F_{xx}}{F_{xy}}\right)}_{\coloneqq b}\lambda + 1\right],\end{aligned}$$

where we used that $F_{xy} = F_{yx}$. Assuming $F_{xy} > 0$, then $\tilde{Q}(\lambda) = 0 \Leftrightarrow Q(\lambda) \coloneqq \tilde{Q}(\lambda)/F_{xy} = 0$. We solve for the latter:

$$\begin{aligned}Q(0) &= 1 > 0,\\ Q(1) &= 1 + b + \beta,\\ Q\left(\frac{1}{\beta}\right) &= \frac{1}{\beta} + b\frac{1}{\beta} + 1,\\ \frac{\partial Q(\lambda)}{\partial\lambda} &= 2\beta\lambda + b,\\ \frac{\partial^2 Q(\lambda)}{\partial\lambda^2} &= 2\beta > 0.\end{aligned}$$

Observe also that: (i) if F is concave, then $b < 0$; (ii) $Q(\lambda)$ is strictly convex.

The roots of the latter are given by

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4\beta}}{2\beta}.$$

There are two cases

- $1 + b + \beta < 0$

$$0 < \lambda_1 < 1 < \lambda_2,$$

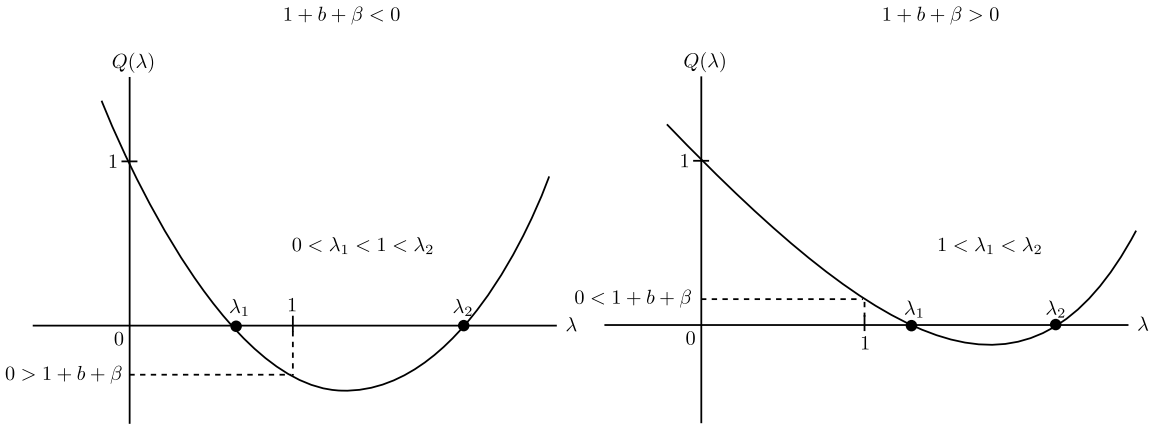
[26] The argument is heuristic since the Euler Equations are satisfied only approximately for the sequence (9.3). Nevertheless, this approximate solution, by the virtue of the implicit function theorem, can be used to construct an exact solution of the Euler Equations that converges to the steady state in a neighbourhood of $\overline{x}$.

$$0 < \lambda_1 < 1 < \lambda_2,$$
$$\lambda_1 = \frac{-b - \sqrt{b^2 - 4\beta}}{2\beta}.$$

- $1 + b + \beta > 0$:

$$1 < \lambda_1 < \lambda_2.$$

See also the figure below.



We can also see how the λ_1 changes with the coefficient b. Taking the first case, note that

$$\begin{aligned}\frac{d\lambda_1}{db} &= \frac{-1 - \frac{b}{\sqrt{b^2-4\beta}}}{2\beta} = \frac{-1 - \frac{b}{\sqrt{b^2-4\beta}}}{2\beta}\frac{\sqrt{b^2-4\beta}}{\sqrt{b^2-4\beta}} \\ &= \frac{-b - \sqrt{b^2-4\beta}}{2\beta}\frac{1}{\sqrt{b^2-4\beta}} = \frac{\lambda_1}{\sqrt{b^2-4\beta}} > 0.\end{aligned}$$

9.1.4 Speed of convergence

The magnitude of $|g'(\overline{x})|$ describes the speed of convergence. To see this, let $\lambda_1 = g'(\overline{x})$ such that $|\lambda_1| < 1$.We can write (9.3) as

$$x_t = \overline{x} + \lambda_1 (x_{t-1} - \overline{x}).$$

Backwards substituting yields

$$\begin{aligned} x_t &= \overline{x} + \lambda_1 ((\overline{x} + \lambda_1 (x_{t-2} - \overline{x})) - \overline{x}) \\ &= \overline{x} + \lambda_1^2 (x_{t-2} - \overline{x}) \\ &= \quad \vdots \\ &= \overline{x} + \lambda_1^t (x_0 - \overline{x}) \\ \Rightarrow x_t - \overline{x} &= \lambda_1^t (x_0 - \overline{x}). \end{aligned}$$

Notice that:

- $\lim_{t\to\infty} x_t - \overline{x} = 0$ since $\lim_{t\to\infty} \lambda_1^t = 0$ given $|\lambda_1| < 1$;
- as $|\lambda_1| \to 1$, convergence is slower. The permanent income hypothesis case is one in which $g'(k^*) = 1$. In this case, the convergence speed is so slow that you, in fact, do not move.

- as $|\lambda_1| \to 0$, convergence is faster. If the cross derivatives, F_{xy} and F_{yx}, are zero (so that the roots are zero), then we would move to the steady state in the next period.

Hence, we realise that the speed of convergence is decreasing in $|\lambda_1|$, or equivalency, it is decreasing in $|g'(\overline{x})|$.

9.1.5 Neoclassical growth model

In the neoclassical model, we have

$$F(x, y) = U(f(x) - y)$$

so that

$$\begin{aligned}
F_x &= f'(x) U'(f(x) - y) \\
F_y &= -U'(f(x) - y) \\
F_{xx} &= f''(x) u'(f(x) - y) + [f'(x)]^2 U''(f(x) - y) \\
F_{yy} &= U''(f(x) - y) \\
F_{xy} &= -f'(x) U''(f(x) - y).
\end{aligned}$$

Recall that the steady-state capital solves $1 = \beta f'(k^*)$. Evaluating these at the steady-state values and substituting into (9.1) gives

$$\begin{aligned}
0 &= F_{yx} + F_{yy} g' + \beta \left(F_{xx} g' + F_{xy} (g')^2 \right) \\
&= -f'U'' + U'' g' + \beta \left(\left(f'U' + (f')^2 U'' \right) g' - f'U'' (g')^2 \right) \\
&= -f'U'' + \left(U'' + \beta f'' Y' + \beta (f')^2 U'' \right) g' - \beta f'U'' (g')^2 \\
&= -U'' \left[f' - \left(1 + \beta f'' \frac{U'}{U''} + \beta (f')^2 \right) g' + \beta f' (g')^2 \right] \\
&= -U'' \left[\frac{1}{\beta} - \left(1 + \frac{1}{\beta} + \beta f'' \frac{U'}{U''} \right) g' + (g')^2 \right] \because f' = 1/\beta \\
&= -U'' \left[\frac{1}{\beta} - \left(1 + \frac{1}{\beta} + \frac{f''}{f'} \frac{U'}{U''} \right) g' + (g')^2 \right] \because \beta = 1/f' \\
&= -U'' \left[\frac{1}{\beta} - \left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'} / \frac{U''}{U'} \right) \right) g' + (g')^2 \right].
\end{aligned}$$

In this discrete-time case, we notice that the expression depends on the elasticity of productivity and elasticity of intertemporal substitution.

The quadratic equation is given by

$$Q(\lambda) = \lambda^2 - \left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'} / \frac{U''}{U'} \right) \right) \lambda + \frac{1}{\beta}.$$

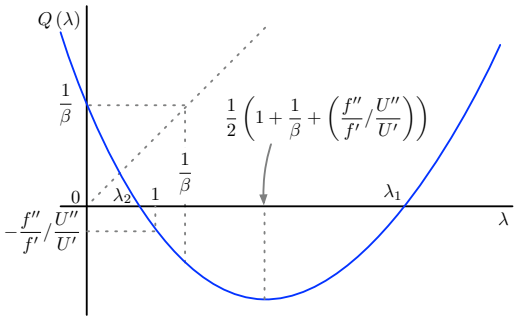
Notice that

$$Q(0) = \frac{1}{\beta} > 0,$$
$$Q(1) = -\frac{f''}{f'}/\frac{U''}{U'} < 0$$
$$Q\left(\frac{1}{\beta}\right) = -\left(\frac{f''}{f'}/\frac{U''}{U'}\right)\frac{1}{\beta} < 0$$

and $Q(\lambda)$ attains a minimum at $Q'(\lambda^*) = 0$, where

$$\lambda^* = \frac{1}{2}\left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'}/\frac{U''}{U'}\right)\right) > 1.$$

See figure below.



Therefore, we have that

$$0 < \lambda_1 < 1 < \frac{1}{\beta} < \lambda_2.$$

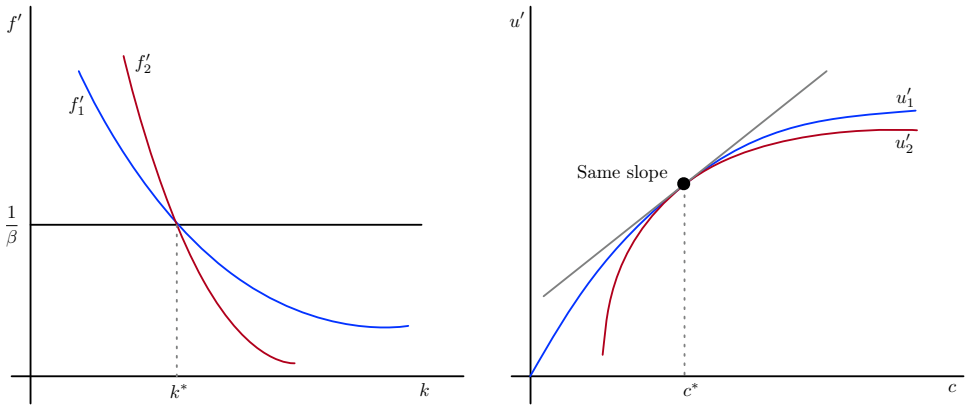
The solution is given by

$$\lambda = \frac{1}{2}\left(\left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'}/\frac{U''}{U'}\right)\right) \pm \sqrt{\left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'}/\frac{U''}{U'}\right)\right)^2 - \frac{4}{\beta}}\right).$$

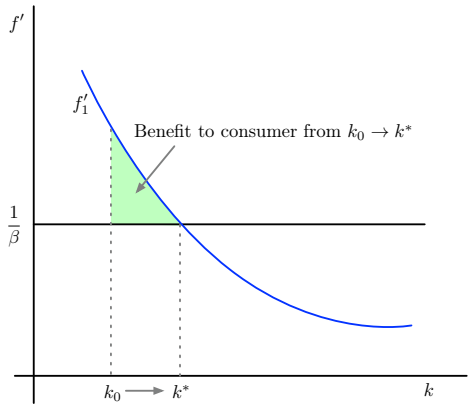
Thus, the root with absolute value less than one is given by

$$\lambda_1 = \frac{1}{2}\left(\left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'}/\frac{U''}{U'}\right)\right) - \sqrt{\left(1 + \frac{1}{\beta} + \left(\frac{f''}{f'}/\frac{U''}{U'}\right)\right)^2 - \frac{4}{\beta}}\right).$$

Suppose we change the production function and the utility function such that the steady states do not change. That is, we change f''/f' and U''/U' but without changing the slope at the steady state. Then, the more curvature there is in U, the more agents prefer to smooth consumption. On the other hand, if u is linear, then agents do not mind consuming today or tomorrow (except for the β discount).



Suppose $k_0 < k^*$, then notice that benefit to the consumer from moving from k_0 to k^* is represented by the shaded area in the figure below (note that, to increase capital stock, agents must abstain from consumption). We therefore see that a higher f'' implies a larger gain so that speed of convergence is faster. In case f' is linear, then it has to coincide with $1/\beta$ and the benefit to consumer is zero. If the production is linear but there is curvature in U, it will take forever to converge. Thus, we see that the speed of convergence depends on the "fight" between the curvature of f and u.



9.2 Stability of discrete-time linear dynamic systems of higher dimensions

Let $\mathbf{x} \in \mathbb{R}^n$ and the function $m : \mathbb{R}^n \to \mathbb{R}^n$ define a dynamic system:

$$\mathbf{x}_{t+1} = m\left(\mathbf{x}_t\right), \forall t \geq 0.$$

Let $\overline{\mathbf{x}}$ be a steady state; i.e.

$$\overline{\mathbf{x}} = m\left(\overline{\mathbf{x}}\right).$$

Consider a first-order approximation of m around $\overline{\mathbf{x}}$:

$$\mathbf{x}_{t+1} = m\left(\overline{\mathbf{x}}\right) + m'\left(\overline{\mathbf{x}}\right)\left(\mathbf{x}_t - \overline{\mathbf{x}}\right).$$

Notice that this analysis is valid globally (i.e. for all $\mathbb{R}^n$) if the system is indeed linear. Alternatively, it is valid in the neighbourhood of the steady state.

Since $\overline{\mathbf{x}} = m(\overline{\mathbf{x}})$, we can write

$$\begin{aligned} \mathbf{x}_{t+1} - \overline{\mathbf{x}} &= m'(\overline{\mathbf{x}})(\mathbf{x}_t - \overline{\mathbf{x}}) \\ \Rightarrow \mathbf{y}_{t+1} &= A\mathbf{y}_t, \end{aligned}$$

where $\mathbf{y}_t = \mathbf{x}_t - \overline{\mathbf{x}}$ and A is the matrix equal to the Jacobian $m'(\overline{\mathbf{x}})$. In this notation, the steady state is $\mathbf{y}^* = \mathbf{0}$.

Diagonalising the matrix $\mathbf{A}$, we obtain

$$A = P\Lambda P^{-1},$$

where Λ is a diagonal matrix with the eigenvalues of A, denoted by λ_i (possibly complex), on its diagonal. As the notation already uses P is invertible and it contains the eigenvectors of A. We can now write the linear system as

$$P^{-1}\mathbf{y}_{t+1} = \Lambda P^{-1}\mathbf{y}_t, \ \forall t \geq 0.$$

Define $\mathbf{z}_t \coloneqq P^{-1}\mathbf{y}_t$, which is a linear combination of the deviations from the steady state using the eigenvectors P. Since P is invertible, there is a one-to-one mapping so that each z corresponds to a unique x, and vice versa. We can then write the system as

$$\begin{aligned} \mathbf{z}_{t+1} &= \Lambda \mathbf{z}_t, \ \forall t \geq 0 \\ \Rightarrow z_{i,t+1} &= \lambda_i z_{it}, \ \forall i = 1, 2, \ldots, n, \ \forall t \geq 0. \end{aligned}$$

We can solve this element by element to obtain that

$$z_{i,t} = \lambda_i^t z_{i,0}, \ \forall i = 1, 2, \ldots, n, \ \forall t \geq 0.$$

Let us consider the case where all the eigenvalues are real. The following proposition says that, if the sequence generated by the dynamic system is to converge to the steady state, then it must be that the initial conditions $\mathbf{x}_0$ belong to a particular linear subspace. The dimension of this subspace is equal to the number of eigenvalues that are larger than one in absolute value ($n - m$ in the notation used above).

Proposition 98. *Let λ_i be such that, for $i = 1, 2, \ldots, m$, we have $|\lambda_i| < 1$ and for $i = m+1, m+2, \ldots, n$, we have $|\lambda_i| \geq 1$. Thus, the eigenvalues of A are ordered so that the first m are smaller than one. Consider the sequence*

$$\mathbf{x}_{t+1} = \overline{\mathbf{x}} + A(\mathbf{x}_t - \overline{\mathbf{x}}) \ \forall t \geq 0$$

for some initial condition $\mathbf{x}_0$. Then,

$$\lim_{t \to \infty} \mathbf{x}_t = \overline{\mathbf{x}},$$

if and only if the initial condition $\mathbf{x}_0$ satisfies

$$\mathbf{x}_0 = P\hat{\mathbf{z}}_0 + \overline{\mathbf{x}},$$

where $\hat{\mathbf{z}}_0$ is a vector with its $n - m$ last coordinates equal to zero; i.e.

$$\hat{\mathbf{z}}_{i,0} = 0 \ \forall i = m+1, m+2, \ldots, n$$

and where the remaining elements of $\hat{\mathbf{z}}_0$ are arbitrary.

Proof. Recall that

$$z_{i,t} = \lambda_i^t z_{i,0},\ \forall i = 1, 2, \ldots, n,\ \forall t \geq 0.$$

For any $|\lambda_i| > 1$ with positive initial value, notice that the sequence is exploding. Hence, in order for the system to converge, it must be that for all i with $|\lambda_i| > 1$, $z_{i,0} = 0$. Recall that

$$\begin{aligned}\mathbf{z}_t = P^{-1}\mathbf{y}_t = P^{-1}\left(\mathbf{x}_t - \overline{\mathbf{x}}\right) &\Rightarrow P\mathbf{z}_t = \mathbf{x}_t - \overline{\mathbf{x}} \\ &\Rightarrow \mathbf{x}_t = P\mathbf{z}_t + \overline{\mathbf{x}} \\ &\Rightarrow \mathbf{x}_0 = P\mathbf{z}_0 + \overline{\mathbf{x}}.\end{aligned}$$

Therefore, it must be the case that $\mathbf{x}_0 = P\hat{\mathbf{z}}_0 + \overline{\mathbf{x}}$ for the system to converge. ■

Remark 20. If λ_i can be complex, then we would only consider the real part of the complex root and the conditions are the same with respect to the real part.

9.2.1 Log-linearisation

Suppose that the system of equations is given by

$$G\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right) = \mathbf{0}, \tag{9.4}$$

where $\mathbf{x}_t = (x_{j,t})_{j=1}^m \in \mathbb{R}^{m\times 1}$ are m variables consisting of both state and control variables and $G : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ are m explicit/implicit equations for the m variables. If the system has a stationary point, denoted $\overline{\mathbf{x}}$, it satisfies

$$G\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right) = \mathbf{0}.$$

Note that

$$x_{j,t} = e^{\log x_{j,t}} = e^{\log x_{j,t} - \log x_{j,t}^* + \log x_{j,t}^*} = \left(e^{\log x_{j,t} - \log x_{j,t}^*}\right) x_{j,t}^*.$$

Define $\widehat{x}_{j,t} \coloneqq \log x_{j,t} - \log x_{j,t}^*$ as the log-deviation from the steady state; i.e., percentage deviation of $x_{j,t}$ from x_j^*. For each $j \in \{1, 2, \ldots, m\}$, write

$$0 = G_j\left(\mathbf{x}_t, \mathbf{x}_{t+1}\right) \equiv G_j\left(e^{\hat{\mathbf{x}}_t}\overline{\mathbf{x}}, e^{\hat{\mathbf{x}}_{t+1}}\overline{\mathbf{x}}\right),$$

where

$$e^{\hat{\mathbf{x}}_t}\overline{\mathbf{x}} \coloneqq \begin{bmatrix} e^{\widehat{x}_{1,t}}x_{1,t}^* & e^{\widehat{x}_{2,t}}x_{2,t}^* & \cdots & e^{\widehat{x}_{m,t}}x_{m,t}^* \end{bmatrix}^\top.$$

To log-linearise, we use first-order Taylor expansion around the stationary point:

$$\mathbf{x}_t = \mathbf{x}_{t+1} = \overline{\mathbf{x}} \Leftrightarrow \hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t+1} = \mathbf{0}.$$

For each $j \in \{1, 2, \ldots, m\}$, Taylor expansion gives

$$\begin{aligned}0 \simeq \underbrace{G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}_{=0} &+ \sum_{i=1}^m \left(\frac{\partial G_j\left(e^{\hat{\mathbf{x}}_t}\overline{\mathbf{x}}, e^{\hat{\mathbf{x}}_{t+1}}\overline{\mathbf{x}}\right)}{\partial x_{i,t}} e^{\hat{\mathbf{x}}_t} x_i^* \Big|_{\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t+1} = 0} \right) \left(\hat{x}_{i,t} - 0\right) \\ &+ \sum_{i=1}^m \left(\frac{\partial G_j\left(e^{\hat{\mathbf{x}}_t}\overline{\mathbf{x}}, e^{\hat{\mathbf{x}}_{t+1}}\overline{\mathbf{x}}\right)}{\partial x_{i,t+1}} e^{\hat{\mathbf{x}}_{t+1}} x_i^* \Big|_{\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_{t+1} = 0} \right) \left(\hat{x}_{i,t+1} - 0\right) \\ = \sum_{i=1}^m \frac{\partial G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{i,t}} x_i^* \hat{x}_{i,t} &+ \sum_{i=1}^m \frac{\partial G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{i,t+1}} x_i^* \hat{x}_{i,t+1}.\end{aligned}$$

For each $t \in \mathbb{N}$, define an $m \times m$ matrix G_t^* as

$$G_1^* \coloneqq \left[\frac{\partial G_i\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{j,t}} x_j^*\right]_{ij}, \; G_2^* \coloneqq \left[\frac{\partial G_i\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{j,t+1}} x_j^*\right]_{ij}$$

Using this matrix notation, we can rewrite the log-linearised system of equations succinctly as

$$G_t^* \hat{\mathbf{x}}_t + G_{t+1}^* \hat{\mathbf{x}}_{t+1} \simeq \mathbf{0}.$$

We "approximate"; i.e. we assume that the equation above holds with equality.

If G_2^* is invertible, then we can write

$$\hat{\mathbf{x}}_{t+1} = -\left(G_2^*\right)^{-1} G_1^* \hat{\mathbf{x}}_t = M \hat{\mathbf{x}}_t,$$

where $M \coloneqq -(\mathbf{G}_2^*)^{-1}\mathbf{G}_1^*$ does not depend on t since the partial derivatives are evaluated at the stationary point.

Backwards substitution yields

$$\hat{\mathbf{x}}_t = M \hat{\mathbf{x}}_{t-1} = M^2 \hat{\mathbf{x}}_{t-2} = \cdots = M^t \hat{\mathbf{x}}_0. \tag{9.5}$$

To solve the this difference equation we diagonalise the matrix M. Let $\boldsymbol{\lambda} \in \mathbb{R}^m$ denote the m *distinct* and *real* eigenvalues of M, and denote the corresponding $m \times m$ matrix of eigenvectors as $E = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m]$. By definition, $\mathbf{e}_i$ is an eigenvector of M if

$$M\mathbf{e}_i = \lambda_i \mathbf{e}_i,$$

which means that we can write

$$ME = E\Lambda,$$

where $\Lambda \coloneqq \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_m]$. Furthermore, if E is invertible, then,

$$\begin{aligned}
M &= E\Lambda E^{-1}, \\
\hat{\mathbf{x}}_{t+1} &= E\Lambda E^{-1} \hat{\mathbf{x}}_t \\
&= E\Lambda E^{-1} E\Lambda = E\Lambda^2 E^{-1} \hat{\mathbf{x}}_{t-1} \\
&= \quad \vdots \\
&= E\Lambda^{t+1} E^{-1} \hat{\mathbf{x}}_0 \\
\Rightarrow E^{-1}\hat{\mathbf{x}}_{t+1} &= \Lambda^{t+1} E^{-1} \hat{\mathbf{x}}_0.
\end{aligned}$$

Define $\mathbf{z}_t \coloneqq E^{-1}\hat{\mathbf{x}}_{t+1}$ so we can express above as

$$\mathbf{z}_{t+1} = \Lambda^{t+1} \mathbf{z}_0.$$

But because Λ is a diagonal matrix, each of the m equations are independent; i.e. (rolling back one period)

$$z_{i,t} = \lambda_{i,t}^t z_{i,0}, \; i = 1, 2, \ldots, m.$$

Note that

$$\begin{aligned}\hat{\mathbf{x}}_t &= E\mathbf{z}_t = E\Lambda^t\mathbf{z}_0 \\ &= \begin{bmatrix} e_{11} & e_{22} & \cdots & e_{m1} \\ e_{12} & e_{22} & \cdots & e_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1m} & e_{2m} & \cdots & e_{mm} \end{bmatrix} \begin{bmatrix} \lambda_1^t & 0 & \cdots & 0 \\ 0 & \lambda_2^t & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m^t \end{bmatrix} E^{-1}\hat{\mathbf{x}}_0 \\ &= \begin{bmatrix} \mathbf{e}_1\lambda_1^t & \mathbf{e}_2\lambda_2^t & \cdots & \mathbf{e}_m\lambda_m^t \end{bmatrix} E^{-1}\hat{\mathbf{x}}_0\end{aligned}$$

Taking the transpose of both sides gives

$$\begin{aligned}\Rightarrow \hat{\mathbf{x}}_t^\top &= \underbrace{\hat{\mathbf{x}}_0^\top \left(E^{-1}\right)^\top}_{=\boldsymbol{\mu}_{1\times m}} \begin{bmatrix} \lambda_1^t\mathbf{e}_1' \\ \lambda_2^t\mathbf{e}_2' \\ \vdots \\ \lambda_m^t\mathbf{e}_m' \end{bmatrix}_{m\times m} \\ &= \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_m \end{bmatrix} \begin{bmatrix} \lambda_1^t e_{11} & \lambda_1^t e_{12} & \cdots & \lambda_1^t e_{1m} \\ \lambda_2^t e_{21} & \lambda_2^t e_{22} & \cdots & \lambda_2^t e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_m^t e_{m1} & \lambda_m^t e_{m2} & \cdots & \lambda_m^t e_{mm} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=1}^m \mu_j\lambda_j^t e_{j1} & \sum_{j=1}^m \mu_j\lambda_j^t e_{j2} & \cdots & \sum_{j=1}^m \mu_j\lambda_j^t e_{jm} \end{bmatrix}\end{aligned}$$

$$\Rightarrow \begin{bmatrix} \hat{x}_{1,t} \\ \hat{x}_{2,t} \\ \vdots \\ \hat{x}_{m,t} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^m \mu_j\lambda_j^t e_{j1} \\ \sum_{j=1}^m \mu_j\lambda_j^t e_{j2} \\ \vdots \\ \sum_{j=1}^m \mu_j\lambda_j^t e_{jm} \end{bmatrix} = \sum_{j=1}^m \begin{bmatrix} \mu_j\lambda_j^t \\ \mu_j\lambda_j^t \\ \vdots \\ \mu_j\lambda_j^t \end{bmatrix} \mathbf{e}_j = \sum_{i=1}^m \mu_i\lambda_i^t\mathbf{e}_i$$

That is, we have

$$\hat{x}_{i,t} = \sum_{j=1}^m \mu_j\lambda_j^t e_{ji} \;\forall i \in \{1,\ldots,m\}.$$

Stacking over the is gives

$$\hat{\mathbf{x}}_t = \begin{bmatrix} \sum_{j=1}^m \mu_j\lambda_j^t e_{j1} \\ \sum_{j=1}^m \mu_j\lambda_j^t e_{j2} \\ \vdots \\ \sum_{j=1}^m \mu_j\lambda_j^t e_{jm} \end{bmatrix} = \sum_{j=1}^m \begin{bmatrix} \mu_j\lambda_j^t \\ \mu_j\lambda_j^t \\ \vdots \\ \mu_j\lambda_j^t \end{bmatrix} \mathbf{e}_j = \sum_{i=1}^m \mu_i\lambda_i^t\mathbf{e}_i.$$

Thus, we can express $\hat{\mathbf{x}}_t$ as a linear combination of the initial values of $\hat{\mathbf{x}}_0$ with weights given by the eigenvalues and the eigenvectors.

Since we assumed that the eigenvalues are distinct, we can pin down μ_i using the initial condition $\hat{\mathbf{x}}_0$ by solving

$$\hat{\mathbf{x}}_0 = \sum_{i=1}^m \mu_i\mathbf{e}_i.$$

Given that the system converges, it must be the case that, if $|\lambda_i| \geq 1$, then $\mu_i = 0$. We say that the system is *saddle path stable* if not all eigenvalues are in the unit circle.

9.2.2 Log-linearisation versus linearisation

There is no rule that says how one should choose between log-linearisation or linearisation. Of course, log-linearisation will not be appropriate if we expect the variables to be negative (e.g. the real interest rate, the inflation rate). However, in general, the approximation tends to be more accurate with log-linearisation if variables are positive. Another advantage of log-linearisation is that we need not worry about the units of the variable as we can interpret in terms of percentage changes with log-linearisation.

