Math Camp for Incoming Doctoral Students 2022

Department of Economics

Stanford University

Instructor: Mitchell Watt

Note to the reader

Please contact me (at mwatt@stanford.edu*) if you notice any errors or have comments or suggestions. These notes are based partially on older versions of the notes written by Adem Dugalic, Michael Pollmann, Laurence Wong, Pete Troyan, Clayton Featherstone, Joe Romano and others. I have also borrowed heavily from the texts and references listed on page 3. Thank you to Zi Yang Kang for these LaTeX style files.*

Out of consideration for the environment, please consider **not** printing these notes.

Contents

Foreword

Welcome to math camp, the Department of Economics and Stanford University!

Our goal in math camp is to familiarize you with the mathematical concepts and notation that you will encounter in your first year at Stanford, and presumably thereafter in your economics career. A second (equally important) goal is to give you an opportunity to get to know Stanford and the people you will be spending a lot of time with and collaborating with over the coming years. These notes are related to the former objective. I will leave the latter objective to your social chairs!

In response to feedback, we have made some small changes to math camp this year. These changes mostly consist of reducing the coverage of material that is covered in detail in later courses, while increasing the coverage of material that continues to challenge students in core courses.

Designing a syllabus for a math camp is somewhat tricky. The main reasons are *heterogeneity* in:

- the mathematical backgrounds of PhD students in economics[1]—some entering students will have taken pure mathematics degrees in the past, while others will have taken just enough mathematics courses to find their way here;
- what students want to do with mathematics—some students are likely to go on to study microeconomic theory, econometric theory or conduct other very mathematically intensive research, while other students will conduct research using other methods and only want the bare minimum mathematics needed to get through; and
- the types of mathematics used for different areas of research—the vital mathematical tools used by microeconomists differs from those of macroeconomists and econometricians, even before thinking about heterogeneity within these fields.

Therefore, we are focusing on material that you will almost surely encounter in your first year of economics and in particular on the results and techniques you will need to use in problem sets

[1] Moreover, there are some non-economics PhD students taking the math camp - while these notes are geared at economics PhD students, I will keep this in mind while teaching.

and exams for your core courses. Many of these are also fundamental to research in certain areas of economics.

My standing assumption in these notes is that you have taken mathematics at the undergraduate level up to real analysis and linear algebra *but* that you may be very rusty or didn't understand the material well in the first place. That is, I will be treating the topics we cover as if you have seen them before but don't understand them well, and I will put a special emphasis on how the topics are likely to be useful to you in economics (which may be different from the way they are taught in real analysis classes for mathematicians).

A big change I am making in response to feedback is a reduction in the emphasis on proofs of technical results and increasing the emphasis on methods and problem-solving. We will still cover some proofs where they are necessary to have a deep understanding of the material or where they illustrate an interesting or generalizable proof technique. Where the proofs were already available in older versions of the notes, I have included them here—but we will not cover all of these in-class.

Given the focus on problem-solving, there are many problems included in these notes. I will ask you to work on some of these in class and others I hope you will look into outside of class. I will post solutions to the problems at the end of the math camp.

I hope these notes will serve as a reference for you as you embark on your core courses and your career in economics. Here are some additional references you might be interested in:

- For a textbook covering the basics with an emphasis on economic applications, Angel de La Fuente's *Mathematical Methods and Models for Economists* is a good option.

- For linear algebra, an excellent reference is *Linear Algebra Done Right* by Sheldon Axler. It is a good introduction to linear algebra from the perspective of linear transformations (downplaying the role of matrices and especially determinants).

- For analysis, the classic reference is Walter Rudin's *Principles of Mathematical Analysis*. An exceptional, but very mathematically sophisticated, guide to analysis is *Infinite-Dimensional Analysis: A Hitchhiker's Guide* by Charalambos Aliprantis and Kim Border. It is hard to imagine any mathematics of use in economics that is not contained in this book, although it is a very complicated guide.

- Some other good references geared to economics are the mathematics chapters in *Recursive Methods in Economic Dynamics* by Nancy Stokey, Robert Lucas and Edward Prescott, and the appendices of the excellent *Microeconomic Foundations I: Choice and Competitive Markets* by David Kreps.

- On convexity, there are several excellent textbooks by R. Tyrrell Rockafellar: *Convex Analysis* for a solid grounding and *Variational Analysis* (with Roger J-B Wets) is the authoritative guide. For a focus on optimization, *Convex Optimization* by Stephen Boyd and Lieven Vandenberghe is also excellent.

The major sources of material in this document are as follows:

- Previous versions of math camp notes written by Adem Dugalic, Michael Pollmann, Laurence Wong, Pete Troyan, Clayton Featherstone, Joe Romano and others,
- *Linear Algebra Done Right* by Sheldon Axler,
- *Mathematical Methods and Models for Economists* by Angel de la Fuente,
- *Real Analysis with Economic Applications* by Efe A. Ok,
- *Introduction to Modern Economic Growth* by Daron Acemoglu,
- *Recursive Methods in Economic Dynamics* by Nancy L. Stokey and Robert E. Lucas,
- *Linear and Nonlinear Programming* by David G. Luenberger and Yinyu Ye, and
- many Wikipedia articles!

Part I

Preliminaries

1

Logic and Sets

Contents

1.1 Logic

Logic is one of the foundations of mathematics. Think about it logically: without choosing a system of logic, how can we determine if some mathematical claim is true or false? You could spend your whole life on the study of mathematical logic, but in these notes we will cover only the very basic conventions of mathematical logic that are used regularly by economists.

The basic objects of logic are **propositions**.

Definition 1.1.1. A proposition is a logical statement that is either **true** or **false**.

We often use Roman capital letters P, Q, R to stand for propositions.

Example. The statement

$$P: \quad 3 < 5$$

is true.

The statement

$$Q: \quad 3 \geq 5$$

is false.

The statement

$$Rx: \quad x \geq 0$$

is true if x is larger or equal to 0, and false if x is less than 0. ♣

This latter example is an example of a **propositional variable**, whose truth depends on the value of x.

Definition 1.1.2. We define three common binary logical operators: $\wedge$ (and), $\vee$ (or), $\neg$ (not). Let

$$\begin{aligned} R:&\quad P\wedge Q &&\text{(P and Q)}\\ S:&\quad P\vee Q &&\text{(P or Q)}\\ T:&\quad \neg P &&\text{(not P),}\end{aligned}$$

then

- $R = P \wedge Q$ is true if *both* P and Q are true; but false if either P or Q is false;
- $S = P \vee Q$ is true if either P is true, Q is true, or both P and Q are true; while it is false if both P and Q are false;
- $T = \neg P$ is true if P is false, and false if P is true.

If you have trouble remembering which of $\wedge$ and $\vee$ is “and” and which is “or” it might help you to remember that $\vee$ stands for the (first letter of) Latin “vel” – which means “or.” If your Latin is a little rusty, you may think of $\wedge$ as a capital “A” (with the horizontal bar missing), short for “and.”

Definition 1.1.3. We say P implies Q, or $P \implies Q$, if Q is always true when P is true.