If we were to linearise (9.4) around the stationary point $\overline{\mathbf{x}}$, then we obtain

$$0 \simeq \sum_{i=1}^{m} \frac{\partial G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{i,t}} \left(x_{i,t} - x_i^*\right) + \sum_{i=1}^{m} \frac{\partial G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{i,t+1}} \left(x_{i,t+1} - x_i^*\right).$$

For each $t \in \mathbb{N}$, define an $m \times m$ matrix $\tilde{G}_t^*$ as

$$\tilde{G}_1^* \coloneqq \left[\frac{\partial G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{i,t}}\right]_{ij}, \; \tilde{G}_2^* \coloneqq \left[\frac{\partial G_j\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}\right)}{\partial x_{i,t+1}}\right]_{ij},$$

Then

$$\tilde{G}_1^* \left(\mathbf{x}_t - \overline{\mathbf{x}}\right) + \tilde{G}_2^* \left(\mathbf{x}_{t+1} - \overline{\mathbf{x}}\right) \simeq \mathbf{0}.$$

So, if $\tilde{\mathbf{G}}_2^*$ is invertible,

$$\begin{aligned} \mathbf{x}_{t+1} - \overline{\mathbf{x}} &= \left(\tilde{G}_2^*\right)^{-1} \tilde{G}_1^* \left(\mathbf{x}_t - \overline{\mathbf{x}}\right) \\ &= \tilde{M} \left(\mathbf{x}_t - \overline{\mathbf{x}}\right). \end{aligned}$$

9.3 Stability of continuous-time linear dynamic systems of one dimension

This section complements the derivation of the optimal decision rules for a one-dimensional continuous-time dynamic problem. We wish to characterise the optimal decision rule $\dot{k} = g(k)$ to determine the rate of change of the state as a function of the level of the state. In the counterpart to the difference equation we obtained in the discrete-time case, we obtain a differential equation for the function g in the continuous-time case.

In the continuous-time case, we consider the state representation—i.e. a differential equation with respect to the state k—as opposed to the standard representation of the Euler Equation as a (second-order) differential equation on the state at its derivatives with respect to time.

We then use g to study the local dynamics of k—i.e. dynamics of the state variable k close to the steady state value $\bar{k}$, which we summarise by $g'(\bar{k})$. We will find an algebraic (quadratic) equation for $g'(\bar{k})$ and study sufficient conditions for local stability; i.e. for $g'(\bar{k}) < 0$. We will use the neoclassical growth model as an illustration of the general principle.

Notation: We mainly use k instead of x in the subsection.

9.3.1 General continuous-time framework

We work with the continuous-time model as defined before where we choose the derivative of the state with respect to time, $\dot{x}(t)$, in each time period to maximise:

$$\begin{aligned} V^{*}\left(x_{0}\right) \coloneqq \max_{\dot{x}(t)} \quad & \lim_{T\to\infty} \int_{0}^{T} e^{-\rho t} F\left(x\left(t\right), \dot{x}\left(t\right)\right) \mathrm{d}t \\ & s.t. \quad \dot{x}\left(t\right) \in \Gamma\left(x\left(t\right)\right) \ \forall t \geq 0, \\ & \quad x_{0} \text{ given.} \end{aligned}$$

Recall that EE in the continuous time case is given by

$$\begin{aligned} & F_{x}\left(x\left(t\right), \dot{x}\left(t\right)\right) + \rho F_{\dot{x}}\left(x\left(t\right), \dot{x}\left(t\right)\right) \\ & = F_{\dot{x}x}\left(x\left(t\right), \dot{x}\left(t\right)\right) \dot{x}\left(t\right) + F_{\dot{x}\dot{x}}\left(x\left(t\right), \dot{x}\left(t\right)\right) \ddot{x}\left(t\right), \ \forall t \geq 0. \end{aligned} \tag{9.6}$$

To simplify notation, we write EE as

$$H\left(\ddot{x}\left(t\right), \dot{x}\left(t\right), x\left(t\right)\right) = 0, \ \forall t \geq 0,$$

where the function is defined in the obvious way.

The TC is given by

$$0 = \lim_{T\to\infty} e^{-\rho T} F_{\dot{x}}\left(x\left(T\right), \dot{x}\left(T\right)\right) x\left(T\right).$$

Time domain We first review the analysis of the problem with respect to time. We are looking for a path of $k\left(t\right)$—i.e. capital as a function of time—that:

(i) starts at the initial condition

$$k\left(0\right) = k_{0};$$

(ii) satisfies the Euler Equations

$$H\left(\ddot{k}\left(t\right), \dot{k}\left(t\right), k\left(t\right)\right) = 0, \ \forall t > 0; \tag{9.7}$$

(iii) converges to the unique steady state (and hence satisfies transversality)

$$k\left(t\right) \to \bar{k} \text{ as } t \to \infty.$$

State domain Now we consider the analysis of the problem with respect to the state variable. From this perspective, we are looking for a function $\dot{k} = g(k)$—i.e. the rate of change of capital as a function of the level of capital—that:

(i) solves the Euler Equations (note this is not a function of time here any more):

$$H\left(g'\left(k\right) g\left(k\right), g\left(k\right), k\right) = 0, \ \forall k; \tag{9.8}$$

(ii) converges to the steady state (and hence satisfies transversality):[27]

$$g(k) > 0, \text{ if } k < \bar{k},$$
$$g(k) < 0, \text{ if } k > \bar{k}.$$

To see the equivalence between (9.7) and (9.8), differentiate $\dot{k}$ with respect to t, which gives

$$\ddot{k} = \frac{\mathrm{d}\dot{k}}{\mathrm{d}t} = \frac{\mathrm{d}g(k)}{\mathrm{d}t} = g'(k)\frac{\mathrm{d}k}{\mathrm{d}t} = g'(k)\dot{k} = g'(k)g(k).$$

9.3.2 Linearisation around the steady state

Using Taylor expansion on the (non-linear) law of motion for capital $\dot{k}(t) = g(k(t))$ around $k = \bar{k}$:

$$\dot{k}(t) = g(k) \simeq g\left(\bar{k}\right) + g'(k)\left(k - \bar{k}\right).$$

Since $g(\bar{k}) = 0$, we obtain the following differential equation:

$$\dot{k} = g'\left(\bar{k}\right)\left(k - \bar{k}\right).$$

9.3.3 Condition for convergence

Let us focus on values of k close to the steady state $\bar{k}$. The condition for convergence are

$$g(k) > 0, \text{ if } k < \bar{k},$$
$$g(k) < 0, \text{ if } k > \bar{k},$$

which only need to hold in a neighbourhood of $\bar{k}$. The conditions above simply say that capital must be increasing when below the steady state, and decreasing if above the steady state. Given this interpretation, it follows that we can write the equivalent condition for convergence as

$$g'\left(\bar{k}\right) \equiv \frac{\partial g\left(\bar{k}\right)}{\partial k} < 0. \tag{9.9}$$

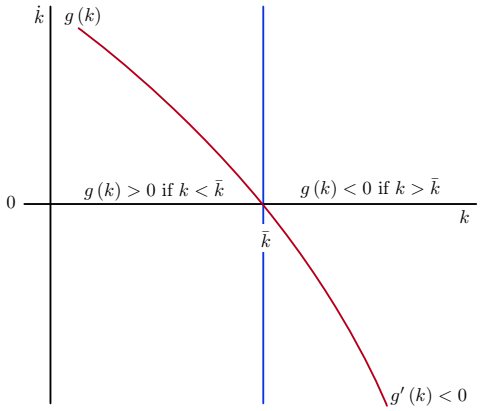
See also the figure below

[27] This condition can be extended to the m dimensional case:

$$\|g(k)\| \text{ is decreasing in } \left\|k - \bar{k}\right\|.$$

More generally, the condition requires that there exists a function $L : \mathbb{R}^m \to R_+$ with $L(k) = 0$ if and only if $k = 0$ and

$$L(g(k))\, g'(k) < 0, \ \forall k.$$



9.3.4 Obtaining $g'(k)$

As in the discrete time case, we use the Euler equation (in state formation) to approximate $g'(k)$ around a steady state $\bar{k}$. Recall

$$H\left(g'\left(k\right)g\left(k\right),g\left(k\right),k\right)=0,\ \forall k. \tag{9.10}$$

The steps are the same as before.

(i) Differentiate the Euler equation with respect to x. This yields a quadratic equation in $g'\left(x\right)$.

(ii) Evaluate the resulting quadratic equation at the steady state, $\overline{x}$.

Differentiating (9.10) with respect to k gives

$$\left(g''\left(k\right)g\left(k\right)+\left(g'\left(k\right)\right)^{2}\right)H_{\ddot{k}}+g'\left(k\right)H_{\dot{k}}+H_{k}=0.$$

Evaluating this expression at the steady state $k=\bar{k}$, where $g(\bar{k})=\dot{k}=0$ so that $g'(\bar{k})g(\bar{k})=\ddot{k}=0$, the expression above simplifies to

$$H_{\dot{k}}\left(0,0,\bar{k}\right)\left(g'\left(\bar{k}\right)\right)^{2}+H_{\dot{k}}\left(0,0,\bar{k}\right)g'\left(\bar{k}\right)+H_{k}\left(0,0,\bar{k}\right)=0. \tag{9.11}$$

Notice that this is a quadratic equation in $g'(\bar{k})$.

We now derive the expressions for the coefficients of the quadratic equation. Recall that

$$H\left(\ddot{k},\dot{k},k\right)=F_{k}\left(k,\dot{k}\right)+\rho F_{\dot{k}}\left(k,\dot{k}\right)-F_{k\dot{k}}\left(k,\dot{k}\right)\dot{k}-F_{\dot{k}\dot{k}}\left(k,\dot{k}\right)\ddot{k}=0.$$

Differentiating this with respect to the three arguments gives

$$\begin{aligned}
H_{\ddot{k}}&=-F_{\dot{k}\dot{k}}\left(k,\dot{k}\right),\\
H_{\dot{k}}&=F_{k\dot{k}}\left(k,\dot{k}\right)+\rho F_{\dot{k}\dot{k}}\left(k,\dot{k}\right)-F_{k\dot{k}\dot{k}}\left(k,\dot{k}\right)\dot{k}-F_{k\dot{k}}\left(k,\dot{k}\right)-F_{\dot{k}\dot{k}\dot{k}}\left(k,\dot{k}\right)\ddot{k},\\
H_{k}&=F_{kk}\left(k,\dot{k}\right)+\rho F_{\dot{k}k}\left(k,\dot{k}\right)-F_{\dot{k}k\dot{k}}\left(k,\dot{k}\right)\dot{k}-F_{\dot{k}\dot{k}k}\left(k,\dot{k}\right)\ddot{k}.
\end{aligned}$$

Evaluating the derivatives at the steady state $k = \bar{k}$ where $\dot{k} = \ddot{k} = 0$, they simplify to

$$\begin{aligned} H_{\dot{k}\dot{k}} &= -F_{\dot{k}\dot{k}}\left(\bar{k}, 0\right), \\ H_{\dot{k}} &= F_{k\dot{k}}\left(\bar{k}, 0\right) + \rho F_{\dot{k}\dot{k}}\left(\bar{k}, 0\right) - F_{\dot{k}k}\left(\bar{k}, 0\right) = \rho F_{\dot{k}\dot{k}}\left(\bar{k}, 0\right), \\ H_{k} &= F_{kk}\left(\bar{k}, 0\right) + \rho F_{\dot{k}k}\left(\bar{k}, 0\right). \end{aligned}$$

Hence, we can now write (9.11) as

$$-F_{\dot{k}\dot{k}}\left(\bar{k}, 0\right)\left(g'\left(\bar{k}\right)\right)^2 + \rho F_{\dot{k}\dot{k}}\left(\bar{k}, 0\right) g'\left(\bar{k}\right) + \left(F_{kk}\left(\bar{k}, 0\right) + \rho F_{\dot{k}k}\left(\bar{k}, 0\right)\right) = 0.$$

Since the expression is quadratic, there may potentially be two candidate values for $g'(k)$.

Let $\lambda = g'(\overline{x})$, we can define the quadratic equation:

$$Q\left(\lambda\right) \coloneqq \left(-F_{\dot{k}\dot{k}}\right)\lambda^2 + \left(\rho F_{\dot{k}\dot{k}}\right)\lambda + \left(F_{kk} + \rho F_{\dot{k}k}\right). \tag{9.12}$$

As in the discrete time case, the roots to $Q(\lambda)$ come in *almost reciprocal pairs.*[28]

Proposition 99. *If λ_1 solves $Q(\lambda_1) = 0$, then so does $\lambda_2 = -\lambda_1 + \rho$.*

Proof. Suppose that λ_1 solves $Q(\lambda_1) = 0$; i.e.

$$Q\left(\lambda_1\right) = \left(-F_{\dot{k}\dot{k}}\right)\lambda_1^2 + \left(\rho F_{\dot{k}\dot{k}}\right)\lambda_1 + \left(F_{kk} + \rho F_{\dot{k}k}\right) = 0.$$

Consider $Q(-\lambda_1 + \rho)$:

$$\begin{aligned} Q\left(-\lambda_1 + \rho\right) &= \left(-F_{\dot{k}\dot{k}}\right)\left(-\lambda_1 + \rho\right)^2 + \left(\rho F_{\dot{k}\dot{k}}\right)\left(-\lambda_1 + \rho\right) + \left(F_{kk} + \rho F_{\dot{k}k}\right) \\ &= \left(-F_{\dot{k}\dot{k}}\right)\lambda_1^2 + \left(-F_{\dot{k}\dot{k}}\right)\rho^2 - \left(-F_{\dot{k}\dot{k}}\right)2\lambda_1\rho - \left(\rho F_{\dot{k}\dot{k}}\right)\lambda_1 + \rho^2 F_{\dot{k}\dot{k}} + \left(F_{kk} + \rho F_{\dot{k}k}\right) \\ &= \left(-F_{\dot{k}\dot{k}}\right)\lambda_1^2 + 2\rho F_{\dot{k}\dot{k}}\lambda_1 - \rho F_{\dot{k}\dot{k}}\lambda_1 + \left(F_{kk} + \rho F_{\dot{k}k}\right) \\ &= \left(-F_{\dot{k}\dot{k}}\right)\lambda_1^2 + \left(\rho F_{\dot{k}\dot{k}}\right)\lambda_1 + \left(F_{kk} + \rho F_{\dot{k}k}\right) \\ &= Q\left(\lambda_1\right) = 0. \end{aligned}$$

■

The theorem means that, if $\lambda_1 < 0$, then $\lambda_2 > 0$. Hence, if we find a solution

$$Q\left(g'\left(\bar{k}\right)\right) = 0 \text{ with } g'\left(\bar{k}\right) < 0,$$

then this is *the* solution since the other root would give that $g'(\bar{k}) > 0$ which we know is not stable (recall (9.9)).

The solution satisfies the EE and TC since it converges to the steady state. The fact that it converges to the steady state also justifies the use of the approximation (i.e. linearisation) of the law of motion for capital since k stays in the neighbourhood of $\bar{k}$. It also means that there is *at most* one stable solution, which is reassuring since a convex problem should have at least one solution.

However, if $\lambda_1 > 0$ and $\lambda_2 > 0$, then the system is *not* locally stable. In this case, local arguments alone do not suffice to identify which one of the roots of Q is the solution for $g'(\bar{k})$, but we know that one of them gives the value of $g'(\bar{k})$.

We now analyse under what circumstances local stability (i.e. $g'(\bar{k}) < 0$) can be obtained in a one-dimensional continuous-time setting. Recall from (9.12) that

$$Q\left(\lambda\right) \coloneqq \left(-F_{\dot{k}\dot{k}}\right)\lambda^2 + \left(\rho F_{\dot{k}\dot{k}}\right)\lambda + \left(F_{kk} + \rho F_{\dot{k}k}\right).$$

[28] Reciprocal here means that one root is positive and the other is negative (ignoring the ρ term).

which can be written as

$$Q(\lambda) = \left(-F_{\dot{k}\dot{k}}\right)\left[\lambda^2 - \rho\lambda - \frac{F_{kk} + \rho F_{\dot{k}k}}{F_{\dot{k}\dot{k}}}\right]$$

so that

$$\begin{aligned} Q(0) &= F_{kk} + \rho F_{\dot{k}k}, \\ Q(\rho) &= Q(0), \\ \frac{\partial Q(\lambda)}{\partial \lambda} &= \left(-F_{\dot{k}\dot{k}}\right)(2\lambda - \rho), \\ \frac{\partial^2 Q(\lambda)}{\partial \lambda^2} &= -2F_{\dot{k}\dot{k}} > 0, \end{aligned}$$

where we assume that F is strictly concave so that $F_{\dot{k}\dot{k}} < 0$. These imply that $Q(\lambda)$ is strictly convex, U-shaped that has a negative slope as it crosses the y-axis. Since $Q(\rho) = Q(0)$, it attains its minimum between $[0, \rho]$.

The solutions are:

$$\lambda = g'(k) = \frac{\rho \pm \sqrt{\rho^2 + 4\frac{-\left(F_{kk} + \rho F_{\dot{k}k}\right)}{\left(-F_{kk}\right)}}}{2}. \tag{9.13}$$

From (9.13), we have the following cases:

- $-(F_{kk} + \rho F_{\dot{k}k})/(-F_{\dot{k}\dot{k}}) > 0$:

  $$\lambda_1 < 0 < \rho < \lambda_2$$

  so that $g'\left(\bar{k}\right) = \lambda_1$ is the locally stable steady state.

- $-(F_{kk} + \rho F_{\dot{k}k})/(-F_{\dot{k}\dot{k}}) < 0$:

  $$0 < \lambda_1 < \lambda_2 < \rho,$$

  which means that steady states are not locally stable.

Remark 21. The previous proposition also holds for higher dimensions. If the state is of dimension m, then there will be $2m$ roots satisfying

$$\lambda_{m+i} = -\lambda_i + \rho, \ \forall i = 1, 2, \ldots, m.$$

In the m dimensional case, the roots are *not* simply $\partial g_i / \partial k_i$. The roots are eigenvalues of the matrix of the derivatives of g. In the m dimensional case, we have

$$\dot{k}_i = g_i(k), \ \forall i = 1, 2, \ldots, m,$$

or, in vector notation,

$$\dot{\mathbf{k}} = \left(\dot{k}_1, \dot{k}_2, \ldots, \dot{k}_m\right) = \mathbf{g}(\mathbf{k}) = \mathbf{g}(k_1, k_2, \ldots, k_m)$$

and

$$\mathbf{g}(k_1, k_2, \ldots, k_m) = \left(g_1(\mathbf{k}), g_2(\mathbf{k}), \ldots, g_m(\mathbf{k})\right).$$

Then, $G'(\mathbf{k})$ is the matrix

$$G'(\mathbf{k}) = \begin{bmatrix} \frac{\partial g_1}{\partial k_1} & \frac{\partial g_1}{\partial k_2} & \cdots & \frac{\partial g_1}{\partial k_m} \\ \frac{\partial g_2}{\partial k_1} & \frac{\partial g_2}{\partial k_2} & \cdots & \frac{\partial g_2}{\partial k_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial k_1} & \frac{\partial g_m}{\partial k_2} & \cdots & \frac{\partial g_m}{\partial k_m} \end{bmatrix}.$$

The eigenvalues of the matrix above control the (local) behaviour of k around $\bar{k}$.

9.3.5 Speed of convergence

The magnitude of $|g'(\bar{k})|$ describes the speed of convergence. To see this, recall the linearised law of motion:

$$\dot{k} = g'\left(\bar{k}\right)\left(k - \bar{k}\right).$$

Rewriting this as

$$\frac{\mathrm{d}k}{\mathrm{d}t} = -g'\left(\bar{k}\right)\bar{k} + g'\left(\bar{k}\right)k,$$

which is differential equation in k.

Solving the differential equation To solve this, guess that the solution is of the form

$$k\left(t\right) = \exp\left[g'\left(\bar{k}\right)t\right]c\left(t\right).$$

Then,

$$\begin{aligned}\frac{\mathrm{d}k\left(t\right)}{\mathrm{d}t} &= g'\left(\bar{k}\right)\exp\left[g'\left(\bar{k}\right)t\right]c\left(t\right) + c'\left(t\right)\exp\left[g'\left(\bar{k}\right)t\right] \\ &= g'\left(\bar{k}\right)k\left(t\right) + c'\left(t\right)\exp\left[g'\left(\bar{k}\right)t\right].\end{aligned}$$

Substituting this into the expression for $\mathrm{d}k/\mathrm{d}t$,

$$\begin{aligned}-g'\left(\bar{k}\right)\bar{k} + g'\left(\bar{k}\right)k\left(t\right) &= g'\left(\bar{k}\right)k\left(t\right) + c'\left(t\right)\exp\left[g'\left(\bar{k}\right)t\right] \\ \Leftrightarrow c'\left(t\right) &= -g'\left(\bar{k}\right)\bar{k}\exp\left[-g'\left(\bar{k}\right)t\right].\end{aligned}$$

Integrating both sides with respect to t

$$\begin{aligned}\int c'\left(t\right)\mathrm{d}t &= \int -g'\left(\bar{k}\right)\bar{k}\exp\left[-g'\left(\bar{k}\right)t\right]\mathrm{d}t \\ \Rightarrow c\left(t\right) &= \bar{k}\exp\left[-g'\left(\bar{k}\right)t\right] + C.\end{aligned}$$

Hence,

$$\begin{aligned}k\left(t\right) &= \exp\left[g'\left(\bar{k}\right)t\right]\left(\bar{k}\exp\left[-g'\left(\bar{k}\right)t\right] + C\right) \\ &= \bar{k} + \exp\left[g'\left(\bar{k}\right)t\right]C.\end{aligned}$$

To pin down C, we need a boundary condition, which is that $k\left(0\right) = k_0$.

$$k_0 = \bar{k} + C \Leftrightarrow C = \bar{k} - k_0.$$

Thus, we obtain that

$$k\left(t\right) = \bar{k} + \left(\bar{k} - k_0\right)\exp\left[g'\left(\bar{k}\right)t\right].$$

Using integrating factor We can solve this first-order differential equation using an integrating factor.[29] The integrating factor is then given by

$$I(t) = \exp\left[\int^t -g'\left(\bar{k}\right) \mathrm{d}t\right] = \exp\left[-g'\left(\bar{k}\right) t\right].$$

The solution is then given by

$$\begin{aligned}\exp\left[-g'\left(\bar{k}\right) t\right] k(t) &= \int^t -g'\left(\bar{k}\right) \bar{k} \exp\left[-g'\left(\bar{k}\right) t\right] \mathrm{d}t + c \\ &= \bar{k} \exp\left[-g'\left(\bar{k}\right) t\right] + c.\end{aligned}$$

Initial condition is that $k(0) = k_0$. Notice that if $t = 0$, then above expression simplifies to

$$k_0 = \bar{k} + c \Rightarrow c = k_0 - \bar{k}.$$

This gives us a linearised version of the law of motion for capital:[30]

$$\begin{aligned}\exp\left[-g'\left(\bar{k}\right) t\right] k(t) &= \bar{k} \exp\left[-g'\left(\bar{k}\right) t\right] + \left(k_0 - \bar{k}\right) \\ \Rightarrow k(t) &= \bar{k} + \left(k_0 - \bar{k}\right) \exp\left[g'\left(\bar{k}\right) t\right].\end{aligned}$$

Thus, the if $g'\left(\bar{k}\right) < 0$, then $k(t) \to \bar{k}$. Moreover, the more negative $g'\left(\bar{k}\right)$, the faster is the convergence.

We can also solve the first-order differential equation,

$$\dot{k} = g'\left(\bar{k}\right)\left(k - \bar{k}\right) = g'\left(\bar{k}\right) k - g'\left(\bar{k}\right) \bar{k}$$

by the usual "guess and verify" method. Guess that the solution is given by

$$k(t) = \exp\left[g'\left(\bar{k}\right) t\right] c(t),$$

where $c(t)$ is the constant of variation. We also need the boundary condition, which in this case is the initial capital level: $k(0) = k_0$. Then,

$$\begin{aligned}\dot{k} &= g'\left(\bar{k}\right) \exp\left[g'\left(\bar{k}\right) t\right] c(t) + \dot{c}(t) \exp\left[g'\left(\bar{k}\right) t\right] \\ &= g'\left(\bar{k}\right) k(t) + \dot{c}(t) \exp\left[g'\left(\bar{k}\right) t\right].\end{aligned}$$

[29] Recall that if we have a differential equation of the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} + P(t) y = Q(t),$$

then we can define the integrating factor to be $I(t)$ such that $I'(t) = P(t) I(t)$. Multiplying both sides by $I(t)$ gives

$$\begin{aligned}I(t) \frac{\mathrm{d}y}{\mathrm{d}t} + P(t) I(t) y = Q(t) I(t) \Rightarrow \frac{\mathrm{d}I(t) y}{\mathrm{d}t} &= Q(t) I(t) \\ \Rightarrow I(t) y &= \int^t Q(t) I(t) \mathrm{d}t + c,\end{aligned}$$

where c depends on the initial condition. To find $I(t)$:

$$\frac{I'(t)}{I(t)} = P(t) \Rightarrow \frac{\mathrm{d} \ln(I(t))}{\mathrm{d}t} = P(t) \Rightarrow \ln(I(t)) = \int^t P(t) \mathrm{d}t \Rightarrow I(t) = \exp\left[\int^t P(t) \mathrm{d}t\right].$$

[30] To verify this, differentiating the expression with respect to t yields

$$\dot{k}(t) = g'\left(\bar{k}\right)\left[k(0) - \bar{k}\right] \exp\left[g'\left(\bar{k}\right) t\right] = g'\left(\bar{k}\right)\left(k(t) - \bar{k}\right).$$

Thus, we need

$$\begin{aligned}
\dot{c}(t)\exp\left[g'\left(\bar{k}\right)t\right] &= -g'\left(\bar{k}\right)\bar{k} \\
\Rightarrow \dot{c}(t) &= -\exp\left[-g'\left(\bar{k}\right)t\right]g'\left(\bar{k}\right)\bar{k} \\
\Rightarrow c(t) = \int \dot{c}(t)\,\mathrm{d}t &= \int -\exp\left[-g'\left(\bar{k}\right)t\right]g'\left(\bar{k}\right)\bar{k}\mathrm{d}t \\
&= -g'\left(\bar{k}\right)\bar{k}\int \exp\left[-g'\left(\bar{k}\right)t\right]\mathrm{d}t \\
&= -g'\left(\bar{k}\right)\bar{k}\left(-\frac{1}{g'\left(\bar{k}\right)}\exp\left[-g'\left(\bar{k}\right)t\right]+C\right) \\
&= \bar{k}\exp\left[-g'\left(\bar{k}\right)t\right]-g'\left(\bar{k}\right)\bar{k}C.
\end{aligned}$$

We therefore have that

$$\begin{aligned}
k(t) &= \exp\left[g'\left(\bar{k}\right)t\right]\left(\bar{k}\exp\left[-g'\left(\bar{k}\right)t\right]-g'\left(\bar{k}\right)\bar{k}C\right) \\
&= \bar{k}-\exp\left[g'\left(\bar{k}\right)t\right]g'\left(\bar{k}\right)\bar{k}C.
\end{aligned}$$

We solve for C using the boundary condition.

$$\begin{aligned}
k_0 &= \bar{k}-\exp\left[g'\left(\bar{k}\right)\cdot 0\right]g'\left(\bar{k}\right)\bar{k}C \\
&= \bar{k}-g'\left(\bar{k}\right)\bar{k}C \\
\Rightarrow C &= \frac{\bar{k}-k_0}{g'\left(\bar{k}\right)\bar{k}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
k(t) &= \bar{k}-\exp\left[g'\left(\bar{k}\right)t\right]g'\left(\bar{k}\right)\bar{k}\frac{\bar{k}-k_0}{g'\left(\bar{k}\right)\bar{k}} \\
&= \bar{k}+\left(k_0-\bar{k}\right)\exp\left[g'\left(\bar{k}\right)t\right]
\end{aligned}$$

as we had before.

Remark 22. (*Half life*) Let us define τ as the time τ that it takes so that the system reaches close to the "half" of the difference between the initial point and the steady state.

$$k(\tau)-\bar{k} = \frac{1}{2}\left(k_0-\bar{k}\right).$$

Substituting for $k(\tau)$ the linearised law of motion for capital yields

$$\begin{aligned}
\bar{k}+\left(k_0-\bar{k}\right)\exp\left[g'\left(\bar{k}\right)\tau\right]-\bar{k} &= \frac{1}{2}\left(k_0-\bar{k}\right) \\
\Rightarrow \exp\left[g'\left(\bar{k}\right)\tau\right] &= \frac{1}{2} \\
\Rightarrow \tau &= -\frac{\ln 2}{g'\left(\bar{k}\right)}.
\end{aligned}$$

9.3.6 Neoclassical growth model

Recall that

$$F\left(k,\dot{k}\right) = U\left(f(k)-\delta k-\dot{k}\right)$$

so that

$$\begin{aligned}
F_k\left(k,\dot{k}\right) &= \left(f'\left(k\right)-\delta\right)U', \\
F_{\dot{k}}\left(k,\dot{k}\right) &= -U', \\
F_{kk}\left(k,\dot{k}\right) &= -\left(f'\left(k\right)-\delta\right)U'', \\
F_{k\dot{k}}\left(k,\dot{k}\right) &= U'',
\end{aligned}$$

where we add one extra derivative

$$F_{kk} = f''\left(k\right)U' + \left(f'\left(k\right)-\delta\right)^2 U''.$$

Evaluating them at the steady state $k=\bar{k}$ using the fact that $\rho = f'\left(\bar{k}\right)-\delta$ gives

$$\begin{aligned}
F_k\left(\bar{k},0\right) &= \left(f'\left(\bar{k}\right)-\delta\right)U' = \rho U', \\
F_{\dot{k}}\left(\bar{k},0\right) &= -U', \\
F_{kk}\left(\bar{k},0\right) &= -\rho U'', \\
F_{k\dot{k}}\left(\bar{k},0\right) &= U'', \\
F_{kk}\left(\bar{k},0\right) &= f''\left(\bar{k}\right)U' + \rho^2 U''.
\end{aligned}$$

The quadratic equation, $Q(\lambda)$, is given by

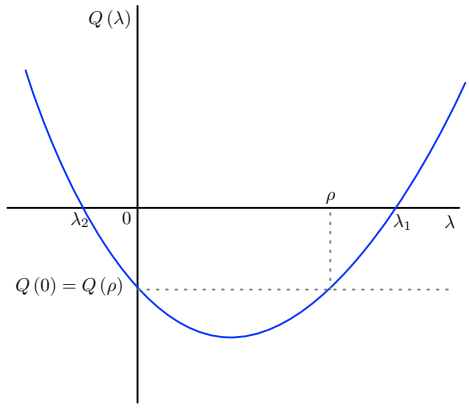
$$\begin{aligned}
Q\left(\lambda\right) &= \left(-U''\right)\lambda^2 + \left(\rho U''\right)\lambda + \left(f''U' + \rho^2 U'' - \rho^2 U''\right) \\
&= \left(-U''\right)\lambda^2 + \left(\rho U''\right)\lambda + \left(f''U'\right) \\
&= \left(-U''\right)\left[\lambda^2 - \rho\lambda - \frac{U'}{U''}f''\right] \\
&= \left(-U''\right)\left[\lambda^2 - \rho\lambda - \frac{U'}{U''}f''\left(\frac{f'}{f'}\frac{c}{c}\frac{k}{k}\right)\right] \\
&= \left(-U''\right)\left[\lambda^2 - \rho\lambda - \left(\frac{cf'}{k}\right)\left(-\frac{U''}{U'}c\right)^{-1}\left(-\frac{f''}{f'}k\right)\right].
\end{aligned}$$

Notice that $-\frac{U''}{U'}c$ is the elasticity of marginal utility (or, the reciprocal of the elasticity of intertemporal substitution) and $-\frac{f''}{f'}k$ is the elasticity of marginal productivity.

Now we check that one root is negative, say λ_2, and the other is then positive and larger than ρ. To do this, note that,

$$\begin{aligned}
Q\left(0\right) &= \left(-U''\right)\left[-\left(\frac{cf'}{k}\right)\left(-\frac{U''}{U'}c\right)^{-1}\left(-\frac{f''}{f'}k\right)\right] < 0 \because U'' < 0, \\
Q\left(\rho\right) &= Q\left(0\right), \\
\frac{\partial Q\left(\lambda\right)}{\partial\lambda} &= -U''\left(2\lambda-\rho\right), \\
\frac{\partial^2 Q\left(\lambda\right)}{\partial\lambda^2} &= -2U'' > 0.
\end{aligned}$$

Figure below plots $Q\left(\lambda\right)$, which visually shows that one root is negative and the other is positive and larger than ρ.



We can, in fact, solve for the roots using the quadratic formula:

$$\lambda = g'(k) = \frac{\rho \pm \sqrt{\rho^2 - 4\left(\frac{cf'}{k}\right)\left(-\frac{U''}{U'}c\right)^{-1}\left(-\frac{f''}{f'}k\right)}}{2},$$

and the negative is root is given by when $\pm$ is $-$.

Example 33. Let us impose some functional forms for the production function and the period utility:

$$f(k) = Ak^{\alpha},$$
$$u(c) = \frac{c^{1-\gamma} - 1}{1-\gamma}.$$

The steady steady equations are:

$$f'\left(\bar{k}\right) = \rho + \delta \Rightarrow \alpha A\bar{k}^{\alpha-1} = \rho + \delta.$$
$$\Rightarrow A\bar{k}^{\alpha} = \frac{\rho+\delta}{\alpha}\bar{k}$$

We can obtain steady-state consumption from the law of motion for capital ($\dot{k}_t = G(k_t, 1) - \delta k_t - c_t$):

$$\bar{c} = A\bar{k}^{\alpha} - \delta\bar{k}.$$

Combining the two expressions, we can write

$$\bar{c} = \frac{(\rho+\delta)}{\alpha}\bar{k} - \delta\bar{k}$$
$$= \left(\frac{\rho + (1-\alpha)\delta}{\alpha}\right)\bar{k}.$$

The elasticities are

$$-\frac{f''}{f'}k = -\frac{(\alpha-1)\alpha A\bar{k}^{\alpha-2}}{\alpha A\bar{k}^{\alpha-1}}k = (1-\alpha),$$
$$-\frac{u''}{u'}c = -\frac{-\gamma c^{-\gamma-1}}{c^{-\gamma}}c = \gamma.$$

Therefore,

$$\left(\frac{cf'}{k}\right)\left(-\frac{U''}{U'}c\right)^{-1}\left(-\frac{f''}{f'}k\right)=\left(\frac{\rho+(1-\alpha)\delta}{\alpha}\right)(\rho+\delta)\frac{1-\alpha}{\gamma}\geq 0.$$

Hence, indeed,

$$Q(0)=Q(\rho)=(-U'')\left[-\left(\frac{\rho+(1-\alpha)\delta}{\alpha}\right)(\rho+\delta)\frac{1-\alpha}{\gamma}\right]<0$$

so that one root is negative.

9.4 Stability of continuous-time linear dynamic systems of higher dimensions

In continuous time and with higher dimensions, we would have

$$\dot{\mathbf{x}}(t)=m(\mathbf{x}(t)),\ \forall t\geq 0.$$

At the steady state,

$$0=m(\overline{\mathbf{x}}).$$

Linear approximation of $\dot{\mathbf{x}}(t)$ yields

$$\begin{aligned}\dot{\mathbf{x}}(t)=m(\mathbf{x}(t))&\simeq m(\overline{\mathbf{x}})+m'(\overline{\mathbf{x}})(\mathbf{x}(t)-\overline{\mathbf{x}})\\&\simeq m'(\overline{\mathbf{x}})(\mathbf{x}(t)-\overline{\mathbf{x}}).\end{aligned}$$

Define $A\coloneqq m'(\overline{\mathbf{x}})$ and $\mathbf{y}(t)\coloneqq\mathbf{x}(t)-\overline{\mathbf{x}}$ so that we may write

$$\dot{\mathbf{y}}(t)=A\mathbf{y}(t).$$

As before, we can diagonalise A as $A=P\Lambda P^{-1}$ and define $\mathbf{z}(t)\coloneqq P^{-1}\mathbf{y}(t)$ to obtain

$$\dot{\mathbf{z}}(t)=\Lambda\mathbf{z}(t),\ \forall t\geq 0,$$

which can be written as

$$\dot{z}_i(t)=\lambda_i z_i(t),\ \forall i=1,2\ldots,n,\ \forall t\geq 0,.$$

This has the solution:[31]

$$z_i(t)=e^{\lambda_i t}z_i(0),\ \forall i=1,2\ldots,n,\ \forall t\geq 0.$$

[31] To see this, rearrange

$$\begin{aligned}\lambda_i=\frac{\dot{z}_i(t)}{z_i(t)}&=\frac{\mathrm{d}}{\mathrm{d}t}\left[\ln z_i(t)\right]\\\Rightarrow\int\lambda_i\mathrm{d}t=\lambda_i t&=\ln z_i(t)+C\\\Rightarrow z_i(t)&=e^{\lambda_i t}e^{-C}.\end{aligned}$$

Given the initial condition $z_i(0)$, $e^{-C}=z_i(0)$.

Proposition 100. *Let λ_i be such that for $i = 1, 2, \ldots, m$, we have $\lambda_i < 0$ and for $i = m+1, m+2, \ldots, n$, we have $\lambda_i \geq 0$. Thus, the eigenvalues of A are ordered so that the first m are negative. Consider the sequence*

$$\dot{\mathbf{x}} = A\left(\mathbf{x}\left(t\right) - \overline{\mathbf{x}}\right), \ \forall t \geq 0$$

for some initial condition $\mathbf{x}(0)$. Then

$$\lim_{t \to \infty} \mathbf{x}\left(t\right) = \overline{\mathbf{x}}$$

if and only if the initial condition $\mathbf{x}(0)$ satisfies

$$\mathbf{x}\left(0\right) = P\hat{\mathbf{z}}\left(0\right) + \overline{\mathbf{x}},$$

where $\hat{\mathbf{z}}(0)$ is a vector with its $n - m$ last coordinates equal to zero; i.e.

$$\hat{z}_i\left(0\right) = 0, \ \forall i = m+1, m+2, \ldots, n$$

and where the remaining elements of $\hat{\mathbf{z}}(0)$ are arbitrary.

Proof. Recall that

$$z_i\left(t\right) = e^{\lambda_i t} z_i\left(0\right), \ \forall i = 1, 2 \ldots, n, \ \forall t \geq 0.$$

For any $\lambda_i > 0$ with positive initial value, the sequence is exploding. Hence, in order for the system to converge, it must be that, for all i with $\lambda_i > 0$, $z_i\left(0\right) = 0$. Recall that

$$\begin{aligned} \mathbf{z}\left(t\right) = P^{-1}\mathbf{y}\left(t\right) = P^{-1}\left(\mathbf{x}\left(t\right) - \overline{\mathbf{x}}\right) &\Rightarrow P\mathbf{z}\left(t\right) = \left(\mathbf{x}\left(t\right) - \overline{\mathbf{x}}\right) \\ &\Rightarrow \mathbf{x}\left(t\right) = P\mathbf{z}\left(t\right) + \overline{\mathbf{x}} \\ &\Rightarrow \mathbf{x}\left(0\right) = P\mathbf{z}\left(0\right) + \overline{\mathbf{x}}. \end{aligned}$$

Therefore, it must be the case that $\mathbf{x}\left(0\right) = P\hat{\mathbf{z}}\left(0\right) + \overline{\mathbf{x}}$ for the system to converge. ■

Remark. Once again, if λ_i can be complex, we only consider the real part of the complex root and the conditions are the same with respect to the real part.

9.5 Saddle path for linearised dynamics

Consider a continuous-time problem with state x, controls u, objective function h, law of motion g and discount factor ρ as done in the notes before. The Hamiltonian is

$$H\left(x, u, \lambda\right) = h\left(x, u\right) + \lambda g\left(x, u\right),$$

where λ is the co-state variable. Let $x, u, \lambda \in \mathbb{R}$. The first-order conditions for the Hamiltonian are:

$$\begin{aligned} 0 &= H_u\left(x, u, \lambda\right), \\ \dot{\lambda} &= \rho\lambda - H_x\left(x, u, \lambda\right), \\ \dot{x} &= g\left(x, u\right). \end{aligned}$$

Using $H_u = 0$ to solve for $u = \mu\left(x, \lambda\right)$ as a function of x and λ, we get

$$H_u\left(x, \mu\left(x, \lambda\right), \lambda\right) = 0.$$

We then obtain the two dimensional dynamic system:

$$\begin{aligned}\dot{\lambda} &= \rho\lambda - H_x\left(x, \mu\left(x,\lambda\right),\lambda\right),\\ \dot{x} &= g\left(x,\mu\left(x,\lambda\right)\right).\end{aligned}$$

Linearising this system, we obtain

$$\left[\begin{array}{c}\frac{\mathrm{d}\hat{\lambda}}{\mathrm{d}t}\\ \frac{\mathrm{d}\hat{x}}{\mathrm{d}t}\end{array}\right] = \left[\begin{array}{cc}a_{11} & a_{12}\\ a_{21} & a_{22}\end{array}\right]\left[\begin{array}{c}\hat{\lambda}\left(t\right)\\ \hat{x}\left(t\right)\end{array}\right],$$

where $\hat{\lambda}(t) = \lambda(t) - \lambda^*$ and $\hat{x}(t) = x(t) - \overline{x}$.

The matrix $A = [a_{ij}]$ as the derivatives of the two-dimensional system displayed above evaluated at the steady-state values. Using the previous result, we can write

$$\left[\begin{array}{c}\hat{\lambda}\left(t\right)\\ \hat{x}\left(t\right)\end{array}\right] = P\left[\begin{array}{c}z_1\left(t\right)\\ z_2\left(t\right)\end{array}\right] = P\left[\begin{array}{c}z_1\left(0\right)e^{\theta_1 t}\\ z_2\left(0\right)e^{\theta_2 t}\end{array}\right],$$

where

$$A = P\Theta P^{-1}$$

and Θ is a diagonal matrix with the eigenvalues of A, denoted by θ_i in its diagonal.

Let us assume that $\theta_1 < 0 < \theta_2$. For the system to converge to the steady state, we must have $z_2(0) = z_2(t) = 0$ for all t, where z is given by

$$\left[\begin{array}{c}z_1\left(t\right)\\ z_2\left(t\right)\end{array}\right] = P^{-1}\left[\begin{array}{c}\hat{\lambda}\left(t\right)\\ \hat{x}\left(t\right)\end{array}\right].$$

Then,

$$\left[\begin{array}{c}z_1\left(0\right)e^{\theta_1 t}\\ 0\end{array}\right] = \left[\begin{array}{cc}x_{11} & x_{12}\\ x_{21} & x_{22}\end{array}\right]\left[\begin{array}{c}\hat{\lambda}\left(t\right)\\ \hat{x}\left(t\right)\end{array}\right] \quad \forall t \geq 0$$

or

$$x_{21}\hat{\lambda}\left(t\right) + x_{22}\hat{x}\left(t\right) = 0.$$

This defines the saddle path as

$$\begin{aligned}\hat{\lambda}\left(t\right) &= -\frac{x_{22}}{x_{21}}\hat{x}\left(t\right),\\ \Rightarrow \lambda\left(t\right) &= \lambda^* - \frac{x_{22}}{x_{21}}\left(x\left(t\right) - \overline{x}\right).\end{aligned}$$

9.6 Slope of the saddle path and of the optimal decision rule (continuous-time, one-dimensional case)

Let the function ϕ satisfying $\lambda = \phi(x)$ be the saddle path. We want to compute

$$\frac{\mathrm{d}\lambda}{\mathrm{d}x} = \phi'\left(\overline{x}\right),$$

where $\overline{x}$ is the steady state. Notice that

$$\phi' = \frac{\mathrm{d}\lambda}{\mathrm{d}x} = \frac{\mathrm{d}\lambda/\mathrm{d}t}{\mathrm{d}x/\mathrm{d}t} \equiv \frac{\dot{\lambda}\left(\lambda,x\right)}{\dot{x}\left(\lambda,x\right)} = \frac{\rho\lambda - H_x\left(x,\mu\left(x,\lambda\right),\lambda\right)}{g\left(x,\mu\left(x,\lambda\right)\right)}$$

At the steady state, $\dot{\lambda} = \dot{x} = 0$ which implies that $\phi' =$ "0/0". Thus, we use L'Hôpital's rule to find this ratio.

$$\begin{aligned}\lim_{x\to\overline{x}} \phi'(x) \equiv \left.\frac{\mathrm{d}\lambda}{\mathrm{d}x}\right|_{x=\overline{x}} &= \left.\frac{\dot{\lambda}(\lambda,x)}{\dot{x}(\lambda,x)}\right|_{x=\overline{x}} \\ &= \left.\frac{\frac{\mathrm{d}}{\mathrm{d}x}\dot{\lambda}(\lambda,x)}{\frac{\mathrm{d}}{\mathrm{d}x}\dot{x}(\lambda,x)}\right|_{x=\overline{x}} = \left.\frac{\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}x}+\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}x}}{\frac{\mathrm{d}\dot{x}}{\mathrm{d}\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}x}+\frac{\mathrm{d}\dot{x}}{\mathrm{d}x}}\right|_{x=\overline{x}} \\ &= \left.\frac{\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda}\phi'+\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}x}}{\frac{\mathrm{d}\dot{x}}{\mathrm{d}\lambda}\phi'+\frac{\mathrm{d}\dot{x}}{\mathrm{d}x}}\right|_{x=\overline{x}}\end{aligned}$$

where

$$\begin{aligned}\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda} &= \rho - H_{xu}\mu_\lambda - H_{x\lambda}, \\ \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}x} &= -H_{xx} - H_{xu}\mu_x, \\ \frac{\mathrm{d}\dot{x}}{\mathrm{d}\lambda} &= g_u\mu_\lambda, \\ \frac{\mathrm{d}\dot{x}}{\mathrm{d}x} &= g_x + g_u\mu_x, \\ 0 &= H_{ux} + H_{uu}\mu_x, \\ 0 &= H_{u\lambda} + H_{uu}\mu_\lambda.\end{aligned}$$

Note that the above comes from differentiating the first-order conditions:

$$\begin{aligned}\dot{\lambda} &= \rho\lambda - H_x\left(x,\mu(x,\lambda),\lambda\right), \\ \dot{x} &= g\left(x,\mu(x,\lambda)\right), \\ 0 &= H_u\left(x,\mu(x,\lambda),\lambda\right).\end{aligned}$$

We therefore have the following quadratic form for ϕ':

$$\begin{aligned}\phi'\left(\frac{\mathrm{d}\dot{x}}{\mathrm{d}\lambda}\phi' + \frac{\mathrm{d}\dot{x}}{\mathrm{d}x}\right) &= \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda}\phi' + \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}x} \\ \Rightarrow (\phi')^2\frac{\mathrm{d}\dot{x}}{\mathrm{d}\lambda} + \phi\left(\frac{\mathrm{d}\dot{x}}{\mathrm{d}x} - \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda}\right) - \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}x} &= 0.\end{aligned}$$

One of the two roots of this quadratic equation is the slope of the saddle path. Unlike when we were analysing convergence, there is no "rule" as to which root will represent the stable steady state. Thus, when using this method, we would usually plot the phase diagram to know the slope of the saddle path from the diagram, and pick the appropriate root using the equation above accordingly.

We can also find the slope of the optimal control rule setting the control as a function of the state. Let

$$u^*(x) = \mu\left(x,\phi(x)\right).$$

Then

$$\left.\frac{\mathrm{d}u^*(x)}{\mathrm{d}x}\right|_{x=\overline{x}} = \mu_x + \mu_\lambda\phi'.$$

Finally, we can find the rate of change of the state as a function of the state (i.e. the speed of

convergence, $g'(x)$):

$$\begin{aligned}\left.\frac{\mathrm{d}\dot{x}}{\mathrm{d}x}\right|_{x=\overline{x}} &= g_x + g_u \left.\frac{\mathrm{d}u^*(\overline{x})}{\mathrm{d}x}\right|_{x=\overline{x}} \\ &= g_x + g_u(\mu_x + \mu_\lambda \phi').\end{aligned}$$

Example 34. (*Neoclassical growth model*) Recall that

$$\begin{aligned}\dot{\lambda} &= \lambda(\rho - (f'(k) - \delta)) \\ \dot{k} &= f(k) - \delta k - c(k).\end{aligned}$$

Since $U'(c) = \lambda$, we can write

$$\dot{k} = f(k) - \delta k - (U')^{-1}(\lambda).$$

Then

$$\phi'(k) \equiv \frac{\mathrm{d}\lambda}{\mathrm{d}k} = \frac{\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda}\phi' + \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}k}}{\frac{\mathrm{d}\dot{k}}{\mathrm{d}\lambda}\phi' + \frac{\mathrm{d}\dot{k}}{\mathrm{d}k}},$$

where

$$\begin{aligned}\frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}\lambda} &= \rho - (f'(k) - \delta) \\ \frac{\mathrm{d}\dot{\lambda}}{\mathrm{d}k} &= \frac{\mathrm{d}\lambda}{\mathrm{d}k}(\rho - f'(k) + \delta) - \lambda f''(k) = \phi'(k)(\rho - f'(k) + \delta) - \lambda f''(k) \\ \frac{\mathrm{d}\dot{k}}{\mathrm{d}\lambda} &= -\frac{\mathrm{d}(U')^{-1}}{\mathrm{d}\lambda} = -\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda} \\ \frac{\mathrm{d}\dot{k}}{\mathrm{d}k} &= f''(k) - \delta - \frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}k} = f''(k) - \delta - \frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\phi'(k).\end{aligned}$$

Hence,

$$\phi'\left(\bar{k}\right) = \frac{\left(\rho - f'\left(\bar{k}\right) - \delta\right)\phi'\left(\bar{k}\right) + \phi'\left(\bar{k}\right)\left(\rho - f'\left(\bar{k}\right) + \delta\right) - \lambda f''\left(\bar{k}\right)}{-\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\phi'\left(\bar{k}\right) + f''\left(\bar{k}\right) - \delta - \frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\phi'\left(\bar{k}\right)}.$$

In the steady state, $\rho = f'(\bar{k}) - \delta$, so we can simplify above to

$$\phi'\left(\bar{k}\right) = \frac{\delta\phi'\left(\bar{k}\right) + \lambda f''\left(\bar{k}\right)}{2\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\phi'\left(\bar{k}\right) - f''\left(\bar{k}\right) + \delta}$$

and so

$$2\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\left(\phi'\left(\bar{k}\right)\right)^2 - \left(f''\left(\bar{k}\right) + \delta\right)\phi'\left(\bar{k}\right) + \left(\delta - \lambda f''\left(\bar{k}\right)\right) = 0$$

This is a quadratic equation in $\phi'(k)$ which we can solves to obtain the slope of the saddle path in (k, λ) space.

$$\phi'\left(\bar{k}\right) = \frac{f''\left(\bar{k}\right) + \delta \pm \sqrt{\left(f''\left(\bar{k}\right) + \delta\right)^2 - 8\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\left(\delta - \lambda f''\left(\bar{k}\right)\right)}}{4\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}}.$$

Recall from section 8.3.3 when we drew the phase diagram that the slope of the saddle path is negative in (k, λ) space. Since $U'' < 0$, it follows that $\mathrm{d}(U')^{-1}(\lambda)/\mathrm{d}\lambda < 0$ and so the negative root

is given by when $\pm$ is $+$.[32] That is,

$$\phi'\left(\bar{k}\right)=\frac{f''\left(\bar{k}\right)+\delta+\sqrt{\left(f''\left(\bar{k}\right)+\delta\right)^{2}+8\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}\lambda f''\left(\bar{k}\right)}}{4\frac{\mathrm{d}(U')^{-1}(\lambda)}{\mathrm{d}\lambda}}.$$

We can also obtain the slope of the saddle path in (k,c) space. Recall that

$$\begin{aligned}\frac{\dot{c}}{c}&=\frac{1}{-\frac{U''(c)c}{U'(c)}}\left(f'\left(k\right)-\delta-\rho\right),\\ \dot{k}&=f\left(k\right)-\delta k-c.\end{aligned}$$

So the slope of the saddle path is

$$\tilde{\phi}'\left(\bar{k}\right)=\left.\frac{\mathrm{d}c}{\mathrm{d}k}\right|_{k=\bar{k}}=\left.\frac{\dot{c}}{\dot{k}}\right|_{k=\bar{k}}=\frac{\frac{1}{-\frac{u''}{u'}}\left(f'\left(k\right)-\delta-\rho\right)}{f\left(k\right)-\delta k-\phi\left(k\right)}.$$

Using L'Hôpital's rule and evaluating at the steady state

$$\tilde{\phi}'\left(\bar{k}\right)=\frac{\frac{1}{-\frac{u''}{u'}}\left(f''\left(\bar{k}\right)\right)+\overbrace{\left(f'\left(\bar{k}\right)-\delta-\rho\right)}^{=0}\frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{1}{-\frac{u''}{u'}}\right)}{\underbrace{f'\left(k\right)-\delta}_{=\rho}-\phi'\left(\bar{k}\right)}.$$

Hence,

$$\left(\rho-\tilde{\phi}'\left(\bar{k}\right)\right)\tilde{\phi}'\left(\bar{k}\right)=\frac{f''\left(\bar{k}\right)}{-\frac{u''}{u'}}\frac{-\left(f'\left(\bar{k}\right)-\delta\right)}{-\left(f'\left(\bar{k}\right)-\delta\right)}=-\frac{\frac{-f''}{f'-\delta}}{-\frac{u''}{u'}}\rho\equiv-b$$

where we used the fact that $f'=\rho$ in the steady state and $b>0$. The quadratic equation is then

$$\tilde{\phi}'\left(\bar{k}\right)^{2}-\rho\tilde{\phi}\left(\bar{k}\right)-b=0.$$

Letting $\gamma\coloneqq\tilde{\phi}'(\bar{k})$,

$$Q\left(\gamma\right)\coloneqq\gamma^{2}-\rho\gamma-b.$$

Observe that

$$\begin{aligned}Q\left(0\right)&=-b<0,\\ Q\left(\rho\right)&=-b<0\\ \left.\frac{\mathrm{d}Q}{\mathrm{d}\gamma}\right|_{\gamma=0}&=-\rho<0,\\ \frac{\mathrm{d}^{2}Q}{\mathrm{d}\gamma^{2}}&=2>0.\end{aligned}$$

Hence, $Q(\gamma)$ has a minimum between $\gamma\in(0,\rho)$, is strictly convex and is downward sloping when $\gamma=0$. So, the figure looks like the one we drew in section 9.3.6. We realise that one root is positive while the other is negative. Recall from section 8.3.3 that the slope of the saddle path is positive

[32] For example, let $U\left(c\right)=\ln c$. Then $U'\left(c\right)=1/c$ so that

$$1/c=\lambda\Leftrightarrow\lambda=\frac{1}{c};$$

i.e. $\left(U'\right)^{-1}\left(\lambda\right)=1/\lambda$. Then, $\mathrm{d}\left(U'\right)^{-1}\left(\lambda\right)/\mathrm{d}\lambda=-1/\lambda^{2}<0$.

in (k, c) space. So,

$$\tilde{\phi}'\left(\bar{k}\right) = \frac{\rho + \sqrt{\rho^2 + 4b}}{2}$$

Since $U'\left(c\left(k\right)\right) = \lambda\left(k\right)$, we can write

$$c^*\left(k\right) = \left(U'\right)^{-1}\left(\lambda\left(k\right)\right)$$

so that

$$\tilde{\phi}'\left(\bar{k}\right) = \left.\frac{\mathrm{d}c^*\left(k\right)}{\mathrm{d}k}\right|_{k=\bar{k}} = \left.\frac{\mathrm{d}\left(U'\right)^{-1}\left(\lambda\left(k\right)\right)}{\mathrm{d}k}\frac{\mathrm{d}\lambda}{\mathrm{d}k}\right|_{k=\bar{k}} \frac{\mathrm{d}\left(U'\right)^{-1}\left(\lambda\left(\bar{k}\right)\right)}{\mathrm{d}k}\phi'\left(\bar{k}\right).$$

10 Principle of Optimality and Dynamic Programming

Bellman's Principle of Optimality provides conditions under which a programming problem expressed in sequence form is equivalent (as defined below) to a two-period recursive programming problem (called the *Bellman* equation). The first part of the note discusses this relation and introduces some concepts and techniques used to solve dynamic programming problems in discrete time. The last part of the note focuses on continuous-time dynamic programming problems and shows how they relate to the Maximum Principle.

10.1 Principle of Optimality

10.1.1 Recursive problem

Recall that the sequence problem is given by:

$$\begin{aligned} V^*(x_0) \coloneqq \max_{(x_{t+1})_{t=0}^{\infty}} & \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}) \\ \text{s.t.} \quad & x_{t+1} \in \Gamma(x_t) \ \ \forall t \geq 0, \\ & x_0 \text{ given.} \end{aligned}$$

Equivalently, we can write the problem as

$$V^*(x_0) = \max_{x^\infty \in \Pi^\infty(x_0)} u((x_{t+1})), \tag{10.1}$$

where

$$\begin{aligned} x^t &\coloneqq (x_0, x_1, \ldots, x_t), \\ \Pi^t(x_0) &\coloneqq \left\{x^t : x_{s+1} \in \Gamma(x_s), s = 0, 1, \ldots, t-1\right\}, \\ u((x_{t+1})) &= \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t F(x_t, x_{t+1}). \end{aligned}$$

We think of x^t as the set of states from period 0 to t, $\Pi^t(x_0)$ as the set of all possible states from period 0 to period t (thus, $\Pi^t(x_0)$ is an increasing set as t increases), and $u((x_{t+1}))$ is the discounted (infinite) sum of the utility.