Note that this definition does not put any restrictions on Q if P is false. In this case, P is said to be a **sufficient condition** for Q, while Q is said to be a **necessary condition** for P.

Definition 1.1.4. We say P if and only if Q, or P iff Q (note the two “f”), or P and Q are equivalent, or P is equivalent to Q, or $P \iff Q$, if $P \implies Q$ and $Q \implies P$.

In this case P is said to be a **necessary and sufficient condition** for Q (and vice versa).

The following truth table summarizes these definitions:

P	Q	$P \wedge Q$	$P \vee Q$	$\neg P$	$P \implies Q$	$P \iff Q$
true	true	true	true	false	true	true
true	false	false	true	false	false	false
false	true	false	true	true	true	false
false	false	false	false	true	true	true

Sometimes, the order of precedence is relevant. Note that typically

$$\neg P \wedge Q$$

is the same as

$$(\neg P) \wedge Q$$

and NOT the same as

$$\neg (P \wedge Q)$$

so that $\neq$ takes highest precedence. Beyond that, in my opinion, when there is room for doubt, it is typically better to use parentheses to clarify the order of operations rather than rely on a mutual understanding.

Based on these definitions, we could show any number of (non-) relations, e.g.

$$\begin{aligned}
\neg (P \wedge Q) &\iff \neg P \vee \neg Q \\
\neg (P \vee Q) &\iff \neg P \wedge \neg Q \\
(P \wedge Q) \wedge R &\iff P \wedge (Q \wedge R) \\
(P \vee Q) \vee R &\iff P \vee (Q \vee R) \\
P \wedge Q &\implies P \vee Q
\end{aligned}$$

Note that to show the equivalence in, e.g., the first statement, we need to show that both

$$\begin{gathered}
\neg (P \wedge Q) \implies \neg P \vee \neg Q \\
\text{and} \\
\neg P \vee \neg Q \implies \neg (P \wedge Q)
\end{gathered}$$

to satisfy the definition of $\iff$.

Exercise 1.1. *Prove the following logical formulas. (That is, show that the formulas always evaluate to 'true', regardless of the truth values of P and Q).*

$$\begin{aligned}
((P \implies Q) \wedge P) &\implies Q \\
((P \implies Q) \wedge \neg Q) &\implies \neg P \\
(\neg Q \implies \neg P) &\iff (P \implies Q)
\end{aligned}$$

The first is called "direct proof" or modus ponens, the second is called modus tollens and is the basis of proofs by contradiction, and the third is the basis of argument by contraposition.

1.2 Sets

The first objects of our analysis are *sets.* A set should be defined precisely in terms of the Zermelo-Fraenkel axioms, which are a list of propositions which mathematicians typically take to be true (e.g., there exists an empty set). We will not go into the murky logical foundations of set theory. Instead, let's adopt the following informal definition.

Definition 1.2.1. A **set** is a collection of objects.

In most versions of formal set theory, these objects are also sets! Even numbers are viewed as sets, as we will see later on.

For any set A, $x \in A$ means that the element x is in the set A. The empty set is denoted by $\emptyset$. It is the unique set with no elements (as its name suggests).

Definition 1.2.2. Let A and B be sets.

(a) If each element of A is also an element of B, then we say that A is a **subset** of B, or write $A \subseteq B$. Equivalently, we may say that B is a **superset** of A, denoted by $B \supseteq A$.

(b) If $A \subseteq B$ and $B \subseteq A$, we say that the sets are **equal**, denoted by $A = B$.

(c) We say that A is a **proper subset** of B, or $A \subset B$, if $A \subseteq B$ and $A \neq B$.

Given two sets, we have various ways of combining them to create new sets.

Definition 1.2.3. Let A and B be sets. We define:

(a) **set union** by $A \cup B = \{x : x \in A \text{ or } x \in B\}$;

(b) **set intersection** by $A \cap B = \{x : x \in A \text{ and } x \in B\}$. If $A \cap B = \emptyset$, then A and B are said to be **disjoint**;

(c) **set minus**, by $A \setminus B = A - B = \{x : x \in A \text{ and } x \notin B\}$;

(d) the **Cartesian product** $A \times B = \{(a, b) : a \in A \text{ and } b \in B\}$;[a] and

(e) the **power set** of A, denoted by $\mathcal{P}(A)$, or 2^A, as the set of all subsets of A, .

[a] Formally, we should first define ordered pairs by $(a, b) = \{a, \{a, b\}\}$, but this is really pedantic.

Often, we are interested in sets that are subsets of some *universal set* $\mathcal{U}$, which might be the set of all real numbers, the set of all functions or the set of all three-legged chairs. With this universe fixed, we may define the **complement** of a set as $A^c = \mathcal{U} - A$, all the objects not in A.

Here are some basic set properties :

Theorem 1.2.4. *Distributive laws:*

(i) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

(ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Theorem 1.2.5. *De Morgan's laws:*

(i) $(A \cup B)^c = A^c \cap B^c$.

(ii) $(A \cap B)^c = A^c \cup B^c$.

Definition 1.2.6. Index families: Suppose for each i in a nonempty set I there corresponds a set A_i. Then we call $\mathcal{A} = \{A_i : i \in I\}$ an index family.

We can equivalently define set operations over index families: $\bigcup_{i \in I} A_i$, $\bigcap_{i \in I} A_i$. De Morgan's law now gives $(\bigcup_{i \in I} A_i)^c = \bigcap_{i \in I} A_i^c$.

Although we agreed to avoid the axiomatic construction of sets, there is one famous axiom related to index families that we should at least mention.

Definition 1.2.7 (Axiom of Choice)**.** If $\{A_i : i \in I\}$ is a nonempty set of nonempty sets, then the Cartesian product $\prod_{i \in I} A_i$ is nonempty.

This axiom is somewhat controversial in mathematics: it is independent of the other more innocuous axioms of set theory (i.e., it cannot be proven from them) and has a few unusual implications (like the Banach-Tarski paradox). On the other hand, dropping the Axiom of Choice brings its own problems, including making infinite-dimensional analysis much harder. When problems in economics are modelled in infinite-dimensional space, theoretical issues related to implications of the Axiom of Choice occasionally arise.

With sets defined, we may now also define a couple more logical operations, called **quantifiers**.

Definition 1.2.8. Let Px be a propositional variable and S a set. The **existential quantifier** is the statement $\exists x \in S, Px$ which is true if **there exists** an $x \in S$ for which the proposition Px is true. The **universal quantifier** $\forall x \in S, Px$ is true if **for all** $x \in S$, the proposition Px is true.

For example, let Px be the statement x is a prime number and let S be the set of even numbers. Then the statement $\exists x \in S, Px$ is true since 2 is an even number. The statement $\forall x \in S, Px$ is false since 4 is an even number which is not prime.

Exercise 1.2. *Prove that*

(a) $(A \cup B) \times C = (A \times C) \cup (B \times C)$*, and*

(b) $(A \cap B) \times (C \cap D) = (A \times C) \cap (B \times D)$*.*

1.3 Relations

Definition 1.3.1. A **relation** R between X and Y is a subset of $X \times Y$. If $(x, y) \in R$, we write xRy. If $X = Y$, then the relation is said to be **on** X.