10.1.2 Bellman equation

The Bellman equation, which is a recursive problem, is to find a function $V : X \to \mathbb{R}$ such that

$$V(x) = \max_{y\in\Gamma(x)} \left[F(x, y) + \beta V(y)\right] \ \ \forall x \in X. \tag{10.2}$$

This is a *functional* equation in which the solution is a function V that must satisfy the equation above for all $x \in X$. Let $g(x)$ denote the maximiser of the right-hand side of (10.2). Then, $g(x)$ satisfies

$$V(x) = F(x, g(x)) + \beta V(g(x)).$$

If the function V were known (or if we know properties of it), then (10.2) is a two-period problem.

10.1.3 Principle of Optimality

The *Principle of Optimality* states that

$$V^*(x) = V(x) \ \ \forall x \in X.$$

That is, the solution to the two-period problem in (10.2) is equivalent to the infinite-dimension problem in (10.1). Once we have V and g, then we have the solution of (10.1) for any initial condition $x_0 \in X$.

Theorem 22. *Suppose $\Gamma(x)$ is nonempty for all $x \in X$ and that, for all $x_0 \in X$ and all $x^\infty \in \Pi^\infty(x_0)$, $\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t F(x_t, x_{t+1})$ exists in $\overline{\mathbb{R}}$. Then, V^* satisfies (10.2). Moreover, if V is a solution to (10.2) and*

$$\lim_{T\to\infty} \beta^T v(x_T) = 0 \ \forall x^\infty \in \Pi^\infty(x_0) \ \ \forall x_0 \in X,$$

then $V^* = V$.

Here, we sketch the basic reasoning behind the Principle of Optimality (this is not a rigorous proof!). Take the case where F is bounded so $|F(x,y)| \leq B < \infty$ for all $(y,x) \in \text{gr}(\Gamma)$. Notice that, for any $x^\infty \in \Pi^\infty(x_0)$, since $\beta \in (0,1)$,

$$\lim_{T\to\infty} \sum_{t=0}^{T} \beta^t |F(x_t, x_{t+1})| \leq \lim_{T\to\infty} \sum_{t=0}^{T} \beta^t B = \frac{B}{1-\beta}.$$

That is,

$$|u((x_{t+1}))| \leq \frac{B}{1-\beta}.$$

Splitting the infinite sum into the first $T-1$ periods and the remaining periods:

$$u((x_{t+1})) = \sum_{t=0}^{T-1} \beta^t F(x_t, x_{t+1}) + \beta^T \left[\sum_{t=T}^{\infty} \beta^{t-T} F(x_t, x_{t+1}) \right].$$

Then, we can write

$$\left| u((x_{t+1})) - \sum_{t=0}^{T-1} \beta^t F(x_t, x_{t+1}) \right| = \beta^T \left| \sum_{t=T}^{\infty} \beta^{t-T} F(x_t, x_{t+1}) \right| \leq \beta^T \frac{B}{1-\beta}.$$

Since β^T can be made arbitrarily small by choosing T sufficiently large, we can approximate $u((x_{t+1}))$—i.e. the value of a plan—by $\sum_{t=0}^{T-1} \beta^t F(x_t, x_{t+1})$ for sufficiently large T.

By definition, (10.2) means that $V(x_0)$ is the maximum value such that

$$V(x_0) \geq F(x_0, x_1) + \beta V(x_1), \ \forall (x_0, x_1) \in \Pi^1(x_0) \tag{10.3}$$

and for some $(x_0, x_1^*) \in \Pi^1(x_0)$, the above holds with equality; i.e.

$$V(x_0) = F(x_0, x_1^*) + \beta V(x_1^*).$$

Since (10.2) holds for all $x \in X$, it must also be the case that

$$V(x_1^*) \geq F(x_1^*, x_2) + \beta V(x_2), \ \forall (x_0, x_1^*, x_2) \in \Pi^2(x_0).$$

Substituting into (10.3),

$$\begin{aligned} V(x_0) &\geq F(x_0, x_1^*) + \beta V(x_1^*) \\ &\geq F(x_0, x_1^*) + \beta \left[F(x_1^*, x_2) + \beta V(x_2)\right] \\ &= F(x_0, x_1^*) + \beta F(x_1^*, x_2) + \beta^2 V(x_2), \ \forall (x_0, x_1^*, x_2) \in \Pi^2(x_0). \end{aligned}$$

Again, there exists some $(x_0, x_1^*, x_2^*) \in \Pi^2(x_0)$ such that

$$V(x_0) = F(x_0, x_1^*) + \beta F(x_1^*, x_2^*) + \beta^2 V(x_2^*).$$

Continuing in a similar fashion, we get

$$V(x_0) \geq \sum_{t=0}^{T-1} \beta^t F(x_t, x_{t+1}) + \beta^T V(x_T) \ \ \forall x^T \in \Pi^T(x_0)$$

and, for some $x^T \in \Pi^T(x_0)$,

$$V(x_0) = \sum_{t=0}^{T-1} \beta^t F(x_t, x_{t+1}) + \beta^T V(x_T). \tag{10.4}$$

Since $|F| \leq B$, $|V| \leq \frac{B}{1-\beta}$ and so

$$\lim_{T \to \infty} \beta^T V(x_T) = 0.$$

Thus, we can approximate the value of $V(x_0)$ arbitrarily well by choosing T sufficiently large; i.e. we conclude that

$$V(x_0) = V^*(x_0).$$

10.2 Bounded Dynamic Programming

We now study the Bellman equation for the case where F is bounded.

Let (S, ρ) be a metric space and let $T : S \to S$ be a self-map. The self-map T is a *contraction with modulus* $\beta \in (0, 1)$ if

$$\rho(Tx, Ty) \leq \beta \rho(x, y) \ \ \forall x, y \in S.$$

Thus, T is a contraction if it "shrinks" the distance between any two points by more than the fraction β.

Remark 23. Let $S = [0, 1]$ so that x and y are scalars between zero and one. Let $\rho(x, y) = |x - y|$ and T be a function. Then, for T to be a contraction with modulus β, it must be that

$$\begin{aligned} &|T(x) - T(y)| \leq \beta |x - y| \\ \Rightarrow &\frac{|T(y) - T(x)|}{|y - x|} \leq \beta. \end{aligned}$$

Hence, this says that the slope of the function T is less than β (and β is the Lipschitz constant). The key here is the following fix-point theorem.

Theorem 23. *(Contraction Mapping Theorem/Banach) If T is a contraction in a metric space (S, ρ) with modulus β, then:*

(i) there is a unique fixed point $s^* \in S$,

$$s^* = T(s^*);$$

(ii) iterations of T converge to the fixed point

$$\rho(T^n s_0, s^*) \leq \beta^n \rho(s_0, s^*), \ \forall s_0 \in S,$$

where $T^{n+1}s = T(T^n s)$.

In our main application, S will be the set of continuous and bounded functions from X to $\mathbb{R}$:

$$S \coloneqq \left\{ f : X \to \mathbb{R}, \ f \text{ is continuous, and } \|f\| \equiv \sup_{x \in X} |f(x)| < \infty \right\}$$

and the metric is the sup norm:

$$\rho(f, g) \equiv \|f - g\| \coloneqq \sup_{x \in X} |f(x) - g(x)|.$$

The following proposition allows us to determine if something is a contraction.

Proposition 101. *(Blackwell Sufficient Conditions) Let S be the space of bounded functions of X and $\|\cdot\|$ the* sup *norm. Let $T : S \to S$. Then, T is a contraction if*

(i) T is monotone; i.e. $Tf(x) \leq Tg(x)$ for any $x \in X$ and g, f such that $f(x) \leq g(x)$ for all $x \in X$;

(ii) T discounts; i.e. there exists $\beta \in (0, 1)$ such that, for any $a \in \mathbb{R}_+$,

$$T(f + a)(x) \leq Tf(x) + a\beta, \ \forall x \in X, f \in S.$$

Proof. First, we can always write

$$f - g = f - g.$$

By the definition of sup norm ($\|f(x)\| = \sup |f(x)|$):

$$\begin{aligned} f - g &\leq \|f - g\| \\ \Rightarrow f &\leq g + \|f - g\|. \end{aligned}$$

Since T is monotone, then

$$Tf \leq T(g + \|f - g\|).$$

Since T discounts, setting $a = \|f - g\|$, we can write

$$\begin{aligned} Tf &\leq T(g + \|f - g\|) \\ &\leq Tg + \beta \|f - g\| \\ \Rightarrow Tf - Tg &\leq \beta \|f - g\|. \end{aligned}$$

Of course, we can reverse the roles of f and g to obtain that

$$\begin{aligned} Tg &\leq Tf + \beta \|g - f\| \\ &= Tf + \beta \|f - g\| \\ \Rightarrow Tg - Tf &\leq \beta \|f - g\|. \end{aligned}$$

Combining both inequalities,

$$\|Tf - Tg\| \leq \beta \|f - g\|;$$

i.e. T is a contraction. ■

We define the *Bellman operator* T as

$$(Tv)(x) = \max_{y \in \Gamma(x)} [F(x,y) + \beta v(y)]. \tag{10.5}$$

Theorem 24. *Assume that F is bounded and continuous and that Γ is continuous and has range that is compact. Let T be the Bellman operator as defined in (10.5). Then, T maps the set of continuous and bounded functions S onto itself. Moreover, T is a contraction.*

Proof. That T maps the set of continuous and bounded follow from the Theorem of the Maximum which we will study in ECON 6701. That T is a contraction follows since T satisfies the Blackwell conditions:

(i) Monotonicity. For $f \geq v$,

$$\begin{aligned}(Tv)(x) &= \max_{y \in \Gamma(x)} [F(x,y) + \beta v(y)] \\ &= F(x, g(x)) + \beta v(g(x)) \\ &\leq F(x, g(x)) + \beta f(g(x)) \because f \geq v \\ &\leq \max_{y \in \Gamma(x)} [F(x,y) + \beta f(y)] = (Tf)(x),\end{aligned}$$

where the last inequality comes from the fact that $Tf(x)$ is the maximum.

(ii) Discounts. For $a > 0$,

$$\begin{aligned}(T(v+a))(x) &= \max_{y \in \Gamma(x)} [F(x,y) + \beta(v(y) + a)] \\ &= \max_{y \in \Gamma(x)} [F(x,y) + \beta v(y)] + \beta a \\ &= (Tv)(x) + \beta a.\end{aligned}$$

■

We mentioned previously that (S, ρ) is a metric space. Moreover, this space is also *complete*; i.e., every Cauchy sequence is convergent.

Corollary 10. *Let S be a complete metric space and $S' \subset S$ be closed. Let T be a contraction on S and $s^* = Ts^*$. If $TS' \subset S'$ (i.e. if $s' \in S'$, then $T(s') \in S'$), then $s^* \in S'$. Moreover, if $S'' \subset S'$ and $TS' \subset S''$ (i.e. if $s' \in S'$, then $T(s') \in S''$), then $s^* \in S''$.*

The first statement of Corollary 10 means that, if applying the operator T to any point in a closed set S' means that the "output" stays within the set, then the fixed point must also be in that set. The second means that, if applying the operator T to any point in S' means that the output is in the set S'', then the fixed point must also be in S''. Note that S'' need not be closed, whereas S' must be closed. Corollary 10 is useful in establishing the properties of the value function V and the optimal policy g.

In what follows, we maintain the following set of assumptions.

Assumption 1. *X is a convex subset of $\mathbb{R}^n$, $\Gamma(x)$ is nonempty, compact for all $x \in X$, Γ is continuous, and F is bounded and continuous, and $\beta \in (0,1)$.*

Theorem 25 (Monotonicity)**.** *In addition to Assumption 1, assume that $F(\cdot, y)$ is increasing and that Γ is increasing (i.e. $\Gamma(x) \subseteq \Gamma(x')$ for $x \leq x'$). Then, the fixed point v^* satisfying $v^* = Tv^*$ is increasing. If $F(\cdot, y)$ is strictly increasing, so is v^*.*

Proof. By Corollary 10, it suffices to show that Tf is increasing if f is increasing. Let $x \leq x'$:

$$\begin{aligned}(Tf)(x) &= \max_{y \in \Gamma(x)} [F(x,y) + \beta f(y)] \\ &= F(x, g(x)) + \beta f(g(x)) \\ &\leq F(x', g(x)) + \beta f(g(x))\end{aligned}$$

since $F(\cdot, y)$ is increasing (the inequality will hold strictly if $F(\cdot, y)$ is strictly increasing). But by definition,

$$\begin{aligned}(Tf)(x) &\leq F(x', g(x)) + \beta f(g(x)) \\ &\leq F(x', g(x')) + \beta f(g(x')) \\ &= \max_{y \in \Gamma(x')} [F(x', y) + \beta f(y)] \\ &= (Tf)(x').\end{aligned}$$

■

Theorem 26 (Concavity)**.** *In addition to Assumption 1, assume and Γ is convex,*[33] *and that F is concave in (x, y). Then, the fixed point v^* satisfying $v^* = Tv^*$ is concave in x. Moreover, if F is strictly concave, so is v^* and the optimal policy correspondence is a single-valued.*

Proof. By Corollary 10, it suffices to show that Tf is concave if f is concave. That is, we wish to show that

$$(Tf)\left(x^\theta\right) \geq (1-\theta)(Tf)(x') + \theta(Tf)(x), \forall \theta \in (0,1).$$

Since F and f are concave then, for any $x, x' \in X$,

$$\begin{aligned}F\left(x^\theta, y^\theta\right) &\geq (1-\theta) F(x', y') + \theta F(x, y), \\ f\left(y^\theta\right) &\geq (1-\theta) f(y') + \theta f(y),\end{aligned}$$

for all $y \in \Gamma(x)$ and $y' \in \Gamma(x')$. Summing the two while multiplying the second by β:

$$\begin{aligned}F\left(x^\theta, y^\theta\right) + \beta f\left(y^\theta\right) &\geq (1-\theta) F(x', y') + \theta F(x, y) + \beta [(1-\theta) f(y') + \theta f(y)] \\ &= (1-\theta) [F(x', y') + \beta f(y')] + \theta [F(x, y) + \beta f(y)]\end{aligned}$$

for all $\theta \in (0,1)$. Above holds when $y = g(x)$ and $y' = g(x')$; i.e.

$$\begin{aligned}F\left(x^\theta, y^\theta\right) + \beta f\left(y^\theta\right) &\geq (1-\theta) [F(x', g(x')) + \beta f(g(x'))] + \theta [F(x, g(x)) + \beta f(g(x))] \\ &= (1-\theta)(Tf)(x') + \theta(Tf)(x).\end{aligned}$$

[33] That is, for any $y \in \Gamma(x)$ and $y' \in \Gamma(x')$, we have

$$\theta y' + (1-\theta) y \in \Gamma\left(\theta x' + (1-\theta) x\right) \ \forall x, x' \in X \ \forall \theta \in (0,1).$$

By definition,

$$\begin{aligned}(Tf)\left(x^{\theta}\right) &= \max_{y\in\Gamma(x^{\theta})}\left[F\left(x^{\theta},y\right)+\beta f\left(y\right)\right]\\ &= F\left(x^{\theta},g\left(x^{\theta}\right)\right)+\beta f\left(g\left(x^{\theta}\right)\right)\\ &\geq F\left(x^{\theta},y^{\theta}\right)+\beta f\left(y^{\theta}\right)\ \ \forall y^{\theta}\in\Gamma\left(x^{\theta}\right).\end{aligned}$$

Hence,

$$(Tf)\left(x^{\theta}\right)\geq(1-\theta)\left(Tf\right)\left(x'\right)+\theta\left(Tf\right)\left(x\right)\ \ \forall\theta\in(0,1).$$ ■

Theorem 27 (Differentiability)**.** *In addition to Assumption 1, assume that $F(x,y)$ is strictly concave in (x,y), Γ is convex, and F is continuously differentiable on* $\text{int}(\text{gr}(\Gamma))$*. If $x_0\in\text{int}(X)$ and $g(x_0)\in\text{int}(\Gamma(x_0))$. Then, the fixed point v^* satisfying $v^*=Tv^*$ is differentiable at x_0 with derivatives given by*

$$v_i\left(x_0\right)=\frac{\partial F}{\partial x_i}\left(x_0,g\left(x_0\right)\right)\ \ \forall i\in\{1,2,\ldots,n\}$$

where $X\subseteq\mathbb{R}^n$.

10.2.1 Envelope: Differentiability of the value function

Let us work out a heuristic argument to find an expression for the derivative of the value function. We will assume that V is differentiable and that the policy function g is also differentiable with respect to x. Assume that $(y,x)\in\text{int}(\text{gr}(\Gamma))$.

First, the first-order condition of the problem

$$\max_{y}F\left(x,y\right)+\beta V\left(y\right)$$

evaluated at the optimum, $y=g(x)$, is

$$F_y\left(x,g\left(x\right)\right)+\beta V'\left(g\left(x\right)\right)=0.$$

Now, differentiating both sides of

$$V\left(x\right)=F\left(x,g\left(x\right)\right)+\beta V\left(g\left(x\right)\right)$$

with respect to x gives

$$\begin{aligned}V'\left(x\right) &= F_x\left(x,g\left(x\right)\right)+F_y\left(x,g\left(x\right)\right)g'\left(x\right)+\beta V'\left(g\left(x\right)\right)g'\left(x\right)\\ &= F_x\left(x,g\left(x\right)\right)+\underbrace{\left(F_y\left(x,g\left(x\right)\right)+\beta V'\left(g\left(x\right)\right)\right)}_{=0\text{ FOC}}g'\left(x\right).\end{aligned}\tag{10.6}$$

Hence,

$$V'\left(x\right)=F_x\left(x,g\left(x\right)\right).$$

This is called the *envelope* condition.

Remark 24. The formal proof (Benveniste and Scheinkman Theorem) requires that V is concave, $F(\cdot,y)\in C^1$ and that $(g(x),x)\in\text{int}(\text{gr}(\Gamma))$. Strictly speaking, the theorem does not require g to be differentiable.

10.2.2 First-order and the envelope conditions

The first-order and the envelope conditions are respectively given by

$$\begin{aligned} 0 &= F_y\left(x, g\left(x\right)\right) + \beta V'\left(g\left(x\right)\right), \\ V'\left(x\right) &= F_x\left(x, g\left(x\right)\right) \end{aligned}$$

for all x such that $(g(x), x) \in \text{int}(\text{gr}(\Gamma))$. Notice that combining the two gives the familiar Euler Equation:

$$0 = F_y\left(x, g\left(x\right)\right) + \beta F_x\left(g\left(x\right), g\left(g\left(x\right)\right)\right).$$

10.2.3 Neoclassical growth model

The Bellman equation for the Neoclassical problem is given by

$$V\left(k\right) = \max_{k' \in [0, f(k)]} \left[U\left(f\left(k\right) - k'\right) + \beta V\left(k'\right)\right].$$

Thus, the first-order condition evaluated at the optimal $k' = g\left(k\right)$ is

$$U'\left(f\left(k\right) - g\left(k\right)\right) = \beta V'\left(g\left(k\right)\right).$$

The envelop condition, evaluated at the optimal, is given by

$$V'\left(\left(k\right)\right) = U'\left(f\left(k\right) - g\left(k\right)\right) f'\left(k\right).$$

Exercise 22. Show that the neoclassical growth model satisfy the assumptions of Theorem *25*.

Exercise 23. Show that the neoclassical growth model satisfy the assumptions of Theorem *26*.

10.3 Continuous-time Bellman equation

Consider the following discrete-time Bellman equation,

$$V\left(x_t\right) = \max_{u_t \in U} \left[\Delta h\left(x_t, u_t\right) + \frac{1}{1 + \Delta\rho} V\left(x_{t+\Delta}\right)\right]$$

subject to

$$x_{t+\Delta} = x_t + \Delta g\left(x_t, u_t\right).$$

We will analyse the continuous-time Bellman equation as a limit of the discrete one.

Notice that, if we simply take the limit as Δ goes to zero, we are simply left with $V(x_t) = V(x_t)$, which is not very useful. Using Taylor expansion (around x_t), we can write

$$\begin{aligned} V\left(x_{t+\Delta}\right) &= V\left(x_t + \Delta g\left(x_t, u_t\right)\right) \\ &= V\left(x_t\right) + V'\left(x_t\right) \Delta g\left(x_t, u_t\right) + o\left(\Delta g\left(x_t, u_t\right)\right), \end{aligned}$$

where $\text{d}(z) = o(z)$ means that $\lim_{z \to 0} \text{d}(z)/z = 0$. Then the Bellman equation is

$$V\left(x_t\right) = \max_{u_t \in U} \left[\Delta h\left(x_t, u_t\right) + \frac{1}{1 + \Delta\rho}\left(V\left(x_t\right) + V'\left(x_t\right) \Delta g\left(x_t, u_t\right) + o\left(\Delta g\left(x_t, u_t\right)\right)\right)\right]$$

Multiplying both sides by the positive constant $1 + \Delta\rho$ yields

$$\left(1 + \Delta\rho\right) V\left(x_t\right) = \max_{u_t \in U} \left[\left(1 + \Delta\rho\right) \Delta h\left(x_t, u_t\right) + V\left(x_t\right) + V'\left(x_t\right) \Delta g\left(x_t, u_t\right) + o\left(\Delta g\left(x_t, u_t\right)\right)\right].$$

We can move $V(x_t)$ inside the max to the left-hand side since it does not depend on u_t,

$$\Delta\rho V(x_t) = \max_{u_t \in U} \left[(1 + \Delta\rho) \Delta h(x_t, u_t) + V'(x_t) \Delta g(x_t, u_t) + o(\Delta g(x_t, u_t)) \right].$$

Dividing both sides by Δ,

$$\rho V(x_t) = \max_{u_t \in U} \left[(1 + \Delta\rho) h(x_t, u_t) + V'(x_t) g(x_t, u_t) + \frac{o(\Delta g(x_t, u_t))}{\Delta} \right].$$

Taking the limit as Δ goes to zero,

$$\rho V(x_t) = \max_{u_t \in U} \left[h(x_t, u_t) + V'(x_t) g(x_t, u_t) \right],$$

where we implicitly assume that the limit as $\Delta \to 0$ of the max with respect to u_t is the same as the max with respect to u_t of the limit as $\Delta \to 0$. Removing the time indices, we obtain the continuous-time version of the Bellman equation,

$$\rho V(x) = \max_{u \in U} \left[h(x, u) + V'(x) g(x, u) \right]. \tag{10.7}$$

Under the regularity conditions, the max of the RHS can be characterised using the following first-order condition for u:

$$0 = h_u(x, u^*(x)) + V'(x) g_u(x, u^*(x)),$$

which defines the optimal decision rule $u^*(x)$. Thus, the following two equations summarise the dynamic programming problem:

$$\begin{aligned} \rho V(x) &= h(x, u^*(x)) + V'(x) g(x, u^*(x)), \\ 0 &= h_u(x, u^*(x)) + V'(x) g_u(x, u^*(x)) \end{aligned} \tag{10.8}$$

for all $x \in X$. Notice that these are two functional equations (i.e. solutions are functions). The functions V and u^* are both functions of x.

10.4 Bellman equation and the Maximum Principle

We now show the sense in which the Bellman equation and the first-order conditions above imply the equations for the Maximum Principle (i.e. Hamiltonian) derived previously.

Recall that in the optimal-control approach, $u^*(t)$ maximises

$$H(x, u, \lambda) = h(x, u) + \lambda g(x, u).$$

From (10.7), in the dynamic programming approach, $u^*(t)$ maximises

$$h(x, u) + V'(x) g(x, u). \tag{10.9}$$

Hence, the two approaches are consistent only if

$$\lambda \equiv V'(x);$$

i.e. the co-state in the calculus of variations approach is the derivative of the value function. This is consistent with the interpretation for the discounted value of the co-state variables offered before: the marginal value of an extra unit of the state variable.

Second, using that $\lambda \equiv V'(x)$, we can see that the first-order condition from optimal-control approach

$$H_u(x, u, \lambda) = 0$$

is equivalent to

$$h_u(x, u) + \lambda g_u(x, u) = 0,$$

which is the derivative of (10.9) with respect to u set to zero.

Third, differentiating (10.8) with respect to time yields

$$\begin{aligned}
\rho V' \dot{x} &= h_x \dot{x} + h_u u^{*\prime} \dot{x} + \left(V'' g + V' g_x + V' g_u u^{*\prime}\right) \dot{x} \\
&= h_x \dot{x} + \left(V'' g + V' g_x\right) \dot{x} + \left(\underbrace{h_u + V' g_u}_{=0 \because \text{FOC}}\right) u^{*\prime} \dot{x} \\
&= h_x \dot{x} + \left(V'' g + V' g_x\right) \dot{x}.
\end{aligned}$$

Outside of the steady state, $\dot{x} \neq 0$, so we can divide through to obtain

$$\rho V' = h_x + V'' g + V' g_x. \tag{10.10}$$

Differentiating $\lambda \equiv V'(x)$ with respect to time yields and using that $\dot{x} = g(x, u)$,

$$\dot{\lambda} = V'' \dot{x} = V'' g.$$

Substituting above and $\lambda \equiv V'(x)$ into (10.10) yields

$$\dot{\lambda} = \rho \lambda - \left(h_x + V' g_x\right),$$

which is equivalent to

$$\dot{\lambda} = \rho \lambda - H_x(x, u, \lambda),$$

which is the law of motion of the co-state variable obtained using the Maximum Principle.

10.5 9-step method

We will focus on stationary problems; i.e. return function and law of motions are time invariant.

9-step method for dynamic programming

(i) Write the sequence problem (SP)
- Choose state variables x and state space X
- Describe the feasibility set $\Gamma(x)$, return function F, and discounting β

(ii) Check basic conditions: feasible set always non-empty + discounting

(iii) Formulate the Bellman equation (BE)

(iv) Check that the Contraction Mapping Theorem (CMT) applies—check conditions for X, Γ, F

(v) Check properties of v (value function) and G (optimal policy correspondence).
- Monotonicity of v
- Concavity of v (if strict, G is single valued, in which case write g)
- Differentiability of v and g

(vi) Euler equation.

(vii) Characterise steady states. Linearise Euler equation to study local stability

(viii) Global stability

(ix) Comparative statics

Consider the following problem:

- Discount factor $\beta \in (0,1)$, where $\beta = 1/(1+\rho)$ and ρ reflects time preference (not interest rate).
- Demand curve given by $p(q)$ that is stationary and satisfies law of demand (i.e. downward sloping). Utility $S(q)$ is given by the area under the demand curve:

$$S(q) = \int_0^q p(z)\, dz.$$

- State variable is x—the stock of "fish" at the beginning of the period.
- Timing: (i) Beginning-of-period stock is x; (ii) q is harvested.
- Law of motion $x_{t+1} = \psi(x_t - q_t)$ gives the next period's stock, where ψ is continuous, strictly increasing, weakly concave and differentiable with $\psi(0) = 0$.

10.5.1 Step 1: Write SP

The sequence problem is given by

$$\begin{aligned} \bar{v}(x) \coloneqq \max_{\{q_t\}_{t=0}^{\infty}} \quad & \sum_{t=0}^{\infty} \beta^t S(q_t) \\ \text{s.t.} \quad & x_{t+1} = \psi(x_t - q_t), \forall t \\ & q_t \in [0, x_t], \end{aligned}$$

where $[0, x_t]$ is the feasibility set, and we have discounting according to β.

In each period, the agent can consume all stock of fish or consume none. So, the feasibility set that describes the next-period stock is given by $\Gamma(x) = [\psi(0), \psi(x)] = [0, \psi(x)]$. Since ψ is continuous and strictly increasing $\varphi \coloneqq \psi^{-1}$ is well defined. We can therefore write

$$\varphi(x_{t+1}) = x_t - q_t \Leftrightarrow q_t = x_t - \varphi(x_{t+1}).$$

We can interpret φ as the amount we leave for reproduction tomorrow. Define the period-return function as

$$F(x_t, x_{t+1}) = S(x_t - \varphi(x_{t+1})).$$

10.5.2 Step 2: Check basic conditions

For each $x \in X$, the feasible set is $\Gamma(x)$ is nonempty. This is sufficient if F is bounded. However, if F is unbounded, some additional conditions are needed to ensure that growth is not "too fast" (relative to β).

10.5.3 Step 3: Formulate BE

$$v(x) \coloneqq \max_{y \in \Gamma(x)} \{S(x - \varphi(y)) + \beta v(y)\},$$

where $\beta v(y)$ is the discounted continuation value.