Example. Let X be the set of alternatives from which an individual can choose. A preference relation $\succsim$ is a relation on X. We read $x \succsim y$ as 'x is at least as good as y' or 'x is weakly preferred to y'. We define two other relations as follows:

(a) Strict preference relation: $x \succ y$ if and only if $x \succsim y$ and not $y \succsim x$.

(b) Indifference relation: $x \sim y$ if and only if $x \succsim y$ and $y \succsim x$.

♣

Definition 1.3.2. A relation R on a set X is

(a) **reflexive** if xRx for all $x \in X$;

(b) **irreflexive** if $\neg xRx$ for all $x \in X$;

(c) **symmetric** if xRy implies yRx for all $x, y \in X$;

(d) **antisymmetric** if xRy and yRx implies $x = y$ for all $x, y \in X$

(e) **asymmetric** if xRy implies not yRx for all $x, y \in X$

(f) **transitive** if xRy and yRz imply xRz for all $x, y, z \in X$;

(g) **complete** if for all $x, y \in X$, we have xRy or yRx.

A reflexive, symmetric and transitive relation is an **equivalence relation** (e.g., set equality).

A reflexive, antisymmetric and transitive relation is a **partial order** (e.g., set inclusion $\subseteq$). If it is also complete, it is a **total order** (e.g., $\geq$ in $\mathbb{N}$).

A reflexive, asymmetric and transitive relation is a **strict partial order** (e.g. strict inclusion $\subset$)

Definition 1.3.3. Let R be an equivalence relation on a set X. Then the **equivalence class** of $x \in X$ is given by

$$[x]_R = \{y \in X : yRx\}.$$

Write X/R for the set of equivalence classes of R in X. It is sometimes called the **quotient** of X by R.

Definition 1.3.4. A **partition** of a set X is a collection $\mathcal{P}$ of nonempty subsets of X such that

(i) for all $x \in X$, there exists $A \in \mathcal{P}$ such that $x \in A$.

(ii) for all $A, B \in \mathcal{P}$, $A \cap B \neq \emptyset$ implies $A = B$.

Theorem 1.3.5. *Let R be an equivalence relation on a set X. Then $\{[x]_R : x \in X\}$ is a partition of X.*

Proof. Since R is reflexive, each element of X is in some equivalence class. Suppose $[x]_R \cap [y]_R \neq \emptyset$, so there exists an element $z \in [x]_R \cap [y]_R$. We need to show that $[x]_R = [y]_R$. Let $x' \in [x]_R$. This means $x'Rx$. Since $z \in [x]_R \cap [y]_R$, we can write zRx and zRy, or by symmetry, xRz and zRy. Using the transitivity twice, we obtain $x'Ry$, so $x' \in [y]_R$. This shows that $[x]_R \subseteq [y]_R$. In a similar manner, we can show that $[y]_R \subseteq [x]_R$, so $[x]_R = [y]_R$. Hence, $\{[x]_R : x \in X\}$ is a partition of X. □

Example. The preference relation $\succsim$ is said to be **rational** if it is complete and transitive. Note that completeness implies reflexiveness. We can then show that

(a) $\succ$ is irreflexive and transitive.

(b) $\sim$ is reflexive, symmetric, and transitive.

Hence, the indifference relation $\sim$ is an equivalence relation. The equivalence classes $[x]_\sim = \{y \in X : y \sim x\}$ are often called indifference curves or indifference sets. By Theorem 6.1.6, we see that indifference sets partition the set X, and that indifference sets do not intersect. ♣

Exercise 1.3 (ECON 202 - Final Exam 2015). *A long-standing finding in psychology is that individuals cannot distinguish small quantity differences and that the probability that a difference is distinguishable depends on the ratio of the two quantities. Suppose there is some $\delta > 1$ such that, given two quantities $x > y > 0$ of a good, the two items can be distinguished if and only if $x > \delta y$.*

Define a preference relation for a single good as follows:

- $x \succ y$ *if* $x > \delta y$,
- $x \succeq y$ *if it is not the case that* $y \succ x$, *and*
- $x \sim y$ *if* $x \succeq y$ *and* $y \succeq x$.

Prove or disprove each of the following:

(a) *The relation* $\succeq$ *is complete.*

(b) *The relation* $\succ$ *is transitive.*

(c) *The relation* $\sim$ *is transitive.*

(d) $(x \succ y \text{ and } y \sim z) \implies x \succeq z$.

Exercise 1.4. *Determine which of the following relations are equivalence relations and describe their equivalence classes:*

(a) xRy *if* $x - y$ *is divisible by 8,*

(b) xRy *if* $x - y$ *is odd,*

(c) xRy *if* $x - y$ *is prime.*

1.4 Functions

Definition 1.4.1. Let X and Y be sets. A **function** between X and Y is a nonempty relation $f \subseteq X \times Y$ such that if $(x, y) \in f$ and $(x, y') \in f$, then $y = y'$. The **domain** and **range** of f

are given by

$$\text{domain } f = \{x \in X : \text{there exists } y \in Y \text{ such that } (x, y) \in f\},$$
$$\text{range } f = \{y \in Y : \text{there exists } x \in X \text{ such that } (x, y) \in f\}.$$

The set Y is referred to as the **codomain** of f. If the domain of f is X, we write $f : X \longrightarrow Y$ and say that f is a function from X to Y, or that f is a mapping from X into Y. If $(x, y) \in f$, we often denote y by $f(x)$.

Definition 1.4.2. Let $f : X \longrightarrow Y$. If $C \subseteq X$, then the **image** of C under f, denoted $f(C)$, is the set given by

$$f(C) = \{f(x) \in Y : x \in C\}.$$

In this notation, $f(X) \subseteq Y$ is the range of f. If $D \subseteq Y$, then the **inverse image** of D under f, denoted $f^{-1}(D)$, is

$$f^{-1}(D) = \{x \in X : f(x) \in D\}.$$

Definition 1.4.3. A function $f : X \longrightarrow Y$ is

(a) **surjective** (or is said to map X **onto** Y) if the range of f is Y.

(b) **injective** (or **one-to-one**) if for all $x, x' \in X$, $f(x) = f(x')$ implies $x = x'$.

(c) **bijective** (or is a **one-to-one correspondence** between X and Y) if it is both surjective and injective.

Theorem 1.4.4. *Let f be a function that maps X into Y. Then we have*

(a) If $D \subseteq Y$, then $f\big(f^{-1}(D)\big) \subseteq D$;

(b) If f maps X onto Y, then $f\big(f^{-1}(D)\big) = D$.

(c) If $C \subseteq X$, then $C \subseteq f^{-1}\big(f(C)\big)$;

(d) If f is one-to-one, then $C = f^{-1}\big(f(C)\big)$.