10.5.4 Step 4: Check that CMT applies

Let $C(X)$ be the space of bounded, continuous function, $\tilde{v}(x) \in C(X)$ for all $x \in X$. Define the Bellman operator as

$$(T\tilde{v})(x) \coloneqq \max_{y \in \Gamma(x)} \{S(x - \varphi(y)) + \beta \tilde{v}(y)\}.$$

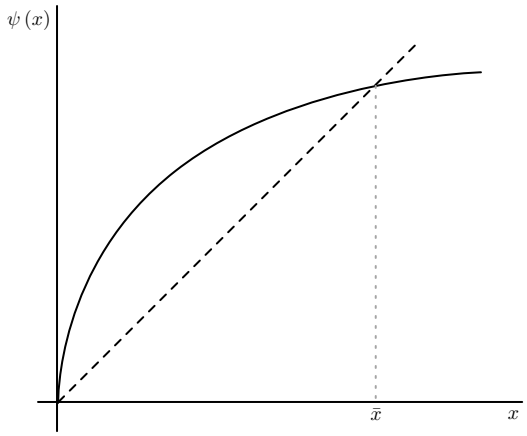
Suppose $\tilde{v}$ is bounded. For CMT to apply, we need that (i) $T\tilde{v}$ is bounded and continuous; (ii) T is a contraction.

- $T\tilde{v}$ is bounded:
  - Either, period return S is bounded—$\lim_{q \to \infty} S(q) \leq \bar{S}$ (S is bounded below by zero by definition);
  - Or, state space X is bounded—e.g. there exists some $\bar{x} > 0$ such that

    $$\psi(x) < x, \forall x > \bar{x}.$$

    We can think of this condition as saying that $\psi(x)$ crosses the 45 degree line at some finite point.[34]

[34] Think of x as δk and $\psi(x)$ as $f(k)$ in the neoclassical growth model.



Define $X = [0, \bar{x}]$ as the state space.

- $T\tilde{v}$ is continuous
  - F varies continuously with x and y—follows from the definition of S as an integral (which guarantees that S is continuous), and φ being continuous.
  - $\Gamma(x)$ varies continuously with state (i.e. Γ is continuous as a correspondence)—follows from the fact that ψ is continuous.
- T is a contraction: Check Blackwell's sufficient condition (T monotone and discounts).

We therefore established that CMT applies, which means that

- Principle of optimality holds; i.e. $v = \bar{v}$.
- T has a unique fixed point v such that $Tv = v$.

10.5.5 Step 5: Check properties of v and G

Monotonicity:

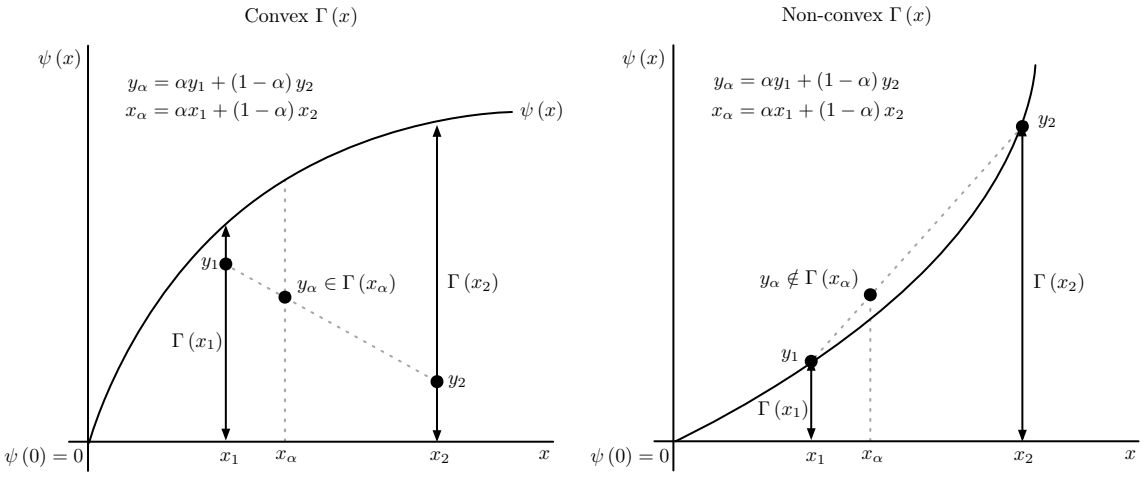
- Requirement 1. $\Gamma(x)$ monotone (i.e. $x' \geq x \Rightarrow \Gamma(x) \subseteq \Gamma(x')$)—holds, since $\Gamma(x) = [0, \psi(x)]$ and ψ is strictly increasing,
- Requirement 2. $F(x, y)$ weakly increasing in x—let $\hat{C}(X) \subset C(X)$ be the set of weakly increasing functions (so the set is closed). Want to show that $\tilde{v} \in \hat{C}(X) \Rightarrow T\tilde{v} \in \hat{C}(X)$—i.e. $T : \hat{C}(X) \to \hat{C}(X)$. Holds since the law of demand holds for $p(q)$ so that S is weakly increasing in x.[35]
- If $F(x, y)$ is strictly increasing in x, then $T\tilde{v}$ is strictly increasing (i.e. the fixed point v lies in the interior of $\hat{C}(x)$).

Concavity:

- Let $\tilde{C}(x) \subset C(x)$ be the set of functions that are weakly concave (so the set is closed). We want to show that $\tilde{v} \in \tilde{C} \Rightarrow T\tilde{v} \in \tilde{C}$—i.e. $T : \tilde{C}(X) \to \tilde{C}(X)$.
- Requirement 1. $F(x, y)$ weakly concave in (x, y)—holds since ψ is weakly concave (which implies that φ is weakly convex and $-\varphi$ is weakly concave).

[35] S is strictly increasing until the demand curve "hits" the x-axis.

- Requirement 2. $\Gamma(x)$ is convex.



- If F is strictly concave in x, then $T\tilde{v}$ is strictly concave (i.e. the fixed point v lies in the interior of $\tilde{C}(x)$). Then, there exists a unique maximiser which is continuous in x (i.e. the optimal policy correspondence G is, in fact, a single-valued function).

Differentiability at $\hat{x}$

- Requirement 1. v is strictly concave.
- Requirement 2. $\hat{x} \in \text{int}(X)$.
- Requirement 3. $g(\hat{x}) \in \text{int}(\Gamma(\hat{x}))$.
- Requirement 4. F is differentiable in x.
- We cannot use the method we used before since the set of differentiable function is not closed (with respect to the sup norm). The idea is to construct a strictly concave, differentiable W that lies everywhere below v and $w(\hat{x}) = v(\hat{x})$. Then W has the same supporting hyperplane as v so that v is differentiable at $\hat{x}$.[36] Fixing $\hat{y} = g(\hat{x})$ and if $\hat{y}$ is feasible in the neighbourhood of $\hat{x}$ (i.e. $\hat{y} \in \text{int}(\Gamma(\hat{x}))$), we can use $w(x) = F(x, \hat{y}) + \beta v(\hat{y})$. Since $\hat{y}$ maximises at $\hat{x}$,[37]

$$w(x) \leq w(\hat{x}) = v(\hat{x}), \ \forall x \neq \hat{x}.$$

  Notice that $\hat{y}$ may not be feasible in the neighbourhood of $\hat{x}$ if v has a kink at $\hat{x}$. In such a case, the inequality above may not hold.

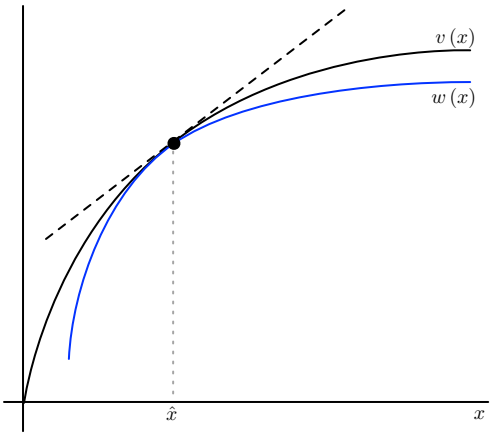
[36] At $x = \hat{x}$, the slope of w and v coincide:

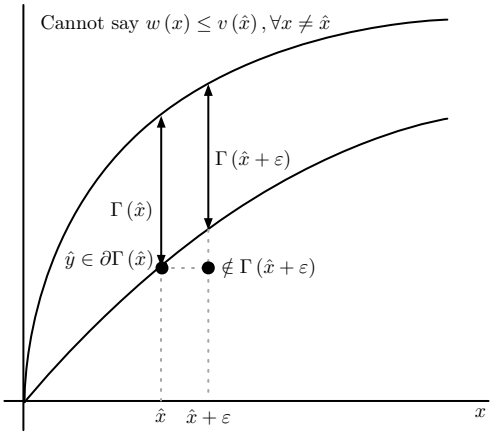
$$\left.\frac{dw(x)}{dx}\right|_{x=\hat{x}} = F_x(\hat{x}, \hat{y}) + \beta v'(\hat{y}) = v'(x).$$

[37] By construction, $w(\hat{x}) = v(\hat{x})$. Finally,

$$\begin{aligned} v(\hat{x}) &= \max_{y \in \Gamma(\hat{x})} F(\hat{x}, y) + \beta v(y) = F(\hat{x}, \hat{y}) + \beta v(\hat{y}) \\ &\geq F(x, \hat{y}) + \beta v(\hat{y}) = w(x) \end{aligned}$$

for all x in the neighbourhood of $\hat{x}$ such that $\hat{y}$ is feasible.





- Differentiability of v does not give differentiability of g. That would require twice differentiability of v. The proof is a "jungle" but it does require at least twice differentiability of F in x.

10.5.6 Step 6: Euler equation

Write the first-order conditions and the envelope condition.

First-order condition The first-order condition with respect to y is given by

$$-S'\left(x-\varphi\left(y^*\right)\right)\varphi'\left(y^*\right)+\beta v'\left(y^*\right) \begin{matrix} \leq \\ = \\ \geq \end{matrix} 0,$$

where: (i) $\leq$ if $y^* = 0 = \psi(0)$ (i.e. $q = x$); (ii) $=$ if $y^* \in (0, \psi(x))$; and (iii) $\geq$ if $y^* = \psi(x)$ (i.e. $q = 0$). Notice that the first-order condition equates the marginal value of consumption today $-S'(\cdot)\varphi'(\cdot)$ against the marginal value of stock in the future $\beta v'(y^*)$.

Envelope condition Substituting in $y^* = g(x)$ into the BE:

$$v(x) = S\left(x-\varphi\left(g(x)\right)\right)+\beta v\left(g(x)\right).$$

Differentiating above with respect to x, while noting that the Envelope Theorem implies that we need not consider derivatives with respect to $g(x)$:

$$v'(x) = S'\left(x-\varphi\left(g(x)\right)\right).$$

Euler equation Euler equation is given by combining the first-order and the envelope conditions. From the envelope condition:

$$v'\left(g(x)\right) = S'\left(g(x)-\varphi\left(g\left(g(x)\right)\right)\right).$$

Substituting this into the first-order condition gives the Euler equation:

$$S'\left(x-\varphi\left(g(x)\right)\right)\varphi'\left(g(x)\right) = \beta S'\left(g(x)-\varphi\left(g\left(g(x)\right)\right)\right).$$

Corner solutions We can think about what conditions would rule out corner solutions.

Consider first the case $y^* = g(x) = 0$. This is optimal if

$$S'(x)\varphi'(0) \geq \beta S'\left(g(x)-\varphi\left(g\left(g(x)\right)\right)\right) = \beta S'\left(-\varphi\left(g(0)\right)\right).$$

where we also used the fact that $\varphi(0) = 0$ and $g(x) = 0$. If the inequality does not hold, then we can rule out this corner solution.

Now consider the case $y^* = g(x) = \psi(x)$. This is optimal if

$$\begin{aligned} S'\left(x - \varphi\left(\psi\left(x\right)\right)\right)\varphi'\left(\psi\left(x\right)\right) &\leq \beta S'\left(\psi\left(x\right) - \varphi\left(g\left(\psi\left(x\right)\right)\right)\right) \\ \Leftrightarrow S'\left(0\right)\varphi'\left(\psi\left(x\right)\right) &\leq \beta S'\left(\psi\left(x\right) - \varphi\left(g\left(\psi\left(x\right)\right)\right)\right) \end{aligned}$$

where we used the fact that $\varphi = \psi^{-1}$. Again, if the inequality does not hold, then we can rule out this corner solution. For example, if we have an Inada condition so that $S'(0) = +\infty$, then we can rule out this corner solution.

10.5.7 Step 7: Characterise steady states

Steady state: Interior In the steady state, $x = g(x) = \bar{x}$, then the Euler equation becomes

$$S'\left(\bar{x} - \varphi\left(\bar{x}\right)\right)\varphi'\left(\bar{x}\right) = \beta S'\left(\bar{x} - \varphi\left(\bar{x}\right)\right),$$

which simplifies to

$$\varphi'\left(\bar{x}\right) = \beta.$$

Above pins down the steady-state value of stock. This can be seen as a rate of return condition—the steady stock is such that the reproduction rate offsets discounting.

Steady state: Corner solutions Steady states at corner solution are possible. If

$$S'\left(x\right)\varphi'\left(0\right) > \beta s'\left(0\right),$$

then

$$\bar{x} = 0.$$

Similarly, if

$$S'\left(0\right)\varphi'\left(\psi\left(x\right)\right) < \beta S'\left(\psi\left(x\right) - \varphi\left(\psi\left(x\right)\right)\right)$$

then $\bar{x} = \psi(\bar{x})$.

10.5.8 Step 8: Global stability

N/A.

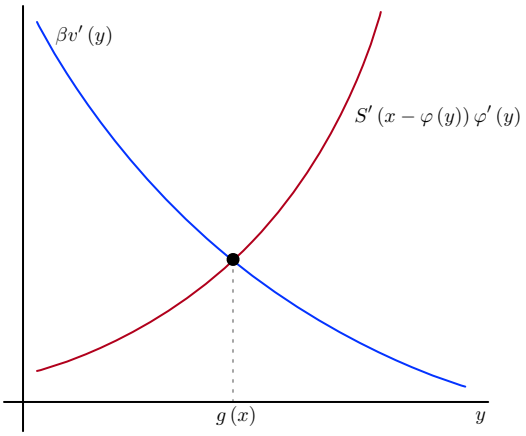
10.5.9 Step 9: Comparative statics

Recall that we are not guaranteed that v is twice differentiable or that $g(x)$ is differentiable. So we need to conduct comparative statics without taking derivatives.

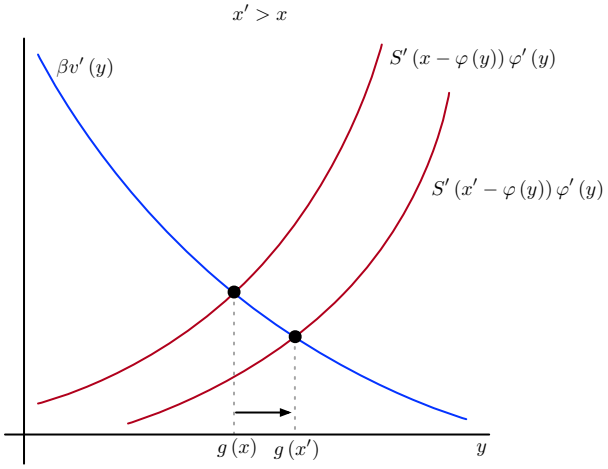
Notice that, from the first-order condition, $y^* = g(x)$ is given by the intersection of $\beta v'(y)$ and $S'(x - \varphi(y))\varphi'(y)$, and

- $\beta v'(\cdot)$ is strictly decreasing in y since v is strictly concave;
- $S'(\cdot)\varphi'(\cdot)$ is increasing in y since: (i) $-\varphi$ is strictly decreasing in y, and we assume that S is strictly concave; (ii) $S' > 0$ since S is (assumed to be) strictly increasing; and (iii) φ is strictly increasing and strictly convex so that $\varphi' > 0$ and φ' is increasing in y.

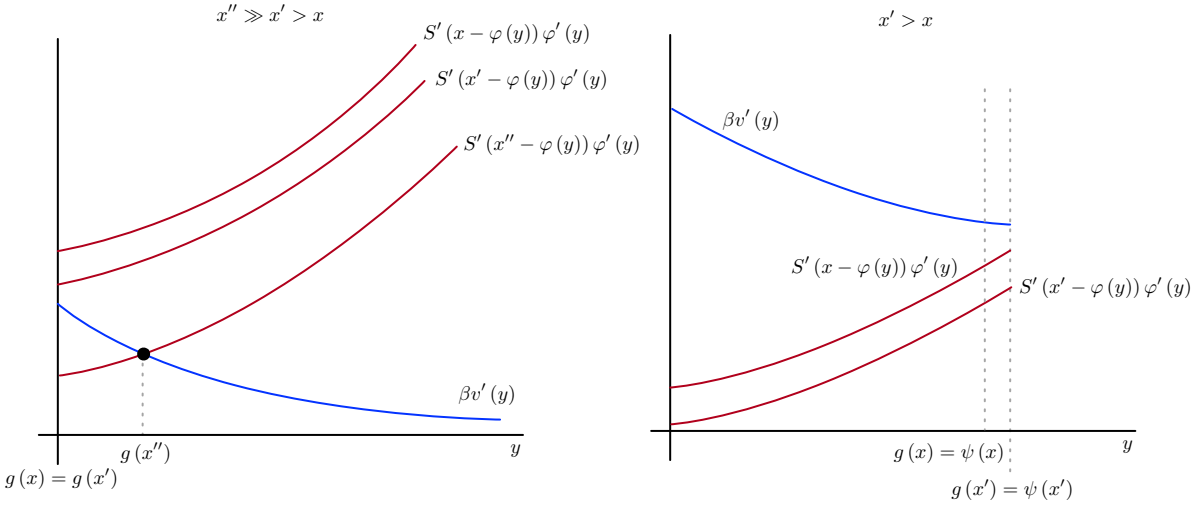
We can plot the two curves.



How would g change if stock today was higher; i.e. $x' > x$. Since S is strictly concave, a higher x implies S' is lower; i.e. the locus $S'(\cdot)\varphi'(\cdot)$ moves down/right. $\beta v'(\cdot)$ is unaffected by x. As can be seen from the figure, this implies that $g(x)$ is increasing in x.



Note that if $g(x)$ is a corner solution, $g(x)$ may not be strictly increasing with x.



- $g(x)$ is at the lower bound zero (figure on the left). The initial value of state is x.

- Suppose x increases to x'. The figure shows the case in which the two curves do not cross at x' so that $g(x) = g(x')$ and g is not strictly increasing. However, observe that $S'(x' - \varphi(y))\varphi(y) < S'(x - \varphi(y))\varphi(y)$ since y remains unchanged. Since S is strictly concave, S' is decreasing so that

$$q' = x' - \varphi(y) > x - \varphi(y) = q,$$

  where we used the fact that $x - \varphi(y)$ equals consumption, q. Hence, we realise that consumption is increasing in x.

- Suppose x increases to x'' such that the two curves now intersect. Then, as we can see from the figure, $g(x'') > g(x)$.

- $g(x)$ is at the upper bound $\psi(x)$ (figure on the right). The initial value of state is x. In this case, an increase in state from x to x' increases the upper bound from $\psi(x)$ to $\psi(x')$ so that $g(x') > g(x)$.

Remark. In general, the lower bound could also move but we may not have monotonicity of $\Gamma(x)$ in such a case.

Part III
Probability and statistics

11 Foundation

11.1 Probability space

We use probability to describe uncertainty, variability, and randomness. By uncertainty, we mean that some *outcome*, ω, is unknown. We refer to the set of all possible outcomes as the *sample space*, Ω, and any subset of outcomes, $E \subseteq \Omega$, as an *event*.

Example 35. Consider a single coin flip. The set of all possible outcomes of a coin flip consists of Heads (H) and Tails (T); i.e., $\Omega = \{H, T\}$. For example, $\{H\}$ is an event, as is $\{T\}$.

Example 36. Consider flipping a coin twice. The set of all possible outcomes is now $\Omega = \{HH, HT, TH, TT\}$. For example, $\{HH\}$ and $\{TH, HH, TT\}$ are both events.

Example 37. Consider surveying the wage of a worker selected randomly from Ithaca. The sample space is $\Omega = [w_{\min}, \infty)$, where $w_{\min}$ is the minimum hourly wage. An example of an event is $\{w \in \mathbb{R} : w \geq 2w_{\min}\}$.

We formalise the set of events as follows. A *σ-algebra* (or a *σ-field*) on a set X is a collection of subsets of X, denoted $\mathcal{F}$, that satisfies the following conditions.[38]

(i) The set itself belongs in $\mathcal{F}$; i.e., $X \in \mathcal{F}$.

(ii) $\mathcal{F}$ is closed under complementation; i.e., $E^c \in \mathcal{F}\ \forall E \in \mathcal{F}$.

(iii) $\mathcal{F}$ is closed under countable unions; i.e., $\bigcup_{k=1}^{\infty} E_k \in \mathcal{F}\ \forall \{E_k\}_{k=1}^{\infty} \subseteq \mathcal{F}$.

Observe that above implies that $\varnothing \in \mathcal{F}$ and that $\mathcal{F}$ is closed (why?) under countable intersections (why?). The pair $(X, \mathcal{F})$ is called a *measurable space*.

Remark 25 (Comparison with topology on X). Recall that a collection of subsets of X, $\mathcal{T}$, is a topology for X if (i) $\varnothing, X \in \mathcal{T}$; (ii) $\mathcal{T}$ is closed under finite intersections; (iii) $\mathcal{T}$ is closed under arbitrary unions.

Example 38. Given a nonempty set of outcomes Ω, we can have many different σ-algebras. For example, the smallest σ-algebra is $\mathcal{F} = \{\varnothing, \Omega\}$, and the largest σ-algebra is $\mathcal{F} = 2^{\Omega}$ (the set of all subsets of $\mathcal{F}$). In case Ω consists of outcomes of two coin tosses, the set

$$\{\varnothing, \{HH\}, \{TT\}, \Omega\}$$

is not a σ-algebra (why?); however, the set

$$\{\varnothing, \{HH\}, \{TH, HT, TT\}, \Omega\}$$

is a σ-algebra.

By giving a topology $\mathcal{T}$ on the sample space Ω, we can define the *Borel algebra* on Ω, denoted $\mathcal{B}(\Omega, \mathcal{T})$, as the smallest σ-algebra that contains $\mathcal{T}$.

[38] Incidentally, an *algebra* on X is a collection of subsets that contains the whole set and is closed under complementation and finite unions.

Example 39. When Ω is countable and $\mathcal{T} = 2^\Omega$, then $\mathcal{B}(\Omega, \mathcal{T}) = 2^\Omega$. If $\Omega = \mathbb{R}$, then $\mathcal{B}(\Omega, \mathcal{T})$ is the collection of all open and closed intervals, their countable unions, intersections, and complements.

We say that an event E *occurs* if the outcome ω is in E; i.e., if $\omega \in E$. Probabilities describe how likely it is for an event to occur. Formally, given a sample space Ω and a σ-algebra, $\mathcal{F}$, a function $P : \mathcal{F} \to [0, 1]$, is a probability function if

(i) $P(\{E\}) \geq 0$ for all $E \in \mathcal{F}$;

(ii) $P(\Omega) = 1$;

(iii) For any countable collection of disjoint events $\{E_1, E_2, \dots, \} \subseteq \mathcal{F}$, $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

The function P is sometimes also called a *probability measure*. We refer to the tuple $(\Omega, \mathcal{F}, P)$ as a *probability space*. These conditions of probability function imply that, for any event E and E':

$$\begin{aligned} P\left(E^c\right) &= 1 - P\left(E\right), \\ P\left(\varnothing\right) &= 0, \\ P\left(E\right) &\leq 1, \\ E \subseteq E' &\Rightarrow P\left(E\right) \leq P\left(E'\right). \end{aligned}$$

Exercise 24. Show that

$$\begin{aligned} &P\left(E \cup E'\right) = P\left(E\right) + P\left(E'\right) - P\left(E \cap E'\right), \\ &P\left(E \cup E'\right) \leq P\left(E\right) + P\left(E'\right), \\ &P\left(E \cap E'\right) \geq P\left(E\right) + P\left(E'\right) - 1. \end{aligned}$$

Given an event F with $P(F) > 0$, the *conditional probability* of E given F, denoted $P(E|F)$, is defined

$$P\left(E|F\right) \coloneqq \frac{P\left(E \cap F\right)}{P\left(F\right)}.$$

Two events E and F are (*statistically*) *independent*, denoted $E \perp\!\!\!\perp F$, if

$$P\left(E \cap F\right) = P\left(E\right) P\left(F\right).$$

Two events are *dependent* if they are not independent. For any two independent events E and F with $P(E), P(F) > 0$,

$$P\left(E|F\right) = P\left(E\right) \text{ and } P\left(F|E\right) = P\left(F\right).$$

Moreover, when events E and F are independent, the following pairs are also independent: $E \perp\!\!\!\perp F^c$, $E^c \perp\!\!\!\perp F$ and $E^c \perp\!\!\!\perp F^c$.

Theorem 28 (Law of Total probability)**.** *Let $\{E_1, E_2, \dots, E_n\}$ be a partition of event space Ω and $P(B_i) > 0$ for all $i \in \{1, 2, \dots, n\}$. Then, for any event F,*

$$P\left(F\right) = \sum_{i=1}^{n} P\left(E_i\right) P\left(F|E_i\right).$$

Proof. Given a partition $\{E_1, E_2, \dots, E_n\}$ of Ω, for any event F, $F = \bigcup_{i=1}^{n} (A \cap E_i)$. Thus,

$$P\left(F\right) = \sum_{i=1}^{n} P\left(A \cap E_i\right) = \sum_{i=1}^{n} P\left(E_i\right) P\left(F|E_i\right),$$

where the first equality follows from the fact that $A \cap E_i$ and $A \cap E_{i'}$ are disjoint for any distinct $i, i' \in \{1, 2, \ldots, n\}$, and the second equality follows from the definition of conditional probability. ■

Theorem 29 (Bayes Rule)**.** *Given events E and F such that $P(E), P(F) > 0$,*

$$P(E|F) = \frac{P(F|E)\,P(E)}{P(F|E)\,P(E) + P(F|E^c)\,P(E^c)}.$$

Proof. Follows from the definition of conditional probability and the law of total probability. ■

11.2 Random variables

A *random variable* is a real-valued outcome and is a function from the sample space Ω to the real line $\mathbb{R}$. You will often see that a random variable written using the upper case and the realisation of a random variable using the lower case.

Example 40. In case of a coin flip, $\Omega = \{H, T\}$ and we may define a random variable as $X : \Omega \to \mathbb{R}$, $\omega \mapsto \mathbb{1}_{\{\omega = H\}}$.

Let Ω be a sample space with a probability function P. A random variable X induces a new sample space in $\mathbb{R}$ and we can define a probability function $P_X : \mathcal{B}(\Omega) \to [0, 1]$ as

$$P_X(X \in B) \coloneqq P(\{\omega \in \Omega : X(\omega) \in B\})$$

for any set B in the Borel algebra of $\mathbb{R}$.

Example 41. Consider three coin flips and let $\Omega = \{H, T\}^3$. We can define a random variable X as the total number of heads. The sample space X is $\mathcal{X} = \{0, 1, 2, 3\}$. Then, $\mathcal{B} = 2^{\mathcal{X}}$ and

$$\begin{aligned} P_X(X = 2) &= P(\{\omega \in \Omega : X(\omega) = 2\}) \\ &= P(\{HHT, HTH, THH\}). \end{aligned}$$

A random variable X always has a *cumulative distribution function* (*CDF*), denoted F_X, given by

$$F_X(x) = P_X(\{X \leq x\}) \ \ \forall x \in \mathbb{R}.$$

The CDF F_X has the following properties:

(i) $F_X(\cdot)$ is nondecreasing;

(ii) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$;

(iii) $F_X(\cdot)$ is right continuous; i.e., $\lim_{x \searrow x_0} F_X(x) = F(x_0)$ for all $x_0 \in \mathbb{R}$.

Moreover, we have

$$P(a < X \leq b) = F_X(b) - F_X(a) \ \text{ and } P(X > b) = 1 - F_X(b).$$

From now on, we simply write P instead of P_X as the probability function of X, and write $F(x)$ instead of $F_X(x)$ as the CDF of X. We write $X \sim F$ to mean that the random variable X is distributed according to CDF F.

We say that a random variable X is *continuous* if F is continuous and that X is *discrete* if F is a step function. Note that a random variable can be neither continuous nor discrete. Given the properties of CDF, the CDF of discrete random variables have jumps. A discrete random variable can only take finite or countably infinite number of values.

Suppose X is a discrete random variable. Its *probability mass function* (*pmf*), denoted f, is given by

$$f(x) \coloneqq P(X = x).$$

To compute the probability of an event, we can sum over all points in that event; e.g.

$$P(X \leq x) = \sum_{\tilde{x} \in \{x' \in X : x' \leq x\}} f(\tilde{x}),$$

When X is a continuous random variable, $P(X = x) = 0$. We therefore work instead with *probability density function* (*pdf*) of X, also denoted f, which is given by

$$f(x) \coloneqq \frac{\mathrm{d}}{\mathrm{d}x} F(x).$$

Note that pdf is well-defined only when F is differentiable. Both pmf and have the following properties:

(i) $f \geq 0$;

(ii) $\sum_x f(x) = 1$ (for pmf) and $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ (for pdf).