(e) If $\{C_\alpha : \alpha \in A\}$ is a family of subsets of X, then $f\left(\bigcup_\alpha C_\alpha\right) = \bigcup_\alpha f(C_\alpha)$;

(f) If $\{D_\alpha : \alpha \in A\}$ is a family of subsets of Y, then $f^{-1}\left(\bigcup_\alpha D_\alpha\right) = \bigcup_\alpha f^{-1}(D_\alpha)$;

(g) $f^{-1}\left(\bigcap_\alpha D_\alpha\right) = \bigcap_\alpha f^{-1}(D_\alpha)$;

(h) $f^{-1}(D^c) = \big(f^{-1}(D)\big)^c$.

Proof. We only prove (a), (b), and (f); the others are similar. If $y \in f\big(f^{-1}(D)\big)$, then $y = f(x)$ for some $x \in f^{-1}(D)$. This means that $y = f(x)$ and $f(x) \in D$, so $y \in D$, which proves (a). If $y \in D$, then $y = f(x)$ for some $x \in X$, and therefore for some $x \in f^{-1}(D)$. This means $y \in f\big(f^{-1}(D)\big)$, which proves (b). Finally, suppose $x \in f^{-1}\left(\bigcup_\alpha D_\alpha\right)$. Then $f(x) \in D_\alpha$ for some α, so $x \in f^{-1}(D_\alpha)$ for some α. Thus $x \in \bigcup_\alpha f^{-1}(D_\alpha)$. Every step in this argument is reversible, establishing (f). □

Definition 1.4.5. Let $f : X \to Y$ and $g : Y \to Z$. The **composite** of f and g, denoted $g \circ f$, is given by

$$g \circ f = \{(x, z) \in X \times Z : \text{there exists } y \in Y \text{ such that } f(x) = y \text{ and } g(y) = z\}.$$

That is, $(g \circ f)(x) = g(f(x))$.

Theorem 1.4.6. *Let $f : X \to Y$ and $g : Y \to Z$. We have:*

(a) If f and g are surjective, then $g \circ f$ is surjective.

(b) If f and g are injective, then $g \circ f$ is injective.

(c) If f and g are bijective, then $g \circ f$ is bijective.

Proof. (a) Since g is surjective, for every $z \in Z$, there exists $y \in Y$ such that $g(y) = z$. Furthermore, since f is also surjective, there exists $x \in X$ such that $f(x) = y$. But $(g \circ f)(x) = g(f(x)) = g(y) = z$. Thus, $g \circ f$ is surjective.

(b) Suppose $(g \circ f)(x) = (g \circ f)(x')$. This means $g(f(x)) = g(f(x'))$. Since g is injective, $f(x) = f(x')$ for all $f(x), f(x') \in Y$. But f is also injective, so $x = x'$ for all $x, x' \in X$. Thus, $g \circ f$ is injective.

(c) Follows directly from parts (a) and (b).

□

Definition 1.4.7. Let $f : X \to Y$ be bijective. Then the **inverse function** of f is the function f^{-1} given by:

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

Theorem 1.4.8. *Let $f : X \longrightarrow Y$ be bijective. Then $f^{-1} : Y \longrightarrow X$ is also bijective.*

Proof. Suppose $f^{-1}(y) = f^{-1}(y')$. Since f is a function, $f(f^{-1}(y)) = f(f^{-1}(y'))$. But this means $y = y'$, so f^{-1} is injective. To show that f^{-1} is surjective, take any $x \in X$. Let $y = f(x)$, then $x \in f^{-1}(y)$. Since f^{-1} is a function (because f is bijective), $x = f^{-1}(y)$. Therefore, f^{-1} is surjective. Thus, f^{-1} is bijective. □

Exercise 1.5. *Decide whether the functions are injective, surjective or bijective. If the function is bijective, then find the inverse.*

(a) $f : \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0}, f(x) = |x|$

(b) $g : \mathbb{N} \longrightarrow \mathbb{N}, g(n) = n + 1$

(c) $h : \mathbb{N}_0 \longrightarrow \mathbb{Z}, h(n) = \begin{cases} \frac{n}{2}, & \textit{for n even} \\ -\frac{n+1}{2}, & \textit{for n odd} \end{cases}$

2

Numbers

Contents

2.1 Constructing the Real Numbers

So far, we have done math without defining numbers! That's a big hole, let's fill it now.

We take the approach summarized by Leopold Kronecker: "God made the integers, all the rest is the work of man." Except we will go back one step and start with the **natural numbers**, $\mathbb{N}$.

There are many equivalent ways to define the natural numbers. The first approach is axiomatic and due to Peano.

Definition 2.1.1. The **Peano axioms** of $\mathbb{N}$ are

(a) 0 is a natural number.

(b) Every natural number has a successor which is also a natural number.

(c) 0 is not the successor of any natural number.

(d) If the successor of x equals the successor of y then $x = y$ (the successor function is injective).

(e) Induction: if a statement is true of 0 and if the truth of that statement for a number implies its truth for that number's successor, then the statement is true for every natural number.

An alternative (arguably best) approach to defining $\mathbb{N}$ is set-theoretic and due to von Neumann (you will hear more about him in ECON 203!), who defines $0 := \{\}$, the empty set, and $n := n - 1 \cup \{n - 1\}$. This is appealing mathematically, but isn't of much interest economically.

The natural numbers are a "commutative semiring": they are closed under addition and multiplication, and satisfy the usual commutative/associative/distributive properties with identities 0 and 1. They also satisfy the following "well-ordering" property with respect to the total order $\geq$ on $\mathbb{N}$.

Theorem 2.1.2. *(**Well-ordering Principle**) The set $\mathbb{N}$ of natural numbers is well-ordered, that is, if T is a non-empty subset of $\mathbb{N}$, then T contains a least element.*

Proof. Arguing by contradiction, assume that there exists a subset T of $\mathbb{N}$ without a minimum. Define

$$S = \{n \in \mathbb{N} \mid n \text{ is a lower bound of } T\}.$$

The set S is nonempty since 1 is a lower bound of every subset of $\mathbb{N}$. Assume that $n \in S$. Then we can show that also $n + 1 \in S$. Indeed, since n is a lower bound of T, we have that $n \leq x$ for all $x \in T$. But since T does not have a minimum, we have that $n \notin T$, implying that $n < x$ for all $x \in T$. This implies that $n + 1 \leq x$ for all $x \in T$, i.e., $n + 1$ is a lower bound of T and it belongs to S. Thus, we have shown that $1 \in S$, and if $n \in S$, then $n + 1 \in S$. Therefore, by induction, $S = \mathbb{N}$.

Now take any $x \in T$ (which exists since T is nonempty). Since $T \subseteq \mathbb{N} = S$, we have that $x \in S$. Therefore, by the definition of S, x is a lower bound of T and since $x \in T$, it follows that x is the minimum of T, which is a desired contradiction. □

From $\mathbb{N}$, we can construct the **integers** $\mathbb{Z}$, by $\{\pm n : n \in \mathbb{N}\}$. The integers have additive inverses (in addition to all the nice properties of $\mathbb{N}$), making it a "commutative ring".

From $\mathbb{Z}$, we construct the **rational numbers** $\mathbb{Q}$ by $\{a/b : a, b \in \mathbb{Z}, b \neq 0\}/\sim$ where $a/b \sim c/d$ if $ad = bc$. We have now added multiplicative inverses (for every number except zero), which makes $\mathbb{Q}$ an "ordered field", in the sense of the following definition.

Definition 2.1.3. A **field** is a set F with two operations, called **addition** and **multiplication**, which satisfy the following "field axioms" for all $x, y, z \in F$:

(A) Axioms for addition:

(A1) Closure of addition: $x + y \in F$.

(A2) Commutative law for addition: $x + y = y + x$.

(A3) Associative law for addition: $x + (y + z) = (x + y) + z$.

(A4) Existence of additive identity: there exists $0 \in F$ such that $x + 0 = x$.

(A5) Existence of additive inverse: there exists $-x \in F$ such that $x + (-x) = 0$.

(M) Axioms for multiplication:

- (M1) Closure of multiplication: $xy \in F$.
- (M2) Commutative law for multiplication: $xy = yx$.
- (M3) Associative law for multiplication: $x(yz) = (xy)z$.
- (M4) Existence of multiplicative identity: there exists $1 \neq 0$ in F such that $x \cdot 1 = x$.
- (M5) Existence of multiplicative inverse: if $x \neq 0$, then there exists $x^{-1} \in F$ such that $x \cdot x^{-1} = 1$.

(D) Distributive law: $x(y + z) = xy + xz$.

An **ordered field** is a field F in which an **order**, denoted by $<$, is defined such that the following "order axioms" are satisfied for all $x, y, z \in F$:

(O1) Trichotomy law: one and only one of the statements $x = y$, $x > y$, or $x < y$ is true.

(O2) If $x < y$ and $y < z$, then $x < z$.

(O3) If $x < y$, then $x + z < y + z$.

(O4) If $x > 0$ and $y > 0$, then $xy > 0$.

So, with all these nice properties, what are the rational numbers missing? A sense of *completeness* with respect to the following operator.

Definition 2.1.4. Let F be an ordered field and $S \subseteq F$.

(a) If there exists $m \in F$ such that $m \geq s$ for all $s \in S$, then S is **bounded above** and m is an **upper bound** for S.

(b) If m is an upper bound for S and $m \in S$, then m is the **maximum** of S, written $m = \max S$.

(c) If m is an upper bound for S and for all $m' < m$, there exists $s' \in S$ such that $s' > m'$, then m is the **supremum** or **least upper bound** of S, written $m = \sup S$.

The lower bound, minimum and infimum ($\inf S$) are defined analogously.

The set of **real numbers** $\mathbb{R}$ is obtained as the (equivalence classes of) partitions of the rationals into **Dedekind cuts**: two nonempty subsets A and A^c such that A is closed downwards ($x < y$ and $y \in A$ implies $x \in A$) and A does not have a maximum. For example, $\sqrt{2}$ may be defined by the Dedekind cut with $A = \{x \in \mathbb{Q} : x < 0 \text{ or } x^2 < 2\}$. This gives the following.

Theorem 2.1.5. *The set of* ***real numbers*** $\mathbb{R}$ *is a* ***complete ordered field****. That is, in addition to the field axioms and the order axioms,* $\mathbb{R}$ *also satisfies the* ***completeness axiom****:*

> *Every nonempty subset S of $\mathbb{R}$ that is bounded above has a least upper bound. That is, $\sup S$ exists and is a real number.*

An interesting fact that we will not prove (or even formally define) is that $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are the canonical examples of each of the algebraic structures we described (i.e., commutative semiring, commutative ring, ordered field and complete ordered field), so that any non-trivial examples of such algebraic structures contains a copy of these canonical examples inside them.

For some applications, $\mathbb{R}$ is still not big enough. For example, not all polynomials with real (or even integer) coefficients have solutions over the field $\mathbb{R}$. To overcome this problem, we can add an imaginary unit i defined by $i^2 = -1$ and let $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$. A downside is that in doing so, we lose the *ordered* property of the field: $\mathbb{C}$ is just a complete field.

Exercise 2.1. *Define the absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ by*

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

Prove that $|x_1 + x_2| \leq |x_1| + |x_2|$ and that

$$|x_1 + x_2 + \ldots + x_n| \leq |x_1| + |x_2| + \ldots + |x_n|$$

for any $n \in \mathbb{N}$ and $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Exercise 2.2. *Let $X, Y \subseteq \mathbb{R}^n$ and $g : X \times Y \to \mathbb{R}$. Show that*

$$\sup_{y \in Y} \inf_{x \in X} g(x, y) \leq \inf_{x \in X} \sup_{y \in Y} g(x, y).$$

2.2 Properties of Real Numbers

Theorem 2.2.1. *Suppose A and B are nonempty subsets of $\mathbb{R}$. Let*

$$C = \{x + y : x \in A \text{ and } y \in B\}.$$

If A and B have suprema, then C has a supremum and

$$\sup C = \sup A + \sup B.$$

Proof. Let $a = \sup A$ and $b = \sup B$. For any $z \in C$, there exists $x \in A$ and $y \in B$ such that $z = x+y$. Thus, $z = x + y \leq a + b$, so $a + b$ is an upper bound for C. By the completeness axiom, $\sup C$ exists and is a real number. Let $c = \sup C$. Since c is the least upper bound, we have $c \leq a + b$.

Now suppose $a + b > c$. This means $a + b - c > 0$. Let

$$\varepsilon = \frac{a + b - c}{2} > 0.$$

Since a is supremum of A, $a - \varepsilon$ is not an upper bound, and there exists $x \in A$ such that $x > a - \varepsilon$. Similarly, there exists $y \in B$ such that $y > b - \varepsilon$. It follows that

$$a + b \geq x + y > a + b - 2\varepsilon = a + b - 2\left(\frac{a + b - c}{2}\right) = c.$$

Thus, we have found $z = x + y$ that is greater than c, a contradiction. Therefore, $a + b \leq c$, and hence, $c = a + b$. □

Theorem 2.2.2. *$\mathbb{N}$ is unbounded above in $\mathbb{R}$.*

Proof. Suppose $\mathbb{N}$ is bounded above. Then by the completeness axiom, $\sup \mathbb{N}$ exists and is in $\mathbb{R}$. Let $m = \sup \mathbb{N}$. Since m is the least upper bound, $m - 1$ is not an upper bound for $\mathbb{N}$. Thus, there exists $n \in \mathbb{N}$ such that $n > m - 1$. But then $n + 1 > m$, and since $n + 1 \in \mathbb{N}$, this contradicts m being an upper bound for $\mathbb{N}$. Thus, $\mathbb{N}$ is unbounded above. □

Theorem 2.2.3. *The following statements are equivalent:*

(a) *$\mathbb{N}$ is unbounded above in $\mathbb{R}$.*

(b) *For each $z \in \mathbb{R}$, there exists $n \in \mathbb{N}$ such that $n > z$.*

(c) *For each $x > 0$ and for each $y \in \mathbb{R}$, there exists $n \in \mathbb{N}$ such that $nx > y$.*

(d) *For each $x > 0$, there exists $n \in \mathbb{N}$ such that $0 < 1/n < x$.*

Proof. Suppose that (a) is true, but (b) does not hold. That is, suppose there exists $z \in \mathbb{R}$ such that $n \leq z$ for all $n \in \mathbb{N}$. But this means $\mathbb{N}$ is bounded above, which contradicts (a). Thus, (a) implies (b).