Moreover, by the fundamental theorem of calculus,

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \;\forall x \in \mathbb{R},$$

and we have

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_a^b f(t)\,\mathrm{d}t.$$

Given a random variable X with pmf or pdf, f, the *support* of X, denoted $\mathrm{supp}(X)$, is given by

$$\mathrm{supp}(X) \coloneqq \{x \in \mathbb{R} : f(x) > 0\};$$

i.e., $\mathrm{supp}(X)$ is set of realisations that can occur with positive probability.

Example 42 (Logistic distribution)**.** Let $F(x) \coloneqq \frac{1}{1+e^{-x}}$ for all $x \in \mathbb{R}$. Then,

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) = \frac{e^{-x}}{(1+e^{-x})^2} = F(x)(1 - F(x)).$$

and

$$P(a < X < b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x - \int_{-\infty}^{a} f(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x.$$

11.2.1 Transformation of random variables

Given random variable X and a function $g : \mathbb{R} \to \mathbb{R}$, $Y = g(X)$ is also a random variable; Y is a transformation of X via g. Observe that given any subset $B \in \mathcal{B}$,

$$P(Y \in B) = P(g(X) \in B)$$

and so the distribution of Y depends on F and g. Let sample spaces for X and Y as

$$\mathcal{X} \coloneqq \mathrm{supp}(X)\;\; \mathcal{Y} \coloneqq g(\mathcal{X}).$$

The function g then defines a mapping $g : \mathcal{X} \to \mathcal{Y}$. The distribution of Y can then be found as

$$\begin{aligned} P(Y \in E) &= P(g(X) \in E) \\ &= P(\{x \in \mathcal{X} : g(x) \in E\}) \\ &= P\left(X \in g^{-1}(E)\right), \end{aligned}$$

where g^{-1} is the inverse mapping of g. Note that if X is a discrete random variable, $\mathcal{X}$ and $\mathcal{Y}$ are both countable; i.e., Y is also a discrete random variable. Its pmf is given by

$$f_Y(y) = \begin{cases} P_Y(Y = y) = \sum_{x \in g^{-1}(y)} P_X(X = x) = \sum_{x \in g^{-1}(y)} f_X(x) & \text{if } y \in \mathcal{Y} \\ 0 & \text{if } y \notin \mathcal{Y} \end{cases}.$$

That is, to obtain the pmf of Y we can proceed as follows: for each $y \in \mathcal{Y}$, identify elements in the set $g^{-1}(y)$ and sum appropriate probabilities.

Exercise 25. Consider a discrete random variable X that has a binomial distribution:

$$f(x) = P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad \forall x \in \{0, 1, \ldots, n\},$$

where n is a positive integer and $p \in [0, 1]$. Find the pmf of $Y = n - X$.

We can use the fact that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) = P(\{x \in \mathcal{X}\} : g(x) \leq y) \\ &= \int_{\{x \in \mathcal{X} : g(x) \leq y\}} f_X(x)\, \mathrm{d}x. \end{aligned}$$

to compute the CDF (or pdf) of Y. A particularly simple case is when g is strictly monotone.

Theorem 30. *Let X be a continuous random variable with pdf f_X, $Y \coloneqq g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is a strictly monotone function such that g^{-1} has a continuous derivative on $\mathcal{Y}$. Then, the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & y \notin \mathcal{Y} \end{cases}.$$

Proof. Let us suppose that g is strictly decreasing (the proof for the case when g is strictly decreasing is analogous). Since g is strictly increasing, $g^{-1} : \mathcal{Y} \to \mathcal{X}$ is a well-defined, strictly decreasing function and so

$$\begin{aligned} F_Y(y) &= P(g(X) \leq y) \\ &= P\left(X \geq g^{-1}(y)\right) \\ &= 1 - F_X\left(g^{-1}(y)\right), \end{aligned}$$

where the last equality uses the fact that X is a continuous random variable. By the chain rule,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = -f_X\left(g^{-1}(y)\right) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) > 0. \qquad \blacksquare$$

Exercise 26. Suppose a random variable X is uniformly distributed on $(0, 1)$. What is its cdf and pdf? Find the pdf of $Y = -\ln X$.

Theorem 31. *Let X have a continuous and strictly increasing CDF, F_X. Define $Y \coloneqq F_X(X)$. Then, Y is uniformly distributed on $(0, 1)$; i.e., $P(Y \leq y) = y$ for all $y \in (0, 1)$.*

Proof. For each $y \in (0,1)$,

$$P(Y \leq y) = P(F_X(X) \leq y) = P\left(X \leq F_X^{-1}(y)\right) = F_X\left(F_X^{-1}(y)\right) = y.$$

At the end points,

$$P(Y \leq y) = 1 \ \forall y \geq 1,$$
$$P(Y \leq y) = 0 \ \forall y \leq 0.$$ ■

Even if g is not monotonic, it may be possible to compute the distribution of $Y = g(X)$ by direct manipulation.

Exercise 27. Let $Y = X^2$, where X is a continuous random variable with support $\mathcal{X} = \mathbb{R}$ and pdf f_X. Compute F_Y and f_y.

11.2.2 Expectations

The *expectation* of a random variable X, denoted $\mathbb{E}[X]$, is given by all its possible realisations averaged according to the probability of each realisation occurring. We think of it as the typical (or expected) value of an observation from the random variable. Provided that it exists, $\mathbb{E}[X]$ is defined as

$$\mathbb{E}[X] \coloneqq \int_{-\infty}^{\infty} x \mathrm{d}F(x) = \begin{cases} \sum_{x \in \mathcal{X}} g(x) f_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}.$$

If $\mathbb{E}[|X|] < \infty$, we say that $\mathbb{E}[X]$ *exists* (or *well defined*). If $\mathbb{E}[|X|] = \infty$, then

$$\mathbb{E}[X] = \begin{cases} \infty & \text{if } I_+ = \infty,\ I_- > -\infty \\ -\infty & \text{if } I_+ < \infty,\ I_- = -\infty \\ \text{undefined} & \text{if } I_+ = \infty,\ I_- = -\infty \end{cases},$$

where

$$I_+ \coloneqq \int_{x>0} g(x) \mathrm{d}F(x), \ I_- \coloneqq \int_{x<0} g(x) \mathrm{d}F(x).$$

Since integration is a linear operation, expectation is a linear operator; i.e., for any $a, b \in \mathbb{R}$,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

You will use the following inequality many times in your first year.

Theorem 32 (Jensen's inequality)**.** *Suppose $g : \mathbb{R} \to \mathbb{R}$ is convex, then for any random variable X,*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Clearly, above implies that if g is concave, then the inequality reverses (right?).

Exercise 28. Use Jensen's inequality to prove that, for any random variable X, and any $0 < r \leq p$,

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|],$$

and that

$$\left(\mathbb{E}\left[|X|^r\right]\right)^{\frac{1}{r}} \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}}.$$

Note that if $\mathbb{E}[|X|^p] < \infty$, then $\mathbb{E}[|X|^r] < \infty$ for $0 < r \leq p$.

11.2.3 Moments

Given any $m \in \mathbb{N}$ and a random variable X, the *mth moment* of X, denoted μ'_m, is defined

$$\mu'_m \coloneqq \mathbb{E}\left[X^m\right],$$

and the *mth central moment* of X, denoted μ_m, is given by

$$\mu_m \coloneqq \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^m\right].$$

The second central moment is simply called the *variance*, denoted $\mathrm{Var}[X]$ *or* σ_X^2 . The *standard deviation* of X is the positive square root of the variance; i.e., $\mathrm{std}(X)$ (or σ_X) is given by $\sqrt{\mathrm{Var}[X]}$.

Exercise 29. Prove that $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ and that $\mathrm{Var}(aX + b) = a^2\mathrm{Var}[X]$ for any $a, b \in \mathbb{R}$.

The first result, in particular, implies

$$\mathrm{Var}[X] \leq \mathbb{E}[X^2].$$

Thus, existence of variance is guaranteed if $\mathbb{E}[X^2] < \infty$.

11.2.4 Quantiles

Expectation is not the only way to measure the "centre" of distributions. Let X have CDF F_X. Given any $\alpha \in [0, 1]$, the *α-quantile* of F_X, denoted $Q(\alpha)$, solves

$$F_X\left(Q\left(\alpha\right)\right) = \alpha \Leftrightarrow P\left(X \leq Q\left(\alpha\right)\right) = \alpha.$$

In particular, we let

$$Q\left(\alpha\right) \coloneqq \min\left\{x \in \mathbb{R} : F_X\left(x\right) \geq \alpha\right\}.$$

(why inf?). The 0.5-quantile for a random variable is called the *median*, and *upper* and *lower quantiles* refer to the 0.75- and 0.25-quantiles, respectively.

11.2.5 Moment generating and characteristic functions

The *moment generating function* (*gmf*) of a random variable X, denoted M_X, is defined as

$$M_X\left(t\right) \coloneqq \mathbb{E}\left[\exp\left(tX\right)\right]$$

provided that the expectation exists (oi.e., is finite) for t in a neighbourhood of 0. For the mgf to exist, the density of X must have thin tails. The following shows that the curvature of M_X at $t = 0$ encodes all moments of the distribution of X.

Theorem 33. *Let M_X be the mgf of a random variable X, then $M(0) = 1$, and for any $m \in \mathbb{N}$,*

$$\frac{\mathrm{d}^m}{(\mathrm{d}t)^m} M_X\left(t\right)\Big|_{t=0} = \mathbb{E}\left[X^m\right].$$

Proof. Using $M(0) = \mathbb{E}[\exp(0)] = 1$. Note that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} M_X(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_{-\infty}^{\infty} \exp(tx)\,\mathrm{d}F(x)\right] \\ &= \int_{-\infty}^{\infty} \left(\frac{\mathrm{d}}{\mathrm{d}t}\exp(tx)\right)\mathrm{d}F(x) \\ &= \int_{-\infty}^{\infty} (\exp(tx)\,x)\,\mathrm{d}F(x)\end{aligned}$$

and evaluating this at $t = 0$ yields $\mathbb{E}[X]$. Other moments can be derived analogously. ■

Because the moment generating function is not necessarily finite, characteristic functions are sometimes used for in some proofs. The *characteristic function* (*cf*) of a random variable X, denoted C_X, is defined as

$$C_X(t) \coloneqq \mathbb{E}[\exp(itX)],$$

where $i \coloneqq \sqrt{-1}$. Since $\exp(iu) = \sin(u) + i\cos(u)$ for any $u \in \mathbb{R}$, C_X exists for any random variable. Moreover,

$$|\exp(iu)| = |\sin(u) + i\cos(u)| = \sqrt{(\sin(u))^2 + (\cos(u))^2} = 1,$$

and

$$\frac{\mathrm{d}^m}{(\mathrm{d}t)^m} C_X(t)\bigg|_{t=0} = i^m \mathbb{E}[X^m].$$

11.3 Bivariate random vector

We first study the bivariate ran case in which $n = 2$. In this case, a random vector is a pair (X, Y) that represents mapping from a sample space to a pair of numbers $(x, y) \in \mathbb{R}^2$.

Example 43. Consider the "experiment" of flipping two coins. The sample space is

$$\Omega \coloneqq \{HH, HT, TH, TT\}.$$

We can define a bivariate random vector (X, Y) via

$$X = \mathbf{1}_{\{H \text{ on the first coin}\}},\ Y = \mathbf{1}_{\{H \text{ on the second coin}\}}.$$

Hence, each sample pint in Ω is associated with a pair of numbers; e.g., HH is associated with $(1, 1)$.

11.3.1 Joint distributions

The *joint distribution function* of a random vector (X, Y) is

$$F_{X,Y}(x, y) = P(X \le x,\ Y \le y) = P(\{X \le x\} \cap \{Y \le y\}).$$

It has properties that are analogous to the univariate case:

(i) $F(x, y)$ is non-decreasing in each argument;

(ii) $\lim_{x\to-\infty} F(x, y) = \lim_{y\to-\infty} F(x, y) = 0$ and $\lim_{x\to\infty,\ y\to\infty} F(x, y) = 1$;

(iii) $0 \le F \le 1$;

(iv) $P(a < X \le b,\ c < Y \le d) = F(b, d) - F(b, c) - F(a, d) + F(a, c)$.

As before, a random vector (X, Y) is *continuous* if its joint distribution function is continuous, and is *discrete* if it is a step function.

When (X, Y) is a discrete bivariate random vector, its *joint probability mass function* (*joint pmf*) is given by

$$f_{X,Y}(x, y) = P(X = x,\ Y = y).$$

Given a subset $E \subseteq \mathbb{R}^2$,

$$P((X, Y) \in E) = \sum_{(x,y)\in E} f_{X,Y}(x, y).$$

When (X, Y) is a continuous bivariate random vector, its *joint probability density function* (*joint pdf*) is given by

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$$

and it satisfies $f \geq 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u, v)\mathrm{d}u\mathrm{d}v = 1$. By the Fundamental Theorem of Calculus,

$$P(a \leq X \leq b,\ c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y$$

and

$$F(x, y) = P(X \leq x,\ Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u, v)\,\mathrm{d}u\mathrm{d}v.$$

Given a subset $E \subseteq \mathbb{R}^2$,

$$P((X, Y) \in E) = \int_{(x,y)\in E} f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y.$$

11.3.2 Marginal distributions

Sometimes, we are only interested in one dimension of the bivariate random variable. Given a bivariate random variable (X, Y), the *marginal distribution* of X is

$$\begin{aligned} F_X(x) &= P(X \leq x) = P\left(\left\{(u, v) \in \mathbb{R}^2 : u \leq x\right\}\right) \\ &= P\left(\left\{(u, v) \in \mathbb{R}^2 : u \leq x,\ v \leq \infty\right\}\right) = P(X \leq x,\ Y \leq \infty) \\ &= \lim_{y\to\infty} F_{X,Y}(x, y). \end{aligned}$$

In the discrete case, the *magical probability mass function* of X

$$\begin{aligned} f_X(x) &= P((X, Y) \in \{(x, y) : y \leq \infty\}) \\ &= \sum_{y\in\mathbb{R}} f_{X,Y}(x, y). \end{aligned}$$

In the continuous case, the *marginal probability density function* of X is

$$\begin{aligned} f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} P(X \leq x,\ Y \leq \infty) \\ &= \frac{\mathrm{d}}{\mathrm{d}x} \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_{X,Y}(u,y)\,\mathrm{d}u\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x} \left(\int_{-\infty}^{x} f_{X,Y}(u,y)\,\mathrm{d}u \right) \mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y, \end{aligned}$$

where the last line uses the Leibniz rule for differentiating integrals.

Given a g that maps (X,Y) to $\mathbb{R}$, the expectation of $\mathrm{g}(X,Y)$ is

$$\mathbb{E}\left[g(X,Y)\right] = \begin{cases} \sum_{(x,y)\in\mathbb{R}^2} g(x,y)\, f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y & \text{if } (X,Y) \text{ is continuous} \end{cases}.$$

Exercise 30. Compute $\mathbb{E}[g(X)]$ when (X,Y) has joint distribution pdf (or pmf) $f_{X,Y}$.

11.3.3 Conditional distributions

When observing a random vector (X,Y), the values of X and Y might be related (e.g., X is years of education and Y is wage). In such case, knowledge of the value of X might gives us information about Y (or not). Conditional distribution is a way to formalise what we can learn about one dimension when we know something about the other dimension.

Let (X,Y) be a bivariate random vector. Let us consider the *conditional distribution of Y given* $X = x$, denoted $F_{Y|X}(\cdot|x)$. When X has a discrete distribution,

$$F_{Y|X}(y|x) = P(Y \leq y | X = x) = \frac{P(Y \leq y,\ X = x)}{P(X = x)}$$

for any x such that $P(X = x) > 0$ (check that $F_{Y|X}$ is a valid cumulative distribution function). When Y is also discrete, the *conditional probability mass function* of Y given $X = x$, denoted $f_{Y|X}(y|x)$, is

$$f_{Y|X}(y|x) = P(Y = y | X = x) = \frac{P(Y = y,\ X = x)}{P(X = x)} = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

provided that $f_X(x) > 0$ (check that $f_{Y|X}$ is a valid pmf). When instead $F_{Y|X}$ is differentiable with respect to y and $P(X = x) > 0$, the *conditional probability density function* of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y} F_{Y|X}(y|x).$$

If (X,Y) is a continuous bivariate random vector with joint probability density function $f_{X,Y}$ and marginal probability density functions f_X and f_Y, then

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

provided that $f_X(x) > 0$ (check that $f_{Y|X}$ is a valid pdf). In this case, $f_{Y|X}$ should be understood as

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y} F_{Y|X}(y|x),$$

where

$$F_{Y|X}(y|x) = \lim_{\epsilon \searrow 0} P(Y \leq y | x - \epsilon \leq X \leq x + \epsilon).$$

Since $f_{Y|X}$ as a function of y is a pdf (or a pmf), we can use it in the same way as we did previously to derive conditional expectations of $Y|X$ or some function g of $Y|X$. Similarly, one can derive conditional moments (e.g., variance) as before. Importantly, whoever, conditional moments such as $\mathbb{E}[Y|X = x]$ and $\text{Var}[Y|X = x]$ are both functions of x and hence $\mathbb{E}[Y|X]$ and $\text{Var}[Y|X]$ are both random variables, the values of which depend on the realisation of X.

Proposition 102. *Let X and Y be two random variables. For any $r \geq 1$ such that $\mathbb{E}[|Y|^r] < \infty$,*

$$\mathbb{E}\left[|\mathbb{E}[Y|X]|^r\right] \leq \mathbb{E}\left[|Y|^r\right].$$

Proof. Note $g(u) \coloneqq |u|^r$ is convex for $r \geq 1$. Hence, Jensen's inequality says

$$|\mathbb{E}[Y|X]|^r \leq \mathbb{E}\left[|Y|^r | X\right].$$

Taking expectation of both sides with respect to X yields the result. ■

11.3.4 Independence

We say that two random variables X and Y are (*statistically*) *independent*, denoted $X \perp\!\!\!\perp Y$, if

$$F_{X,Y} = F_X(x) F_Y(y) \ \ \forall x, y \in \mathbb{R}^2.$$

Suppose that X and Y are independent.

- $f_{X,Y}(x, y) = f_X(x) f_Y(y) \ \ \forall x, y \in \mathbb{R}^2$ and hence

  $$f_{Y|X}(y|x) = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y);$$

  i.e., knowledge that $X = x$ gives no information about Y.
- For any functions $g, h : \mathbb{R} \to \mathbb{R}$, $g(X)$ and $h(Y)$ are independent.
- For any functions $g, h : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[|g(X)|], \mathbb{E}[|h(X)|] < \infty$,

  $$\mathbb{E}[g(X) h(Y)] = \mathbb{E}[g(X)] \mathbb{E}[h(Y)].$$

- Letting M_X and M_Y denote the moment generation functions of X and Y, respectively, the MGF of a random variable $Z \coloneqq X + Y$ is

  $$M_Z(\cdot) = M_X(\cdot) M_Y(\cdot).$$

Theorem 34 (Law of Iterated Expectations)**.** *Suppose $\mathbb{E}[|Y|] < \infty$. Then,*

$$\mathbb{E}_{Y,X}[Y] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[Y|X]\right].$$

Proof. Let $f_{X,Y}$ be the joint probability density of a bivariate random vector (X, Y). Then,

$$\begin{aligned}
\mathbb{E}[Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)\,\mathrm{d}y\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_X(x) f_{Y|X}(y|x)\,\mathrm{d}y\mathrm{d}x \\
&= \int_{-\infty}^{\infty} f_X(x) \underbrace{\int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,\mathrm{d}y}_{=\mathbb{E}[Y|X=x]}\mathrm{d}x \\
&= \mathbb{E}[\mathbb{E}[Y|X]].
\end{aligned}$$

■

Proposition 103 (Variance decomposition)**.** *Suppose* $\mathbb{E}[|Y|^2] < \infty$*. Then,*

$$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathbb{E}[Y|X]].$$

Proof. We use the definition of variance, linearity of expectation operator, and law of iterated expectations to obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}[Y|X]] &= \mathbb{E}\left[\left(\mathbb{E}\left[Y^2|X=x\right]\right) - \left(\mathbb{E}[Y|X=x]\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y^2|X=x\right]\right] - \mathbb{E}\left[\left(\mathbb{E}[Y|X=x]\right)^2\right] \\
&= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[\left(\mathbb{E}[Y|X=x]\right)^2\right] \\
&= \mathbb{E}\left[Y^2\right] - \left(\mathbb{E}[Y]\right)^2 + \left(\mathbb{E}[Y]\right)^2 - \mathbb{E}\left[\left(\mathbb{E}[Y|X=x]\right)^2\right] \\
&= \mathrm{Var}[Y] - \left(\left(\mathbb{E}[\mathbb{E}[Y|X=x]]\right)^2 - \mathbb{E}\left[\left(\mathbb{E}[Y|X=x]\right)^2\right]\right) \\
&= \mathrm{Var}[Y] - \mathrm{Var}[\mathbb{E}[Y|X=x]].
\end{aligned}$$

■

11.3.5 Covariance and correlation

Given two random variables X and Y such that $\mathbb{E}[|X|^2], \mathbb{E}[|Y|^2] < \infty$, the *covariance* between X and Y, denoted $\mathrm{Cov}(X, Y)$, is given by

$$\begin{aligned}
\mathrm{Cov}[X,Y] &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].
\end{aligned}$$

(make sure you can derive the second line from the first!) The *correlation* between X and Y, denoted $\mathrm{Corr}(X, Y)$, is given by

$$\mathrm{Corr}[X,Y] := \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}$$

If $\mathrm{Cor}[X, Y] = 0 \Leftrightarrow \mathrm{Cov}[X, Y] = 0$, we say that X and Y are *uncorrelated*. Otherwise, they are correlated. Covariance and correlation measure a particular kind of linear relationship between X and Y.

Theorem 35. *Let* X *and* Y *be two random variables* X *and* Y *such that* $\mathbb{E}[|X|^2], \mathbb{E}[|Y|^2] < \infty$*.*

(i) $-1 \leq \mathrm{Corr}[X, Y] \leq 1$*;*

(ii) $|\mathrm{Corr}[X, Y]| = 1$ *if and only if there exists a nonzero* $a \in \mathbb{R}$ *and* $b \in \mathbb{R}$ *such that* $aX + b = Y$ *holds with probability* 1*. Moreover, if* $\mathrm{Corr}[X, Y] = 1$ *(resp* $= -1$*), then* $a > 0$ *(resp.* $a < 0$*).*

Theorem 36. *Let X and Y be two random variables X and Y such that $\mathbb{E}[|X|^2], \mathbb{E}[|Y|^2] < \infty$. For any $a, b \in \mathbb{R}$,*

$$\text{Var}\left[aX + bY\right] = a^2\text{Var}\left[X\right] + b^2\text{Var}\left[Y\right] + 2ab\text{Cov}\left[X, Y\right].$$

If X and Y are uncorrelated, then

$$\text{Var}\left[aX + bY\right] = a^2\text{Var}\left[X\right] + b^2\text{Var}\left[Y\right].$$

11.4 n-dimensional random vectors

An n-dimensional random vector is a function from a sample space to $\mathbb{R}^n$ for some $n \in \mathbb{N}$.

Let $\mathbf{X}$ be an n-dimensional random vector. The *joint distribution function*, denoted F_X, of $\mathbf{X}$ is given by

$$F\left(\mathbf{x}\right) \coloneqq P\left(\mathbf{X} \leq \mathbf{x}\right) = P\left(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\right).$$

When $\mathbf{X}$ is discrete, the *joint pmf* is $f(\mathbf{x}) = P(\mathbf{X} = x)$. When $\mathbf{X}$ is continuous and differentiable, the *joint pdf* is given by

$$f\left(\mathbf{x}\right) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F\left(\mathbf{x}\right).$$

The joint pdf satisfies The expectation of an $n$dimensional random vector $\mathbf{X}$ is the (column) vector of expectation of its elements; i.e.,

$$\mathbb{E}\left[\mathbf{X}\right] = \begin{bmatrix} \mathbb{E}\left[X_1\right] \\ \mathbb{E}\left[X_2\right] \\ \vdots \\ \mathbb{E}\left[X_n\right] \end{bmatrix}.$$

The *variance-covariance matrix* of $\mathbf{X}$, denoted $\mathbf{\Sigma_X}$, is given by

$$\begin{aligned} \Sigma_{\mathbf{X}} \coloneqq \text{Var}\left[X\right] &= \mathbb{E}\left[\left(\mathbf{X} - \mathbb{E}\left[X\right]\right)\left(\mathbf{X} - \mathbb{E}\left[X\right]\right)^\top\right] \\ &= \begin{bmatrix} \text{Var}\left[X_1\right] & \text{Cov}\left[X_1, X_2\right] & \cdots & \text{Cov}\left[X_1, X_n\right] \\ \text{Cov}\left[X_2, X_1\right] & \text{Var}\left[X_2\right] & \cdots & \text{Cov}\left[X_2, X_n\right] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}\left[X_n, X_1\right] & \text{Cov}\left[X_n, X_2\right] & \cdots & \text{Var}\left[X_n\right] \end{bmatrix}. \end{aligned}$$

Theorem 37. *$\Sigma_{\mathbf{X}}$ is symmetric and positive semi-definite*

Proof. Symmetry of $\mathbf{\Sigma_X}$ follows from symmetry of $\text{Cov}(\cdot, \cdot)$ (check). For positive semi-definiteness, note that for any column vector $\mathbf{z} \in \mathbb{R}^n$,

$$\begin{aligned} \mathbf{z}^\top \Sigma_{\mathbf{X}} \mathbf{z} &= \mathbf{z}' \mathbb{E}\left[\left(\mathbf{X} - \mathbb{E}\left[X\right]\right)\left(\mathbf{X} - \mathbb{E}\left[X\right]\right)^\top\right] \mathbf{z} \\ &= \mathbb{E}\left[\mathbf{z}'\left(\mathbf{X} - \mathbb{E}\left[X\right]\right)\left(\mathbf{X} - \mathbb{E}\left[X\right]\right)^\top \mathbf{z}\right] \geq 0. \end{aligned}$$

■

Theorem 38. *Suppose an n-dimensional random vector $\mathbf{X}$ has expectation $\boldsymbol{\mu}$ and variance-covariance matrix $\Sigma_{\mathbf{X}}$. For any $\mathbf{A} \in \mathbb{R}^{q \times m}$, the random vector $\mathbf{AX}$ is a random vector with mean $\mathbf{A}\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}^\top$.*

Exercise 31. Prove the result above.

The marginal pdf (resp. pmf) of any subset of coordinates of $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ can be

computed by integrating out (resp. summing) the joint pdf (resp. pmf) over all possible values of other coordinates.

The conditional pdf (resp. pmf) of any subset of coordinates of $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ is obtained by dividing the joint pdf (resp. pmf) by the marginal pdf (resp. pmf) of the remaining coordinates.

Proposition 104 (Holder's Inequality)**.** *Let $p > 1$ and $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then for any random vectors X and Y*

$$\mathbb{E}\left|\mathbf{X}^\top \mathbf{Y}\right| \leq \left(\mathbb{E}\left[\|\mathbf{X}\|^p\right]\right)^{\frac{1}{p}} \left(\mathbb{E}\left[\|\mathbf{Y}\|^q\right]\right)^{\frac{1}{q}}.$$

Setting $p = q = 2$ gives the Cauchy-Schwarz inequality.

Proposition 105 (Minkowski's Inequality)**.** *For any random vectors $\mathbf{X}$ and $\mathbf{Y}$ and $p > 0$,*

$$\left(\mathbb{E}\left[\|\mathbf{X} + \mathbf{Y}\|^p\right]\right)^{\frac{1}{p}} \leq \left(\mathbb{E}\left[\|\mathbf{X}\|^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[\|\mathbf{Y}\|^p\right]\right)^{\frac{1}{p}}.$$

In general, for any random vector $\mathbf{X}$ and a constant vector $\mathbf{a}$ and $p > 0$,

$$\mathbb{E}\left[\|\mathbf{X} + \mathbf{a}\|^p\right] \leq C \cdot \left(\mathbb{E}\left[\|\mathbf{X}\|^p\right] + \|\mathbf{a}\|^p\right),$$

where C is a constant that depends on p. For example, if $\mathbf{a} = -\mathbb{E}[\mathbf{X}]$, then

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}\left[X\right]\|^p\right] &\leq C\left(\mathbb{E}\left[\|\mathbf{X}\|^p\right] + \|\mathbb{E}\left[X\right]\|^p\right) \\ &\leq 2C\mathbb{E}\left[\|\mathbf{X}\|^p\right].\end{aligned}$$

We can also about

$$\mathbb{E}\left[\|X + Y\|^p\right] \leq C\left(\mathbb{E}\left[\|\mathbf{X}\|^p\right] + \mathbb{E}\left[\|Y\|^p\right]\right),$$

where C depends on p.