Suppose that (b) is true and let $z = y/x$. Then there exists $n \in \mathbb{N}$ such that $n > y/x$, so $nx > y$. Thus, (b) implies (c).

Suppose that (c) is true and let $y = 1$. Then there exists $n \in \mathbb{N}$ such that $nx > 1$, so $1/n < x$. Since $n \in \mathbb{N}$, $1/n > 0$. Thus, (c) implies (d).

Finally, suppose that (d) is true, but (a) does not hold. That is, there exists $m \in \mathbb{R}$ such that $n < m$ for all $n \in \mathbb{N}$. But this means $1/n > 1/m$ for all $n \in \mathbb{N}$, which contradicts (d) with $x = 1/m$. Thus, (d) implies (a). □

Lemma 2.2.4. *If $x \geq 0$, then there exists $n \in \mathbb{N}$ such that $n - 1 \leq x < n$.*

Proof. Let $T = \{n \in \mathbb{N} : n > x\}$. By Theorem 2.2.3b, T is nonempty. Since $T \subseteq \mathbb{N}$, T has a least element by the well-ordering principle. Let $m = \min T$. Since m is the minimum, $m - 1 \notin T$. Thus, $m - 1 \leq x < m$. □

Theorem 2.2.5. *Let $x, y \in \mathbb{R}$ such that $x < y$. Then there exists $r \in \mathbb{Q}$ such that $x < r < y$.*

This theorem is often summarized as "the rationals are dense in the reals".

Proof. Suppose $x \geq 0$. By Theorem 2.2.3d, there exists $n \in \mathbb{N}$ such that $1/n < y - x$. Thus, $nx + 1 < ny$. By Lemma 2.2.4, there exists $m \in \mathbb{N}$ such that $m - 1 \leq nx < m$. This implies $m \leq nx + 1$. But $nx + 1 < ny$, so $m < ny$. Thus we have $nx < m < ny$, or, what is the same, $x < m/n < y$.

Now suppose $x < 0$ and $y > 0$. Combining these inequalities give $x < 0 < y$.

Finally, suppose $x < 0$ and $y \leq 0$. That is, $x < y \leq 0$, which means $0 \leq -y < -x$. So by the first part of the proof, there exists $r \in \mathbb{Q}$ such that $-y < r < -x$. Thus, $x < -r < y$. □

Lemma 2.2.6. *Let x be a nonzero rational number and y be irrational. Then xy is irrational.*

Proof. Since x is rational, we can write $x = m/n$ for some nonzero integers m and n. Now suppose that xy is rational. Then we can write $xy = p/q$ for some $p, q \in \mathbb{Z}$. It follows that

$$y = \frac{xy}{x} = \frac{p/q}{m/n} = \frac{pn}{qm},$$

so y would also be rational, a contradiction. Thus, xy is irrational. □

Theorem 2.2.7. *Let $x, y \in \mathbb{R}$ and $x < y$. Then there exists an irrational number w such that $x < w < y$.*

That is, the irrationals are also dense in the reals!

Proof. By Theorem 2.2.5, we can obtain a rational number $r \neq 0$ such that

$$\frac{x}{\sqrt{2}} < r < \frac{y}{\sqrt{2}}.$$

It follows that $x < r\sqrt{2} < y$, where $w = r\sqrt{2}$ is irrational by Lemma 2.2.6. □

Exercise 2.3. *For each $n \in \mathbb{N}$, let $I_n = [a_n, b_n]$. Suppose that $I_{n+1} \subseteq I_n$ for each $n \in \mathbb{N}$. Show that $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$.*

Now suppose that $I_n = \left(-\frac{1}{n}, \frac{1}{n}\right)$. Show that $\bigcap_{n=1}^{\infty} I_n = \{0\}$

2.3 Infinities

Definition 2.3.1. Let S and T be sets. We say:

(a) S and T are **equipotent** or have equal cardinality, denoted $S \sim T$ or $|S| = |T|$, if there exists a bijection from S to T, and

(b) S has cardinality less than T, $|S| \leq |T|$, if there exists an injective function from S to T.

These two definitions are consistent by the following important theorem.

Theorem 2.3.2 (Cantor-Schröder-Bernstein Theorem)**.** *Suppose there exists injective functions $f : S \to T$ and $g : T \to S$. Then, there exists a bijective function $h : S \to T$.*

We also have the following natural results about the cardinality of sets.

Theorem 2.3.3. *Let S, T, and U be sets. Then the following properties hold.*

(a) If $S \subseteq T$, then $|S| \leq |T|$.

(b) $|S| = |S|$.

(c) If $|S| \leq |T|$ and $|T| \leq |U|$, then $|S| \leq |U|$.

(d) If $m, n \in \mathbb{N}$ and $m \leq n$, then $|\{1, 2, \dots, m\}| \leq |\{1, 2, \dots, n\}|$.

(e) If $m \in \mathbb{N}$, then $|\{1, 2, \dots, m\}| < |\mathbb{N}|$ and $|\{1, 2, \dots, m\}| < |\mathbb{R}|$.

Proof. (a) The identity function on S is an injection from S into T. Thus, $|S| \leq |T|$.

(b) The identity function on S is a bijection from S onto itself. Thus, $|S| = |S|$.

(c) Suppose $|S| \leq |T|$ and $|T| \leq |U|$. This means there exist injections $f : S \to T$ and $g : T \to U$. Since the composite function of two injections is injective, $g \circ f : S \to U$ is also injective. Thus, $|S| \leq |U|$.

(d) The identity mapping $i \mapsto i$ is an injection.

(e) If S is finite, then clearly there exists an injection from S into $\mathbb{N}$. It is not difficult to see that we cannot find a surjection from S onto $\mathbb{N}$, which is infinite. Thus, $|S| < \aleph_0$.

□

Definition 2.3.4. Let S be a set and write $I_n = \{1, 2, \dots, n\}$. We say

(a) S is **finite** if $S = \emptyset$ or $I_n \sim S$ for some $n \geq 1$ and say its **cardinality** $|S|$ is n.

(b) S is **infinite** if it is not finite.

(c) S is **countably infinite** if $\mathbb{N} \sim S$.

(d) S is **countable** if it is finite or countably infinite.

(e) S is **uncountable** if it is not countable.

The cardinality of $\mathbb{N}$ is denoted by $\aleph_0$ and the cardinality of $\mathbb{R}$ is denoted by $\mathfrak{c}$.