11.5 Normal distribution

If Z is normally distributed, denoted

$$Z \sim N\left(\mu, \sigma^2\right),$$

then, Z has density and cumulative distribution functions given, respectively, by

$$\begin{aligned}f\left(z\right) = \phi\left(z\right) &\coloneqq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{\left(z-\mu\right)^2}{\sigma^2}\right], \\ F\left(z\right) = \Phi\left(z\right) &\coloneqq \int_{-\infty}^{\infty} \phi\left(z\right) \mathrm{d}z.\end{aligned}$$

Note that, there is no analytical expression for $\Phi(z)$; however, its existence is guaranteed by the fundamental theorem of calculus.

11.5.1 Bivariate normal

If Z and Y are two random variables that are *jointly normally distributed*, we say that Z and Y are *bivariate normal*. We denote this case as

$$\begin{bmatrix} Z \\ Y \end{bmatrix} \sim N\left(\boldsymbol{\mu}, \Sigma\right),$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Z \\ \mu_Y \end{bmatrix}, \ \Sigma = \begin{bmatrix} \sigma_Z^2 & \sigma_Z\sigma_Y\rho \\ \sigma_Z\sigma_Y\rho & \sigma_Y^2 \end{bmatrix}.$$

A nice property of bivariate normal random variables is that their linear combinations are also normally distributed.

Proposition 106. *For any $a, b \in \mathbb{R}$, $aZ + bY \sim N(a\mu_Z + b\mu_Y, a^2\sigma_Z^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_Z\sigma_Y)$.*

You can prove the following by "brute force", i.e., by writing out the density functions, but let us prove it using a property of OLS as the best linear predicator (BLP).

Recall that zero covariance does not imply independence between two random variables; however, if we know that the random variables are jointly normally distributed, then zero covariance does indeed imply independence.

Proposition 107. *If $\mathrm{Cov}[Z, Y] = \sigma_Z\sigma_Y\rho = 0$, then Z and Y are independent.*

Recall that the OLS estimator is given by

$$\left(X^\top X\right)^{-1} X^\top \mathbf{y}$$

where $X \in \mathbb{R}^{m\times n}$, $\widehat{\mathbf{b}} \in \mathbb{R}^{n\times 1}$ and $\mathbf{y} \in \mathbb{R}^{m\times 1}$. The estimator solves the a linear system of equations given by $X\mathbf{b} = \mathbf{y}$ (which may not have a solution) as "best" as possible. Now, suppose we treat X and $\mathbf{y}$ as random—in particular, we think of the system as consisting of m realisations from a random variable vector $X \in \mathbb{R}^{1\times n}$ and a random variable $y \in \mathbb{R}$. Recall that

$$\begin{aligned} \mathrm{Var}\left[X\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^\top \left(X - \mathbb{E}\left[X\right]\right)\right] \in \mathbb{R}^{n\times n} \\ \mathrm{Cov}\left[X, y\right] &= \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^\top \left(y - \mathbb{E}\left[y\right]\right)\right] \in \mathbb{R}^{n\times 1}. \end{aligned}$$

Then, the expectation of the OLS estimator of $\widehat{X}\mathbf{b}^\top = \widehat{y}$, where $\widehat{X} = X - \mathbb{E}[X]$ and $\widehat{y} = y - \mathbb{E}[y]$, is given by

$$\mathbb{E}\left[\left(\widehat{X}^\top\widehat{X}\right)^{-1}\widehat{X}^\top\widehat{y}\right] = \left(\mathrm{Var}\left[X\right]\right)^{-1}\mathrm{Cov}\left[X, y\right].$$

Proposition 108. *$Z|Y \sim N(\mu_Z + \beta(Y - \mu_Y), \sigma_Z^2(1 - \rho^2))$, where $\beta \coloneqq \sigma_Z\rho/\sigma_Y$.*

Proof using BLP. We exploit the property of the Best Linear Predictor (BLP) and joint normality. In particular, BLP tells us that the residual is orthogonal (i.e., uncorrelated) with the independent variable. Joint normality then tell us that, in fact, the independent variable and the residuals are independent.

Let β be the best linear predictor of a regression of Z on Y; i.e.,

$$Z - \mu_Z = \beta\left(Y - \mu_Y\right) + U,$$

where

$$\beta \coloneqq \frac{\mathrm{Cov}\left[Z - \mu_Z, Y - \mu_Y\right]}{\mathrm{Var}\left[Y - \mu_Y\right]} = \frac{\sigma_Z\sigma_Y\rho}{\sigma_Y^2} = \frac{\sigma_Z}{\sigma_Y}\rho.$$

Then, U is orthogonal to Y:

$$\begin{aligned} \mathrm{Cov}\left[Y - \mu_Y, U\right] &= \mathrm{Cov}\left[Y - \mu_Y, Z - \mu_Z - \beta\left(Y - \mu_Y\right)\right] = \mathrm{Cov}\left[Y, Z\right] - \beta\,\mathrm{Var}\left[Y\right] \\ &= \sigma_Z\sigma_Y\rho - \frac{\sigma_Z}{\sigma_Y}\rho\sigma_Y^2 = 0. \end{aligned}$$

Since U is a linear function of Z and Y which are jointly normal, U is also normally distributed (and jointly normally distributed with Y). That Y and U are jointly normal together with the fact that $\text{Cov}[Y, U] = 0$ implies that Y and U are, in fact, independent. Therefore, the distribution of U is the same as the distribution of $U|Y$. The distribution of U is given by $U \sim N(0, \sigma_Z^2(1-\rho^2))$ since

$$\begin{aligned}
\mathbb{E}[U] &= \mathbb{E}[Z - \mu_Z - \beta(Y - \mu_Y)] = 0, \\
\text{Var}[U] &= \text{Var}[Z - \mu_Z - \beta(Y - \mu_Y)] = \text{Var}[Z] + \beta^2 \text{Var}[Y] - 2\beta \text{Cov}[Z, Y] \\
&= \sigma_Z^2 + \frac{\sigma_Z^2}{\sigma_Y^2}\rho^2\sigma_Y^2 - 2\frac{\sigma_Z}{\sigma_Y}\rho\sigma_Z\sigma_Y\rho = \sigma_Z^2\left(1-\rho^2\right).
\end{aligned}$$

Finally, given that

$$Z = \mu_Z + \beta(Y - \mu_Y) + U,$$

it follows that $Z|Y \sim N(\mu_Z + \beta(Y - \mu_Y), \sigma_Z^2(1-\rho^2))$. ■

11.5.2 Multivariate normal

We now extend the result above to the multivariate case. Suppose

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \boldsymbol{\Sigma} \\ \Sigma_{12}^\top & \Sigma_{22} \end{bmatrix}\right),$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are $k_1 \times 1$ and $k_2 \times 1$ random vectors with $k_1 + k_2 = k$. Our goal is to find the distribution of $\mathbf{Z}_1$ conditional on $\mathbf{Z}_2$.

Consider running the (population) regression:

$$\mathbf{Z}_1 = \boldsymbol{\beta}\mathbf{Z}_2 + \mathbf{U}.$$

The BLP, $\boldsymbol{\beta}$, is given by

$$\begin{aligned}
\boldsymbol{\beta} &= \text{Cov}[\mathbf{Z}_1, \mathbf{Z}_2] \text{Var}[\mathbf{Z}_2]^{-1} \\
&= \Sigma_{12}\Sigma_{22}^{-1}.
\end{aligned}$$

We can therefore define the residual as

$$\mathbf{U} = \mathbf{Z}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Z}_2.$$

We can show that $\mathbf{U}$ is uncorrelated with $\mathbf{Z}_2$:

$$\begin{aligned}
\text{Cov}[\mathbf{U}, \mathbf{Z}_2] &= \text{Cov}\left[\mathbf{Z}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Z}_2, \mathbf{Z}_2\right] \\
&= \text{Cov}[\mathbf{Z}_1, \mathbf{Z}_2] - \Sigma_{12}\Sigma_{22}^{-1}\text{Var}[\mathbf{Z}_2] \\
&= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = \mathbf{0}.
\end{aligned}$$

Moreover, since $\mathbf{U}$ is a linear combinations of the components of the normal vector $\mathbf{Z}$, it is jointly normally distributed with $\mathbf{Z}_2$. Hence, $\mathbf{U}$ is independent of of $\mathbf{Z}_2$.

The expected value of $\mathbf{U}$ is

$$\begin{aligned}
\mathbb{E}[\mathbf{U}] &= \mathbb{E}[\mathbf{Z}_1] - \Sigma_{12}\Sigma_{22}^{-1}\mathbb{E}[\mathbf{Z}_2] \\
&= \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2.
\end{aligned}$$

Define

$$\boldsymbol{\varepsilon} \coloneqq \mathbf{U} - \mathbb{E}\left[\mathbf{U}\right],$$

then

$$\begin{aligned} \boldsymbol{\varepsilon} &= \mathbf{Z}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Z}_2 - \left(\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2\right) \\ \Rightarrow \mathbf{Z}_1 - \boldsymbol{\mu}_1 &= \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Z}_2 - \boldsymbol{\mu}_2\right) + \boldsymbol{\varepsilon}, \end{aligned}$$

where $\boldsymbol{\varepsilon}$ is mean zero and (still) independent of $\mathbf{Z}_2$.

The variance of $\boldsymbol{\varepsilon}$ is given by

$$\begin{aligned} \operatorname{Var}\left[\boldsymbol{\varepsilon}\right] &= \operatorname{Var}\left[\mathbf{U} - \mathbb{E}\left[\mathbf{U}\right]\right] = \operatorname{Var}\left[\mathbf{W}\right] = \operatorname{Var}\left[\mathbf{Z}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Z}_2\right] \\ &= \operatorname{Var}\left[\mathbf{Z}_1\right] + \Sigma_{12}\Sigma_{22}^{-1}\operatorname{Var}\left[\mathbf{Z}_2\right]\left(\Sigma_{12}\Sigma_{22}^{-1}\right)' \\ &\quad - \Sigma_{12}\Sigma_{22}^{-1}\operatorname{Cov}\left[\mathbf{Z}_1, \mathbf{Z}_2\right] - \operatorname{Cov}\left[\mathbf{Z}_1, \mathbf{Z}_2\right]\left(\Sigma_{12}\Sigma_{22}^{-1}\right)' \\ &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{12}' - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}. \end{aligned}$$

Thus, we realise that

$$\mathbf{Z}_1|\mathbf{Z}_2 \sim N\left(\boldsymbol{\mu}_1\Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Z}_2 - \boldsymbol{\mu}_2\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}\right)$$

because the variation in $\mathbf{Z}_1$ fixing $\mathbf{Z}_2$ are all driven by variations in $\boldsymbol{\varepsilon}$ which is independent of $\mathbf{Z}_2$.

Moreover, the conditional expectation of $\mathbf{Z}_1 - \boldsymbol{\mu}_1$ given $\mathbf{Z}_2 - \boldsymbol{\mu}_2$ is linear and the coefficients are the coefficient of a linear projection of $\mathbf{Z}_1 - \boldsymbol{\mu}_1$ on $\mathbf{Z}_2 - \boldsymbol{\mu}_2$:

$$\mathbb{E}\left[\mathbf{Z}_1|\mathbf{Z}_2\right] - \boldsymbol{\mu}_1 = \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{Z}_2 - \boldsymbol{\mu}_2\right).$$

11.5.3 Log-normal distribution

A random variable Z is log-normally distributed if $\ln Z$ is normally distributed. The next result is useful when you have a constant absolute risk aversion utility function, i.e., $u(x) = -a\exp(-ax)$ and a normally distributed risk or a constant relative risk aversion utility function, i.e., $u(x) = x^{1-a}/(1-a)$ with $a > 0$, and a log-normally distributed risk.

Proposition 109. *Suppose Z is log-normally distributed; i.e., $\ln Z \sim N(\mu, \sigma^2)$. Then,*

$$\mathbb{E}\left[Z^{-\gamma}\right] = \exp\left[-\gamma\mu + \frac{1}{2}\gamma^2\sigma^2\right].$$

Proof. Define $z \coloneqq \ln Z$.

$$\begin{aligned}
\mathbb{E}\left[Z^{-\gamma}\right] &= \mathbb{E}\left[\exp\left(\ln\left(Z^{-\gamma}\right)\right)\right] = \mathbb{E}\left[\exp\left[-\gamma z\right]\right] \\
&= \int_{-\infty}^{\infty} \exp\left[-\gamma z\right] \phi\left(z\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\gamma z\right] \exp\left[-\frac{\left(z-\mu\right)^2}{2\sigma^2}\right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{2\sigma^2\gamma z + z^2 + \mu^2 - 2z\mu}{2\sigma^2}\right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{z^2 + 2\left(-\mu + \sigma^2\gamma\right) z + \mu^2}{2\sigma^2}\right] \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{\left(z + \left(-\mu + \sigma^2\gamma\right)\right)^2 - \left(-\mu + \sigma^2\gamma\right)^2 + \mu^2}{2\sigma^2}\right] \mathrm{d}z \\
&= \exp\left[-\frac{-\left(\mu - \sigma^2\gamma\right)^2 + \mu^2}{2\sigma^2}\right] \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{\left(z - \left(\mu - \sigma^2\gamma\right)\right)^2}{2\sigma^2}\right] \mathrm{d}z}_{=1} \\
&= \exp\left[-\frac{-\mu^2 - \sigma^4\gamma^2 + 2\mu\sigma^2\gamma + \mu^2}{2\sigma^2}\right] \\
&= \exp\left[-\gamma\mu + \frac{1}{2}\gamma^2\sigma^2\right].
\end{aligned}$$

■

11.6 Other relatives of normal distributions

- Let $(X_i)_{i=1}^n$ be a collection of n random variables that are iid normally distributed. Then, $\sum_{i=1}^n X_i^2$ has a *chi-squared* distribution with n degrees of freedom.
- Let X and Y be two independent random variables that has a chi-squared distribution with degrees of freedoms n_x and n_y respectively. Then,

  $$F \coloneqq \frac{\frac{X}{n_x}}{\frac{Y}{n_y}}$$

  has an F-distribution.
- If $U \sim N(0,1)$ and $V \sim N(0,1)$ and they are independent, then $X \coloneqq \frac{U}{V}$ is has *Cauchy* distribution with pdf

  $$f\left(x\right) = \frac{1}{\pi\left(1 + x^2\right)}.$$

  Cauchy distribution has no integer moments.

12 Relationships between common distributions

12.1 Uniform distribution and other distributions

Let F be a CDF. Then, the *quantile function* of F is $F^{-1} : \mathbb{R} \to \mathbb{R}$ defined as

$$F^{-1}(\alpha) \coloneqq \min\{z \in \mathbb{R} : F(z) \geq \alpha\}.$$

Observe that F^{-1} is well-defined because F is nondecreasing and right-continuous, which imply that the set of values with cumulative probability equal to or higher than α is closed.

The following tells us that we can build random variable with any CDF F from a uniformly distributed random variable.

Proposition 110. *Let $U \sim$ Uniform$[0,1]$ and F be a CDF. Then, $Z = F^{-1}(U)$ is a random variable with CDF F.*

Exercise 32. Consider a logistic random variable with CDF

$$F(x) = \frac{1}{1+e^{-x}}.$$

Show that $X = \ln(\frac{U}{1-U})$ is a logistic random variable when U is uniformly distributed on $[0,1]$.

12.1.1 Bernoulli, Binomial and Poisson distributions

A random variable Z_i follows a *Bernoulli distribution* with parameter p, $Z \sim$ Bernoulli(p), if it takes the value 1 with probability p and 0 with probability $1-p$. Its probability mass function is given by

$$f_{Z_i}(z) = p^z (1-p)^{1-z} = pz + (1-p)(1-z) \;\; \forall z \in \{0,1\}$$

and the cumulative distribution function is given by

$$F_{Z_i}(z) = \begin{cases} 0 & \text{if } z < 0 \\ 1-p & \text{if } 0 \leq z < 1 \\ 1 & \text{if } z \geq 1 \end{cases}.$$

We think of Bernoulli distribution describing a single coin toss and p as the probability that we see (say) head. More generally, we think of Bernoulli as describing the outcome of a single experiment with binary outcomes. The first two moments are:

$$\begin{aligned} \mathbb{E}[Z] &= 1 \cdot p + 0 \cdot (1-p) = p, \\ \operatorname{Var}[Z] &= \mathbb{E}\left[Z^2\right] - (\mathbb{E}[Z])^2 = \mathbb{E}[Z](1 - \mathbb{E}[Z]) = p(1-p), \end{aligned}$$

A random variable Z follows a *Binomial distribution* with parameters (n,p), denoted $Z \sim$ Binomial(n,p), if its probability mass function is given by

$$f_Z(z) = \binom{n}{z} p^z (1-p)^{n-z}$$

and the cumulative distribution function is given by

$$F_Z(z) = \sum_{i=1}^{\lfloor z \rfloor} \binom{n}{i} p^i (1-p)^{n-i}.$$

We think of binomial distribution as describing the distribution of the number of heads when we toss a coin n times independently. More generally, it describes the distribution of the number of successes in n independently identically distributed experiments with binary outcomes. Note that

$$\text{Binomial}(1,p) \overset{\text{d}}{=} \text{Bernoulli}(p).$$

If $(Z_i)_{i=1}^n$ is a collection of independently identically distributed random variables that each follows Bernoulli(p), then

$$\sum_{i=1}^{n} Z_i \overset{\text{d}}{=} \text{Binomial}\left(n,p\right).$$

The first two moments are

$$\begin{aligned}\mathbb{E}\left[Z\right] &= \mathbb{E}\left[\sum_{i=1}^{n} Z_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[Z_i\right] = np,\\ \text{Var}\left[Z\right] &= np\left(1-p\right).\end{aligned}$$

A random variable Y follows a Poisson distribution with parameter λ, denoted $Y \sim \text{Poisson}(\lambda)$, if its probability mass function is given by

$$f_Y\left(y\right) = \frac{\lambda^y e^{-\lambda}}{y!} \ \forall y \in \{0\} \cup \mathbb{N}.$$

and the cumulative distribution function is given by

$$F_Y\left(y\right) = e^{-\lambda} \sum_{i=1}^{\lfloor y \rfloor} \frac{\lambda^i}{i!}.$$

We think of Poisson distribution as describing the probability of a given number of occurrences of an event in a fixed amount of time when the events occur at a constant rate independently of the time of the last event. Its first two moments are

$$\mathbb{E}\left[Y\right] = \lambda = \text{Var}\left[Y\right].$$

To shows this, use the Mclaurin series for the exponential function from (5.2): $e^x = \sum_{n=0}^{\infty} x^n/n!$. Then,

$$\sum_{y=0}^{\infty} f\left(y\right) = \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} e^{\lambda} = 1.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[Y\right] &= \sum_{y=0}^{\infty} y f_Y\left(y\right) = \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} = \sum_{y=1}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!}\\ &= \lambda e^{-\lambda} \sum_{y=1}^{\infty} \frac{\lambda^{y-1}}{(y-1)!} = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}\\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda.\end{aligned}$$

Similar computations gives that $\text{Var}[Y] = \lambda$ (try it!).

The probability mass function for Poisson distribution can also be written recursively as follows:

$$\Pr(Y=y)=\frac{\lambda^y e^{-\lambda}}{y!}=\frac{\lambda}{y}\frac{\lambda^{y-1}e^{-\lambda}}{(y-1)!}=\frac{\lambda}{y}\Pr(Y=y-1), \tag{12.1}$$

where $\Pr(Y=y)$ is the probability measure associated with the event $Y=y$. When n is large and p is small, Poisson distribution approximates binomial distribution with parameters (n,p). To see this, suppose $Z\sim\text{Binomial}(n,p)$. Then,

$$\begin{aligned}\Pr(Z=z)&=\binom{n}{z}p^z(1-p)^{n-z}=\frac{n!}{z!(n-z)!}p^z(1-p)^{n-z}\\&=\frac{p}{1-p}\frac{n-z+1}{z}\frac{n!}{(z-1)!(n-(z-1))!}p^{z-1}(1-p)^{n-(z-1)}\\&=\frac{p}{1-p}\frac{n-z+1}{y}\Pr(Z=z-1)\\&=\frac{np-p(y-1)}{z-pz}\Pr(Z=z-1).\end{aligned}$$

Letting $\lambda=np$, for small p, we have

$$\Pr(Z=z)=\frac{\lambda-p(z-1)}{y-pz}\Pr(Z=z-1)\approx\frac{\lambda}{z}\Pr(Z=z-1),$$

which corresponds to the Poisson recursion in (12.1). It suffices therefore to establish that $\Pr(Z=0)\approx\Pr(Y=0)$. Since

$$\Pr(Z=0)=(1-p)^n=\left(1-\frac{np}{n}\right)^n=\left(1-\frac{\lambda}{n}\right)^n$$

and $\lim_{n\to\infty}(1-\lambda/n)^n=\exp[-\lambda]$, for sufficiently large n. we have

$$\Pr(Z=0)\approx\exp(-\lambda)=\Pr(Y=0).$$

Thus, when n is large and p is small (i.e., when the number of experiment per period is large and the probability of event occurring is small), then

$$\lim_{n\to\infty:np=\lambda}Z\overset{\text{d}}{=}\lim_{n\to\infty:np=\lambda}\sum_{i=1}^{n}Z_i\overset{\text{d}}{=}\text{Poisson}(\lambda).$$

12.1.2 Poisson and exponential

A random variable Z follows an exponential distribution with parameter λ if its probability density function is given by

$$f_Z(z)=\lambda e^{-\lambda z}$$

and and the cumulative distribution function is given by

$$F_Z(z)=1-e^{-\lambda z}.$$

Its first two moments are given by

$$\mathbb{E}[Z]=\frac{1}{\lambda},\ \text{Var}[Z]=\frac{1}{\lambda^2}.$$

Recall that Poisson distribution describes the probability of a given number of occurrences of an event in a fixed amount of time. The exponential distribution describes the length of time between the occurrences of these events (which is constant).

To see this, let Δ be the "fixed amount of time" and suppose that the event occurs at rate λ at any point in time. Then, on average, there will be $\lambda\Delta$ occurrences of events per Δ units of time. The Poisson distribution associated with this is given by

$$\Pr(Y = y) = \frac{(\lambda\Delta)^y e^{-\lambda\Delta}}{y!}$$

and $\Pr(Y = 0) = e^{-\lambda\Delta}$ is the probability of no occurrence during Δ units of time. Now let Z describe the time it takes for the first occurrence of the event. Then, the probability that the event occurs after Δ units of time is equivalent to the probability of no occurrence of the event during Δ units of time; i.e., $\Pr(Z > \Delta) = \Pr(Y = 0)$. Thus, we can write the probability that the event occurs by period Δ is

$$\Pr(Z \leq \Delta) = 1 - \Pr(Z > \Delta) = 1 - \Pr(Y = 0) = 1 - e^{-\lambda\Delta}.$$

Hence, Z follows an exponential distribution with parameter λ. Note that since Poisson is *memory-less* (i.e., the rate of arrival is constant), time between any two events follows an exponential distribution with parameter λ.

Remark 26. Incidentally, suppose that Z_1 and Z_2 follows Poisson distribution with parameter λ_1 and λ_2 respectively and they are independent. We can interpret Z_1 and Z_2 as describing the time until first occurrence of event of types 1 and 2 respectively. Let $Z \coloneqq Z_1 + Z_2$, then

$$\begin{aligned}
F_Z(z) &= \Pr(Z_1 + Z_2 \leq z) = \Pr(\{Z_1 \leq z\} \cap \{Z_2 \leq z - Z_1\}) \\
&= \sum_{i=0}^{z} \Pr(\{Z_1 = i\} \cap \{Z_2 = z - i\}) = \sum_{i=0}^{z} \Pr(\{Z_1 = i\}) \Pr(\{Z_2 = z - i\}) \\
&= \sum_{i=0}^{z} \frac{\lambda_1^i e^{-\lambda_1}}{i!} \frac{\lambda_2^{z-i} e^{-\lambda_2}}{(z-i)!} = \sum_{i=0}^{z} \left(\frac{z!}{i!\,(z-i)!} \lambda_1^i \lambda_2^{z-i} \right) \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \\
&= \left(\sum_{i=0}^{z} \binom{z}{i} \lambda_1^j \lambda_2^{z-j} \right) \frac{e^{-(\lambda_1+\lambda_2)}}{z!} = \frac{(\lambda_1 + \lambda_2)^z e^{-(\lambda_1+\lambda_2)}}{z!},
\end{aligned}$$

where, in second line, we use the fact that Z_1 and Z_2 are independent, and in the last line, we used the fact that

$$(\lambda_1 + \lambda_2)^z = \sum_{i=0}^{z} \binom{z}{i} \lambda_1^j \lambda_2^{z-j} \ \forall z \in \{0\} \cup \mathbb{N}.$$

Thus, $Z = Z_1 + Z_2$ follows a Poisson distribution with parameter $\lambda_1 + \lambda_2$. It follows that the time until the first event of either type 1 or 2 occurs is given by exponential distribution with parameter $\lambda_1 + \lambda_2$.

By independence, the probability density for (Z_1, Z_2) is given by

$$f_{(Z_1,Z_2)}(z_1, z_2) = f_{Z_1}(z_1) f_{Z_2}(z_2) = \lambda_1 e^{-\lambda_1 z_1} \lambda_2 e^{-\lambda_2 z_2} = \lambda_1 \lambda_2 e^{-(\lambda_1 z_1 + \lambda_2 z_2)}.$$

Then the probability that event associated with Z_1 occurs before the event associated with Z_2 is

given by

$$\begin{aligned}\Pr\left(Z_1 < Z_2\right) &= \int_{(z_1,z_2)\in[0,\infty)\times[z_1,\infty)} f_{(Z_1,Z_2)}\left(z_1,z_2\right)\mathrm{d}\left(z_1,z_2\right)\\ &= \int_0^\infty \int_{z_1}^\infty \lambda_1\lambda_2 e^{-(\lambda_1 z_1+\lambda_2 z_2)}\mathrm{d}z_2\mathrm{d}z_1 = \int_0^\infty \lambda_1 e^{-\lambda_1 z_1}\left(\int_{z_1}^\infty \lambda_2 e^{-\lambda_2 z_2}\mathrm{d}z_2\right)\mathrm{d}z_1\\ &= \int_0^\infty \lambda_1 e^{-\lambda_1 z_1}\left(e^{-\lambda_2 z_1}\right)\mathrm{d}z_1 = \int_0^\infty \lambda_1 e^{-(\lambda_1+\lambda_2)z_1}\mathrm{d}z_1\\ &= -\frac{\lambda_1}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)z_1}\Big|_0^\infty = \frac{\lambda_1}{\lambda_1+\lambda_2},\end{aligned}$$

where we used Fubini's theorem in the second line.

12.2 Gamma distribution

Recall the gamma function:

$$\Gamma\left(\alpha\right) \coloneqq \int_0^\infty e^{-t}t^{\alpha-1}\mathrm{d}t,\ \alpha>0.$$

Observe that the integrand, $e^{-t}t^{x-1}$ is nonnegative, hence it follows that

$$f\left(t\right) \coloneqq \frac{e^{-t}t^{\alpha-1}}{\Gamma\left(\alpha\right)}\ \forall t\in(0,\infty)$$

is a probability density function. Now apply the change of variables via $x \coloneqq \beta t \Leftrightarrow t = x/\beta$ so that $\mathrm{d}t = \beta^{-1}\mathrm{d}x$. Hence,

$$\int_0^\infty e^{-t}t^{\alpha-1}\mathrm{d}t = \frac{1}{\beta}\int_0^\infty e^{-\frac{x}{\beta}}\left(\frac{x}{\beta}\right)^{\alpha-1}\mathrm{d}x$$

and so

$$f\left(x|\alpha,\beta\right) = \frac{e^{-\frac{x}{\beta}}\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta^{-1}\Gamma\left(\alpha\right)} = \frac{1}{\Gamma\left(\alpha\right)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$$

for all $x\in(0,\infty)$, $\alpha>0$ and $\beta>0$. Above is the probability density associated with *gamma distribution with parameters* (α,β). We refer to α as the *shape parameter* since it mostly influences the peakedness of the distribution and β as the *scale parameter* since it mostly influences the spread of the distribution.

The first two moments of a random variable Z that follows a gamma distribution with parameters (α,β), denoted $Z\sim\mathrm{Gamma}(\alpha,\beta)$, are given by

$$\mathbb{E}\left[Z\right] = \alpha\beta,\ \mathrm{Var}\left[Z\right] = \alpha\beta^2.$$

Gamma distribution relates to many commonly used distributions. Let $Z\sim\mathrm{Gamma}(\alpha,\beta)$.