Theorem 2.3.5. *$\mathbb{Z}$ is countably infinite.*

Proof. Define $f : \mathbb{N} \to \mathbb{Z}$ by

$$f(n) = \begin{cases} (n-1)/2 & \text{if } n \text{ is odd,} \\ -n/2 & \text{if } n \text{ is even.} \end{cases}$$

It is not difficult to see that f is bijective. Thus, $\mathbb{N} \sim \mathbb{Z}$, so $\mathbb{Z}$ is countably infinite. □

Theorem 2.3.6. *Every subset of a countable set is countable.*

Proof. Let S be a countable set and let $T \subseteq S$. If T is finite, there is nothing to prove. Suppose T is infinite. This implies that S is countably infinite. Since $\mathbb{N} \sim S$, we can write the elements of S as $x_1, x_2, \ldots$. Now define

$$A = \{n \in \mathbb{N} : x_n \in T\}.$$

Let n_1 be the smallest element in A. Having chosen $n_1, n_2, \ldots, n_{k-1}$, let n_k be the smallest element in A greater than n_{k-1}. Then the function $f : \mathbb{N} \longrightarrow T$ defined by $f(k) = x_{n_k}$ is a bijection. Hence, T is countably infinite. □

Theorem 2.3.7. *Let S be a nonempty set. Then the following are equivalent:*

(a) S is countable.

(b) There exists an injection $f : S \longrightarrow \mathbb{N}$.

(c) There exists a surjection $g : \mathbb{N} \longrightarrow S$.

Proof. Suppose S is countable. Then there exists a bijection $h : J \longrightarrow S$, where $J = I_n$ for some $n \in \mathbb{N}$ if S is finite and $J = \mathbb{N}$ if S is infinite. In either case, h^{-1} is at least an injection from S to $\mathbb{N}$. Thus, (a) implies (b).

Suppose there exists an injection $f : S \longrightarrow \mathbb{N}$. Then f is a bijection from S onto $f(S)$, so f^{-1} is a bijection from $f(S)$ onto S. Let p be any fixed member of S. We define $h : \mathbb{N} \longrightarrow S$ by

$$h(n) = \begin{cases} f^{-1}(n), & \text{if } n \in f(S), \\ p, & \text{if } n \notin f(S), \end{cases}$$

which is a surjection. Thus, (b) implies (c).

Finally, suppose there exists a surjection $g : \mathbb{N} \longrightarrow S$. Define $h : S \longrightarrow \mathbb{N}$ by

$$h(s) = \text{the smallest } n \in \mathbb{N} \text{ such that } g(n) = s,$$

which is a bijection from S onto $h(S)$. Since $h(S) \subseteq \mathbb{N}$, $h(S)$ is countable. But $h(S) \sim S$, so S is also countable. Thus, (c) implies (a). □

Theorem 2.3.8. $\mathbb{N} \times \mathbb{N}$ *is countable.*

Proof. By Theorem 9.3.11, it suffices to show that there is an injection from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$. Let f be defined as

$$f(m, n) = 2^m 3^n.$$

To see that f is injective, suppose $2^m3^n = 2^p3^q$. If $m < p$, then $3^n = 2^{p-m}3^q$, contradicting the fact that 3^n is odd for all n. We arrive at a similar contradiction if $m > p$, so we must have $m = p$. This implies $3^n = 3^q$. Hence, $n = q$. □

Theorem 2.3.9. *A countable union of countable sets is countable.*

Proof. Let $\{A_j\}_{j\in J}$ be an indexed family of countable sets, where the index J is countable. Since empty sets contribute nothing to the union, we may assume that all the sets are nonempty. Since each A_j is a countable set, there exists, for each j, a surjection $f_j : \mathbb{N} \longrightarrow A_j$. Similarly, there exists a surjection $g : \mathbb{N} \longrightarrow J$. Now, define

$$h : \mathbb{N} \times \mathbb{N} \longrightarrow \bigcup_{j\in J} A_j$$

by the equation

$$h(k, m) = f_{g(k)}(m).$$

It is not difficult to see that h is surjective. Since there exists a bijection between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$, the countability of the union follows. □

Theorem 2.3.10. $\mathbb{Q}$ *is countable.*

Proof. Denote the set of positive rationals and negative rationals by $\mathbb{Q}^+$ and $\mathbb{Q}^-$, respectively. Consider first $\mathbb{Q}^+$. Any member of $\mathbb{Q}^+$ can be written uniquely as m/n, where $m, n \in \mathbb{N}$, $n \neq 0$, and m and n have no common prime divisors. Define $f : \mathbb{Q}^+ \longrightarrow \mathbb{N}$ by:

$$f(m/n) = 2^m3^n.$$

As shown previously, f is injective, so $\mathbb{Q}^+$ is countable. The function $g : \mathbb{Q}^+ \longrightarrow \mathbb{Q}^-$ defined by $g(r) = -r$ is clearly bijective. Thus, $\mathbb{Q}^+ \sim \mathbb{Q}^-$, so $\mathbb{Q}^-$ is countable. Since $\mathbb{Q} = \mathbb{Q}^+ \cup \{0\} \cup \mathbb{Q}^-$, $\mathbb{Q}$ is countable. □

Theorem 2.3.11 (Cantor)**.** $\mathbb{R}$ *is uncountable.*

Proof. Since every subset of a countable set is countable, it is enough for us to prove that a subset $J = (0, 1)$ of $\mathbb{R}$ is uncountable. Suppose J is countable, then we could list its members as

$x_1, x_2, x_3, \ldots$. Since each element of J has an infinite decimal expansion, we can write:

$$\begin{aligned} x_1 &= 0.a_{11}a_{12}a_{13}\ldots \\ x_2 &= 0.a_{21}a_{22}a_{23}\ldots \\ x_3 &= 0.a_{31}a_{32}a_{33}\ldots \\ &\vdots \end{aligned}$$

where each $a_{ij} \in \{0, 1, ..., 9\}$. We now construct a real number $y = 0.b_1b_2b_3...$ by defining

$$b_n = \begin{cases} 2, & \text{if } a_{nn} \neq 2, \\ 3, & \text{if } a_{nn} = 2. \end{cases}$$

Clearly, $y \in (0, 1)$. However, y is not one of the numbers x_n, since it differs from x_n in the nth decimal place. This contradicts our assumption that J is countable, so J is uncountable. □

Corollary 2.3.12. *The set of irrational numbers is uncountable.*

Proof. Since $\mathbb{R}$ is uncountable, $\mathbb{Q}$ is countable, and $\mathbb{R}$ is the union of the rationals and irrationals, it follows immediately that the set of irrational numbers is uncountable. □

Theorem 2.3.13. $\aleph_0 < \mathfrak{c}$.

Proof. Since $\mathbb{N} \subseteq \mathbb{R}$, we have $\aleph_0 \leq \mathfrak{c}$. In fact, since $\mathbb{N}$ is countable and $\mathbb{R}$ is uncountable, $\mathbb{N}$ and $\mathbb{R}$ cannot be equipotent. Hence, we have $\aleph_0 < \mathfrak{c}$. □

Exercise 2.4. *Let X be a set and $f : X \longrightarrow 2^X$. Show that f is not surjective.*

Exercise 2.5. *Let $X = \mathbb{R}^2$ and define a relation on X by $(x_1, y_1) \succeq (x_2, y_2)$ if $x_1 > x_2$ or $x_1 = x_2$ and $y_1 \geq y_2$.*

(a) Show that X is a complete and transitive preference relation.