- Suppose α is an integer and $Y\sim\mathrm{Poisson}(x/\beta)$. Then, for any $x\in(0,\infty)$,

$$\Pr\left(Z\le z\right) = \Pr\left(Y\ge\alpha\right).$$

- Suppose $\alpha = p/2$ for some integer p and $\beta=2$, then Z follows a χ^2-distribution with p degrees of freedom.

- Suppose $\alpha=1$, then Z follows an exponential distribution with parameter $1/\beta$.

12.3 Conjugate distribution property

Consider a problem of inferring a distribution for a parameter $\theta \in \Theta$ given some data Z. The posterior belief is given by $f(\theta|Z)$, which, by Bayes rule, can be written as

$$f(\theta|Z=z) = \frac{f(z|\theta)\,\pi(\theta)}{\int_{\tilde{\theta}\in\Theta} f\left(z|\tilde{\theta}\right)\pi\left(\tilde{\theta}\right)\mathrm{d}\tilde{\theta}},$$

where π is the prior belief over Θ and $f(Z|\theta)$ is the conditional density of Z given θ. If the posterior distribution, $f(\theta|Z=z)$, is in the same probability distribution family as the prior probability distribution, $\pi(\theta)$, the prior and posterior are then called *conjugate distributions* and the prior is called a *conjugate prior*. We will demonstrate below that normally distributed prior gives rise to normally distributed posterior.

At each date $t \in \mathbb{N}$, suppose that a Bayesian decision maker observes $Z_t = \theta + W_t$, where each W_t are identically and independently distributed according to $N\left(0,\sigma^2\right)$. Let the decision maker's time-t belief be that $\theta \sim N(\mu_t, \gamma_t^2)$ after observing $Z_1, Z_2, \ldots, Z_t$, but before observing Z_{t+1}.

The density conditional on θ and observed values of Z_t's is given by

$$\psi(Z_{t+1}=z|\theta, Z_t, Z_{t-1}, \ldots, Z_1) = \psi(Z_{t+1}=z|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(z-\theta)^2}{2\sigma^2}\right],$$

where that Z_t's are iid implies the first equality. The prior density for θ is given by

$$\pi_t(\theta) = \frac{1}{\sqrt{2\pi\gamma_t^2}}\exp\left[-\frac{(\theta-\mu_t)^2}{2\gamma_t^2}\right].$$

By Bayes rule, posterior density is given by

$$\begin{aligned}\pi_{t+1}(\theta|Z_{t+1}=z) &= \frac{\psi(Z_{t+1}=z|\theta, Z_t, Z_{t-1}, \ldots, Z_1)\,\pi_t(\theta)}{\int_{\mathbb{R}}\psi(Z_{t+1}=z|\theta, Z_t, Z_{t-1}, \ldots, Z_1)\,\pi_t(\theta)\,\mathrm{d}\theta} \\ &= \frac{\psi(Z_{t+1}=z|\theta)\,\pi_t(\theta)}{\int_{\mathbb{R}}\psi(Z_{t+1}=z|\theta)\,\pi_t(\theta)\,\mathrm{d}\theta},\end{aligned}$$

where the second equality follows from the fact that Z_{t+1} is independent of past values conditional

on θ. Consider the term

$$
\begin{aligned}
&\psi\left(Z_{t+1}=z|\theta\right)\pi_t\left(\theta\right)\\
&=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(z-\theta\right)^2}{2\sigma^2}\right]\right)\left(\frac{1}{\sqrt{2\pi\gamma_t^2}}\exp\left[-\frac{\left(\theta-\mu_t\right)^2}{2\gamma_t^2}\right]\right)\\
&=\frac{1}{2\pi\sigma\gamma_t}\exp\left[-\frac{1}{2\gamma_t^2\sigma^2}\left(\gamma_t^2\left(z-\theta\right)^2+\sigma^2\left(\theta-\mu_t\right)^2\right)\right]\\
&=\frac{1}{2\pi\sigma\gamma_t}\exp\left[-\frac{1}{2\gamma_t^2\sigma^2}\left(\gamma_t^2\left(z^2-2\theta z+\theta^2\right)+\sigma^2\left(\theta^2-2\theta\mu_t+\mu_t^2\right)\right)\right]\\
&=\frac{1}{2\pi\sigma\gamma_t}\exp\left[-\frac{1}{2\gamma_t^2\sigma^2}\left(\left(\gamma_t^2+\sigma^2\right)\theta^2-2\theta\left(\gamma_t^2 z+\sigma^2\mu_t\right)+\gamma_t^2 z^2+\sigma^2\mu_t^2\right)\right]\\
&=\frac{1}{2\pi\sigma\gamma_t}\exp\left[-\frac{1}{2\gamma_t^2\sigma^2}\left(\left(\gamma_t^2+\sigma^2\right)\left(\theta^2-2\theta\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)+\gamma_t^2 z^2+\sigma^2\mu_t^2\right)\right]\\
&=\frac{1}{2\pi\sigma\gamma_t}\exp\left[-\frac{1}{2\gamma_t^2\sigma^2}\left(\left(\gamma_t^2+\sigma^2\right)\left(\theta^2-2\theta\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}+\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2-\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2\right)\right.\right.\\
&\qquad\qquad\left.\left.\times\quad\gamma_t^2 z^2+\sigma^2\mu_t^2\right)\right]\\
&=\frac{1}{2\pi\sigma\gamma_t}\exp\left[-\frac{1}{2\gamma_t^2\sigma^2}\left(\left(\gamma_t^2+\sigma^2\right)\left(\theta-\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2-\left(\gamma_t^2+\sigma^2\right)\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2+\gamma_t^2 z^2+\sigma^2\mu_t^2\right)\right]\\
&=\frac{1}{\sqrt{2\pi\left(\gamma_t^2+\sigma^2\right)}}\exp\left[-\frac{-\left(\gamma_t^2+\sigma^2\right)\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2+\gamma_t^2 z^2+\sigma^2\mu_t^2}{2\gamma_t^2\sigma^2}\right]\\
&\quad\times\underbrace{\frac{1}{\sqrt{2\pi\frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2}}}\exp\left[-\frac{1}{2}\frac{\left(\theta-\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2}{\frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2}}\right]}_{\sim N\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2},\frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2}\right)}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\int_\Theta\psi\left(Z_{t+1}=z|\theta\right)\pi_t\left(\theta\right)d\theta\\
&=\frac{1}{\sqrt{2\pi\left(\gamma_t^2+\sigma^2\right)}}\exp\left[-\frac{-\left(\gamma_t^2+\sigma^2\right)\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2+\gamma_t^2 z^2+\sigma^2\mu_t^2}{2\gamma_t^2\sigma^2}\right]\underbrace{\int_\Theta\varphi\left(\theta\right)d\theta}_{=1}\\
&=\frac{1}{\sqrt{2\pi\left(\gamma_t^2+\sigma^2\right)}}\exp\left[-\frac{-\left(\gamma_t^2+\sigma^2\right)\left(\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2+\gamma_t^2 z^2+\sigma^2\mu_t^2}{2\gamma_t^2\sigma^2}\right].
\end{aligned}
$$

Thus, the posterior is given by

$$
\pi_{t+1}\left(\mathrm{d}\theta|Z_{t+1}=z\right)=\frac{1}{\sqrt{2\pi\frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2}}}\exp\left[-\frac{1}{2}\frac{\left(\theta-\frac{\gamma_t^2 z+\sigma^2\mu_t}{\gamma_t^2+\sigma^2}\right)^2}{\frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2}}\right].
$$

We therefore realise that the posterior is normally distributed according to:

$$\pi_{t+1}(\theta) \sim N\left(\mu_{t+1}, \frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2}\right).$$

Note when prior and posterior distributions are the same, they are called conjugate priors.

$$\gamma_{t+1}^2 = \frac{\gamma_t^2\sigma^2}{\gamma_t^2+\sigma^2},\ \mu_{t+1} = \frac{\gamma_t^2 z + \sigma^2\mu_t}{\gamma_t^2+\sigma^2}. \tag{12.2}$$

We can define the expression in terms of precisions, defined as the inverse of the variance:

$$\tau_t^\gamma \coloneqq \frac{1}{\gamma_t^2},\ \tau^\sigma \coloneqq \frac{1}{\sigma^2}.$$

Then,

$$\tau_{t+1}^\gamma = \frac{\gamma_t^2+\sigma^2}{\gamma_t^2\sigma^2} = \frac{1}{\sigma^2} + \frac{1}{\gamma_t^2} = \tau^\sigma + \tau_t^\gamma.$$

Thus, the precision of the posterior is simply the sum of the precision of the prior and $Z_{t+1}|\theta$.

$$\mu_{t+1} = \frac{1}{1+\frac{\sigma^2}{\gamma_t^2}}z + \frac{1}{\left(\frac{\sigma^2}{\gamma_t^2}\right)^{-1}+1}\mu_t = \frac{\tau^\sigma}{\tau^\sigma+\tau_t^\gamma}z + \left(1-\frac{\tau^\sigma}{\tau^\sigma+\tau_t^\gamma}\right)\mu_t.$$

Thus, we see that the posterior mean is a weighted average of the observed $Z_{t+1} = z$ and the prior, where the weights are given by the ratio of precision of between the prior and $Z_{t+1}|\theta$. If the prior is relatively more (less) precise, then the weights placed on $Z_{t+1} = z$ is smaller (larger).

13 Stochastic dominance

Our goal is to give a partial ordering of a set **F** of CDFs. All of the ordering we introduce below are partial orders on **F** (or equivalently, on the set of random variables distributed according to some $F \in \mathbf{F}$).

13.1 First-order stochastic dominance

Proposition 111. *Let F and G be two CDFs with a common support $[a, b]$. The following are equivalent.*

(i) F first-order stochastically dominates *(FOSD) G; i.e., $F \geq_{FOSD} G$.*[39]

(ii) $F(t) \leq G(t)\ \forall t \in [a, b]$.

(iii) $\mathbb{E}[u(Z)] \geq \mathbb{E}[u(Y)]$ *for any increasing function $u : [a, b] \to \mathbb{R}$, where $Z \sim F$ and $Y \sim G$.*

Let $Z \sim F$ and $Y \sim G$. To make sense of (ii), rewrite the condition as $1 - F(t) \geq 1 - G(t)$ and recall that $1 - F(t)$ is the probability that random variables are greater than t. Hence, Z first-order stochastically dominates Y means that, for each possible realisation of the random variables $t \in [a, b]$, the probability that Z has a realisation greater than t is larger than that of Y. (ii) also implies that the F lies everywhere (weakly) below that of G on the support. For (iii), note that the identity function is (weakly) increasing so that FOSD implies that the mean of X is greater than that of Y. The third characterisation can be generalised to deal with random vectors.

13.2 Second-order stochastic dominance

We now introduce a partial order among random variables with a common support and a common mean. We first define the idea of a mean-preserving spread. Let F and G be two CDFs with a common support and suppose $Z \sim F$ and $Y \sim G$.

Proposition 112. *Let F and G be two CDFs with a common support $[a, b]$ and a common mean. The following are equivalent.*

(i) G is a mean-preserving spread *of F, denoted $F \geq_{MPS} G$.*

(ii) $\mathbb{E}[Z|Y] = Y$, *where $Z \sim F$ and $Y \sim G$.*

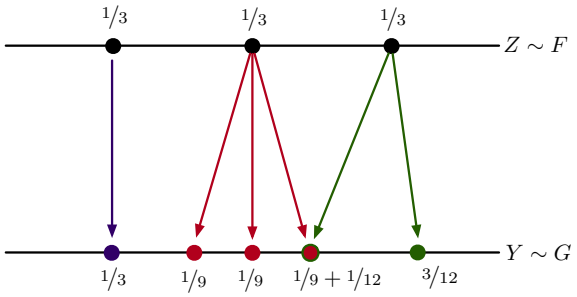
(iii) Y is a garbling of Z; i.e., there exists a random variable W with $\mathbb{E}[W|Z] = 0$ such that $Y \overset{\mathrm{d}}{=} Z + W$.[40]

(iii) implies that if $F \geq_{MPS} G$, then we can obtain $Z \sim F$ from $Y \sim G$ in the following way: (i) draw z from Z; (ii) draw w from W with $\mathbb{E}[W] = 0$; and (iii) $z \coloneqq y + w$. The figure below shows an example of a mean-preserving spread of a random variable Z, which has equal probability mass on three equidistant points. In this example, the distribution of W depends on the realisation of Z. If the left-most point is realised, then $w = 0$ so that probability mass of $Y = Z + W$ lies in the same location as Z (purple). In contrast, if the middle point is realised, then W 'splits' this point into three equidistance points with equal masses (red). Finally, if the right-most point is realised, W splits the mass into two with weight 1/4 on the left and weight 3/4 on the right. The distances are such that the mean is preserved.

[39] Given $Z \sim F$ and $Y \sim G$, we also often say that Z first-order stochastically dominates Y if $F \geq_{FOSD} G$.

[40] $\overset{\mathrm{d}}{=}$ means that the two sides have the same associated CDFs.

Figure 13.1: Mean-preserving spread.



Proposition 113. *Let F and G be two CDFs with a common mean. The following are equivalent.*

(i) F second-order stochastically dominates G; i.e., $F \geq_{SOSD} G$.[41]

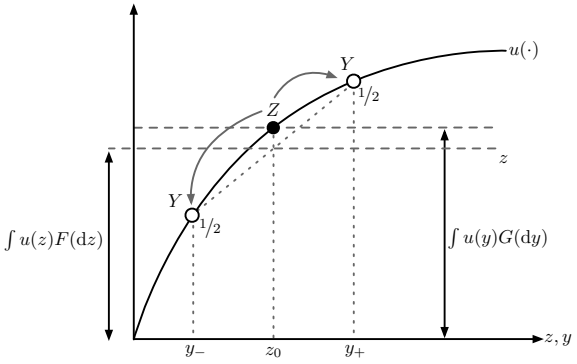
(ii) $\int_a^c F(t)\,\mathrm{d}t \leq \int_a^c G(t)\,\mathrm{d}t \;\forall c \in [a,b]$ *with equality at* $c = b$.

(iii) G *is a mean preserving spread of* F*, i.e.,* $F \geq_{MPS} G$.

(iv) $\mathbb{E}[u(Z)] \leq \mathbb{E}[u(Y)]$ *for any convex function* $u : [a,b] \to \mathbb{R}$*, where* $Z \sim F$ *and* $Y \sim G$.

To understand the connection between (iii) and (iv), consider the following figure in which $Z \sim F$ has a degenerate distribution and $Y \sim G$ is a mean-preserving spread of Z, i.e. $F \geq_{MPS} G$, where G gives probability mass $1/2$ to a low value and a mass $1/2$ to a high value (the two points are equidistance from Z so as to preserve the mean). Let us now consider integrating as per the definition of SOSD. Under F, $\mathbb{E}[u(Z)] = \int u(z)F(\mathrm{d}z)$ corresponds to the vertical height at $\overline{z}$. Under Y, $\mathbb{E}[u(Y)] = \int u(y)G(\mathrm{d}y)$ corresponds to the average of the heights at $\overline{y}_-$ and $\overline{y}_+$. We can see from this figure that the latter is less than the former; i.e. $F \geq_{SOSD} G$. Thinking of Z and Y as lotteries, this also tells us that a risk-averse expected-utility agent prefers lottery Z over lottery Y.

Figure 13.2: $F \geq_{MPS} G$ with $Z \sim F$ and $Y \sim G$.



To understand how (ii), let us consider the case in which the random variable X has support $[0,1]$ with mean $1/2$. Consider first the following cases:

(i) $\underline{X}$ has a degenerate distribution (i.e. $\underline{X}$ is a constant equal to $1/2$);

(ii) $\widetilde{X}$ equals 1 and 0 with equal probabilities;

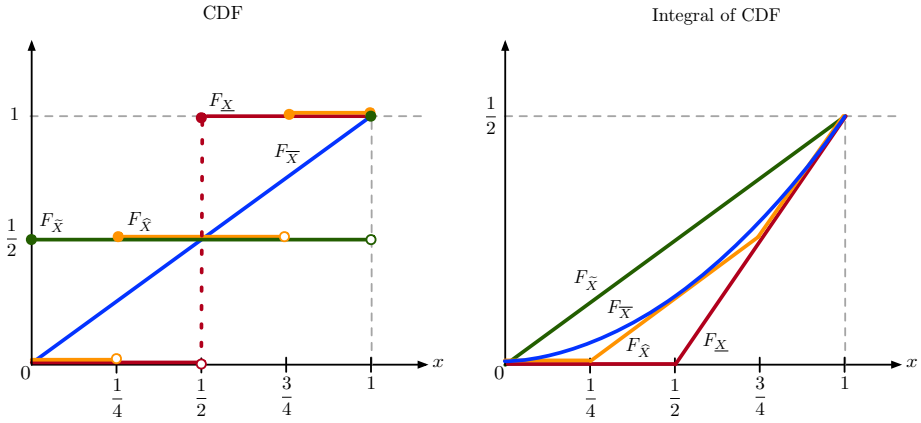
(iii) $\overline{X} \sim$ Uniform$[0,1]$;

[41] Given $Z \sim F$ and $Y \sim G$, we also often say that Z second-order stochastically dominates Y if $F \geq_{FOSD} G$.

(iv) $\widehat{X}$ equals 1/4 and 3/4 with equal probabilities;

We plot the CDF and and the integral of the CDF for each case below.

Figure 13.3: SOSD.



From the figure, we can deduce that

$$F_{\underline{X}} \geq_{SOSD} F_{\widehat{X}} \geq_{SOSD} F_{\overline{X}} \geq_{SOSD} F_{\widetilde{X}},$$

where we may replace $\geq_{SOSD}$ with $\geq_{MPS}$; i.e., $\widetilde{X}$ is a mean-preserving spread of $\overline{X}$, which, in turn, is a mean-preserving spread of $\widehat{X}$, which, in turn, is a mean-preserving spread of $\underline{X}$.

13.2.1 Hazard rate

Let T be a random variable that represents the time until a particular event (e.g., T is the length of unemployment period and the event is finding a job). Let F denote its associated cumulative distribution function so that $F(t) = \Pr(T \leq t)$ denotes the probability that the event has occurred by time t. The survivor function is the probability of surviving past t and define

$$S(t) \coloneqq 1 - F(t) = \Pr(T > t).$$

The conditional probability of leaving the initial state within the time interval t until $t + h$, given survival up to time t is given by

$$\Pr(t \leq T < t + h | T \geq t).$$

Dividing above by h gives the average probability of leaving per unit time period over the interval t until $t + h$. Taking the limit as $h \searrow 0$ gives the *hazard rate function*, denoted $\lambda(t)$; i.e.,

$$\lambda(t) \coloneqq \lim_{h \searrow 0} \frac{\Pr(t \leq T < t + h | T \geq t)}{h}.$$

At each time t, the hazard rate function is the instantaneous rate of event occurring per unit of time. Observe that

$$\Pr(t \leq T < t + h | T \geq t) = \frac{\Pr(\{t \leq T < t + h\} \cap \{t \leq T\})}{\Pr(t \leq T)} = \frac{\Pr(t \leq T < t + h)}{\Pr(t \leq T)} = \frac{F(t + h) - F(t)}{1 - F(t)}.$$

Recalling that F is right-continuous,

$$\lambda(t) = \lim_{h \searrow 0} \frac{\Pr(t \leq T < t+h | T \geq t)}{h} = \lim_{h \searrow 0} \frac{1}{1-F(t)} \frac{F(t+h) - F(t)}{h} = \frac{f(t)}{S(t)}.$$

There exists a bijection between the hazard rate function and the CDF of T. To see this, observe that

$$\lambda(t) = \frac{f(t)}{1-F(t)} = -\frac{\mathrm{d} \ln(1-F(t))}{\mathrm{d}t}.$$

Integrating both sides over $[0, z]$ gives

$$\begin{aligned} \int_0^z \lambda(t)\,\mathrm{d}t &= -\int_0^z \frac{\mathrm{d}\ln(1-F(t))}{\mathrm{d}t}\mathrm{d}t = -\ln(1-F(z)) + \ln(1-F(0)) \\ &= -\ln(1-F(z)), \end{aligned}$$

where we used that $F(0) = 0$. Therefore, it follows that

$$F(z) = 1 - \exp\left(-\int_0^z \lambda(t)\,\mathrm{d}t\right). \tag{13.1}$$

Observe that if we choose the hazard rate function to be constant so that $\lambda(t) \coloneqq \lambda$, then

$$F(z) = 1 - \exp(-\lambda z),$$

which is the CDF for exponential distribution.

13.3 Domination in terms of hazard rate and reverse hazard rates

Given two CDFs F and G, we say that F *dominates* G *in terms of the hazard rate*, denoted $F \geq_{HR} G$, if

$$\lambda_F(z) = \frac{f(z)}{1-F(z)} \leq \frac{g(z)}{1-G(z)} = \lambda_G(z) \;\; \forall z \in \mathbb{R}.$$

Similarly, we say that F *dominates* G *in terms of the reverse hazard rate*, denoted $F \geq_{RHR} G$, if

$$\sigma_F(z) \coloneqq \frac{f(z)}{F(z)} \geq \frac{g(z)}{G(z)} =: \sigma_G(z) \;\; \forall z \in \mathbb{R}.$$

Note that $(\mathbf{F}, \geq_{HR})$ and $(\mathbf{F}, \geq_{RHR})$ are both partially ordered sets.

Domination in terms of hazard rate and reverse hazard rate both imply first-order stochastic dominance.

Proposition 114. *Suppose that $F \geq_{HR} G$ or $F \geq_{RHR} G$ for some CDFs F and G. Then, $F \geq_{FOSD} G$.*

Proof. Recall 13.1. Then, by monotonicity of integral, it follows that if $F \geq_{HR} G$, then

$$F(z) = 1 - \exp\left(-\int_0^z \lambda_F(t)\,\mathrm{d}t\right) \leq 1 - \exp\left(-\int^z \lambda_G(t)\,\mathrm{d}t\right) = G(z) \;\; \forall z \in \mathbb{R}.$$

Hence, $F \geq_{FOSD} G$.

Observe that

$$\sigma_F(z) = \frac{\mathrm{d}\ln F(z)}{\mathrm{d}t} \Rightarrow \int_z^\infty \sigma_F(t)\,\mathrm{d}t = -\ln F(z) \Leftrightarrow F(z) = \exp\left[-\int_z^\infty \sigma_F(t)\,\mathrm{d}t\right],$$

where we used the fact that $\lim_{z\to\infty} F(z) = 1$. Hence, if $F \geq_{RHR} G$, then

$$F(z) = \exp\left[-\int_z^\infty \sigma_F(t)\,\mathrm{d}t\right] \leq \exp\left[-\int_z^\infty \sigma_G(t)\,\mathrm{d}t\right] = G(z) \;\; \forall z \in \mathbb{R}.$$

Hence, $F \geq_{FOSD} G$. ■

13.4 Relation with domination in terms of the likelihood ratio

Given two CDFs F and G, we say that F dominates G in terms of the likelihood ratio, denoted $F \geq_{LR} G$ if

$$\frac{f(z)}{g(z)} \leq \frac{f(y)}{g(y)} \; \forall z, y \in \mathbb{R} : z < y.$$

Equivalently, $F \geq_{LR} G$ is f/g is nondecreasing. If $F \geq_{LR} G$, then the graph of densities f and g can intersect at most once. Note that $(\mathbf{F}, \geq_{LR})$ is a partially ordered set.

We will show that domination in terms of the the likelihood ratio implies domination in terms of both hazard rate and reverse hazard rate. Combined with Proposition 114, it follows that dominance in terms of the likelihood ratio implies first-order stochastic dominance.

Proposition 115. *Suppose that $F \geq_{LR} G$ for some CDFs F and G. Then, $F \geq_{HR} G$ and $F \geq R_{RHR}G$.*

Proof. Suppose $F \geq_{LR} G$. Then for all $z < y$,

$$\frac{f(z)}{g(z)} \leq \frac{f(y)}{g(y)} \Leftrightarrow \frac{f(y)}{f(z)} \geq \frac{g(y)}{g(z)}.$$

Hence, by monotonicity of integration,

$$\frac{1}{\lambda_F(z)} = \frac{1-F(z)}{f(z)} = \int_z^\infty \frac{f(y)}{f(z)}\mathrm{d}y \geq \int_z^\infty \frac{g(y)}{g(z)}\mathrm{d}y = \frac{1-G(z)}{g(z)} = \frac{1}{\lambda_G(z)}.$$

It follows that $\lambda_F(z) \leq \lambda_G(z)$ for all $z \in \mathbb{R}$; i.e., $F \geq_{HR} G$.

We also have that, for all $z < y$,

$$\frac{f(z)}{g(z)} \leq \frac{f(y)}{g(y)} \Leftrightarrow \frac{f(z)}{f(y)} \leq \frac{g(z)}{g(y)}.$$

Hence, by monotonicity of integration,

$$\frac{1}{\sigma_F(y)} = \frac{F(y)}{f(y)} = \int_0^y \frac{f(z)}{f(y)}\mathrm{d}z \geq \int_0^y \frac{g(z)}{g(y)}\mathrm{d}z = \frac{G(y)}{g(y)} = \frac{1}{\sigma_G(y)}.$$

It follows that $\sigma_F(y) \geq \sigma_G(y)$ for all $y \in \mathbb{R}$; i.e., $F \geq_{RHR} G$. ■

To recap, we have shown that

$$\begin{aligned} F \geq_{LR} G &\Rightarrow F \geq_{HR} G \text{ and } F \geq_{RHR} G, \\ F \geq_{HR} G &\Rightarrow F \geq_{FOSD} G, \\ F \geq_{RHR} G &\Rightarrow F \geq_{FOSD} G. \end{aligned}$$

13.5 Order statistics

Let $(Z_i)_{i=1}^n$ be a collection of n independent random variables with associated CDF and density given by F and f, respectively. Let $(Z_{(k)})_{k=1}^n$ be a rearrangement so that $Z_{(1)} \geq Z_{(2)} \geq \cdots \geq Z_{(n)}$.

Each random variable $Z_{(k)}$ is referred to as the *kth order statistics*. Let $F_{(k)}$ denote the distribution of $Z_{(k)}$ with corresponding density function $f_{(k)}$. Our task is to express $f_{(k)}$ as a function of f.

13.5.1 Highest-order statistic

We derive the expression for $F_{(1)}$ and $f_{(1)}$. The event $Z_{(1)} \leq z$ is the same as the event that $Z_i \leq z$ for all $i \in \{1, \ldots, n\}$. Since each Z_i is an independent draw from F, we have

$$F_{(1)}(z) = (F(z))^n \equiv F^n(z).$$

Hence, using the chain rule, the associated density is

$$f_{(1)}(z) = nF^{n-1}(z) f(z).$$

The following is then immediate.

Proposition 116. *If $F \geq_{FOSD} G$, then $F_{(1)} \geq_{FOSD} G_{(1)}$.*

13.5.2 Second-order statistic

We now derive the expression for $F_{(2)}$ and $f_{(2)}$. The event that $Z_{(2)} \leq z$ is the union of the following disjoint events: (i) $Z_i \leq z$ for all $i \in \{1, \ldots, n\}$; (ii) $n-1$ of Z_i's are less than or equal to z and exactly one is greater than y. There are n different ways in which (ii) can occur. Hence,

$$\begin{aligned} F_{(2)}(z) &= F^n(z) + nF^{n-1}(z)(1 - F(z)) \\ &= nF^{n-1}(z) - (n-1)F^n(z). \end{aligned}$$

The associated density function is

$$f_{(2)}(z) = n(n-1)(1 - F(z))F^{n-2}(z) f(z).$$

Proposition 117. *If $F \geq_{FOSD} G$, then $F_{(2)} \geq_{FOSD} G_{(2)}$.*

13.5.3 kth-order statistic

In general, we have the following for the density of the kth order statistic:

$$f_{(k)}(z) = \frac{n!}{(k-1)!(n-k)!} \left(F^{n-k}(z)\right)(1 - F(z))^{k-1} f(z),$$

where the $n!/(k-1)!(n-k!)$ is the number of different ways in which any draws of $\{Z_i\}_{i=1}^n$ could have been the kth order statistic. $F^{n-k}(z)$ is the probability that $n-k$ draws are below z, $1 - F^{k-1}(z)$ is the probability that $k-1$ draws are above z and $f(z)$ is the density associated with drawing z.

Remark 27. The tradition in statistic is to worry about the smallest order statistics; i.e., $Z^{(k)}$ means the kth smallest among $(Z_i)_{i=1}^n$. The density of the kth order statistic is defined as

$$f^{(k)}(z) \coloneqq \frac{n!}{(k-1)!(n-k)!} (F(z))^k (1 - F(z))^{n-k} f(z).$$