(b) Now show that there exists no function $u : X \longrightarrow \mathbb{R}$ such that $x \succeq y$ if and only if $u(x) \geq u(y)$.

Repeat the above exercises for the relation defined by $(x_1, y_1) \succeq (x_2, y_2)$ if and only if either $\min\{x_1, y_1\} > \min\{x_2, y_2\}$ or $\min\{x_1, y_1\} = \min\{x_2, y_2\}$ and $x_1 + y_1 \geq x_2 + y_2$.

Part II

Linear Algebra

3

Linear Algebra

Contents

3.1 Vectors

A vector is basically a list of numbers which can be added together. A formal definition is below.

Definition 3.1.1. A **vector space** (or **linear space**) over the field K is a set V of *vectors* equipped with two operations, vector addition ("+") and scalar multiplication such that:

- Closure: For $a \in K$, $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x}+\mathbf{y} \in V$ and $a\mathbf{x} \in V$;
- Commutativity of addition: For $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$;
- Distributivity of addition: For $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, $(\mathbf{x}+\mathbf{y})+\mathbf{z} = \mathbf{x}+(\mathbf{y}+\mathbf{z})$;
- Additive identity: There is a vector $\mathbf{0} \in V$ such that for any $\mathbf{x} \in V$, $\mathbf{x}+\mathbf{0} = \mathbf{x}$;
- Additive inverse: For any $\mathbf{x} \in V$, $\mathbf{x}+(-1)\mathbf{x} = \mathbf{0}$;
- Distributivity of scalar multiplication: For $a, b \in K$, $\mathbf{x} \in V$, $a(b\mathbf{x}) = (ab)\mathbf{x}$;
- Identity for scalar multiplication: For $\mathbf{x} \in V$, $1\mathbf{x} = \mathbf{x}$;
- Distributive laws: For $a, b \in K$, $\mathbf{x}, \mathbf{y} \in V$, $a(\mathbf{x}+\mathbf{y}) = (a\mathbf{x})+(a\mathbf{y})$ and $(a+b)\mathbf{x} = (a\mathbf{x})+(b\mathbf{x})$.

Example. $\mathbb{Q}^n$, $\mathbb{R}^n$, $\mathbb{C}^n$ for $n \in \{1, 2, \ldots\}$, equipped with usual notions of vector addition and scalar multiplication, is a vector space over the fields $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ respectively. ♣

Example. The set of all bounded functions $B(X)$ from a set X to a field K is also a vector space, where the sum of two functions f and g is given by $(f+g)(x) = f(x)+g(x)$ and $(cf)(x) = cf(x)$. So is the set of all continuous function $C(X)$ and the set of all k-times differentiable functions $C^k(X)$ (assuming X is an open set so that the derivative is well-defined: we will come back to these definitions later on!). ♣

Definition 3.1.2. A **linear combination** of $\mathbf{x}_1, ..., \mathbf{x}_m \in V$ is any sum of scalar multiples of vectors of the form $a_1\mathbf{x}_1 + ... + a_m\mathbf{x}_m$, $a_i \in K$, $\mathbf{x}_i \in V$.

Definition 3.1.3. A **linear subspace** (or **vector subspace**) M of V is a subset of V that is closed under linear combinations. A linear subspace of a vector space is a vector space in its own right.

Example. $\{0\}$ is a linear subspace of $\mathbb{R}$. For any $x \in \mathbb{R}^n$, $\{y : y = \alpha x \text{ for some } \alpha \in \mathbb{R}\}$ is a linear subspace of $\mathbb{R}^n$. ♣

Definition 3.1.4. Let $E \subseteq V$. The *span* of E, denoted span$E$ is the set of all finite linear combinations from E. That is

$$\text{span}E = \{\sum_{i=1}^{m} a_i\mathbf{x}_i : a_i \in K, \mathbf{x}_i \in E, m \in \mathbb{N}\}$$

Definition 3.1.5. A set E of vectors is **linearly dependent** if there are distinct vectors $\mathbf{x}_1, ..., \mathbf{x}_m \in E$ and nonzero scalars $a_1, ..., a_m \in K$ such $\sum_{i=1}^{m} a_i\mathbf{x}_i = 0$. The set of vectors is **linearly independent** if it is not dependent. That is, E is independent if for every set $\mathbf{x}_1, ..., \mathbf{x}_m$ of distinct vectors in E, $\sum_{i=1}^{m} a_i\mathbf{x}_i = 0$ implies $a_1 = ... = a_m = 0$.

Note that the set $\{0\}$ containing only the zero vector is linearly dependent according to this definition (as is any set containing the zero vector).

Theorem 3.1.6. *(Uniqueness of linear combinations) If E is a linearly independent set of vectors and* $\mathbf{z} \in$ *spanE, then* $\mathbf{z}$ *is a unique linear combination of elements of E.*

Proof. If $\mathbf{z} = \mathbf{0}$, the conclusion follows by definition of independence. If $\mathbf{z} \neq \mathbf{0}$, suppose that

$$\mathbf{z} = \sum_{i=1}^{m} a_i \mathbf{x}_i = \sum_{j=1}^{n} b_j \mathbf{y}_j$$

where $\mathbf{x}_i$s are distinct elements of E and $\mathbf{y}_j$s are distinct elements of E (but may overlap with the $\mathbf{x}_i$s), and $a_i, b_j \neq 0$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$. Enumerate $A = \{\mathbf{x}_i : i = 1, \ldots, m\} \cup \{\mathbf{y}_j : j = 1, \ldots, n\}$ as $A = \{\mathbf{z}_k : k = 1, \ldots, p\}$. Then we can rewrite $\mathbf{z} = \sum_{k=1}^{p} \hat{\alpha}_k \mathbf{z}_k = \sum_{k=1}^{p} \hat{\beta}_k \mathbf{z}_k$, where

$$\hat{\alpha}_k = \begin{cases} \alpha_i & \text{if } \mathbf{z}_k = \mathbf{x}_i \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \hat{\beta}_k = \begin{cases} \beta_j & \text{if } \mathbf{z}_k = \mathbf{y}_j \\ 0 & \text{otherwise} \end{cases}$$

Then

$$0 = \mathbf{z} - \mathbf{z} = \sum_{k=1}^{p} \left(\hat{\alpha}_k - \hat{\beta}_k \right) \mathbf{z}_k \Rightarrow \hat{\alpha}_k - \hat{\beta}_k = 0, \; k = 1, \ldots, p$$

since E is independent. Therefore $\hat{\alpha}_k = \hat{\beta}_k$, $k = 1, \ldots, p$, which in turn implies $m = n = p$ and $\{\mathbf{x}_i : i = 1, \ldots, m\} = \{\mathbf{y}_j : j = 1, \ldots, n\}$. □

Definition 3.1.7. A **Hamel basis** for a linear space V is a linearly independent set B such that span$B = V$.

Example. The set of unit coordinate vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n$ is a basis for $\mathbb{R}^n$, called the **standard basis**. Any vector $\mathbf{x} \in \mathbb{R}^n$ can be written uniquely as $\mathbf{x} = \sum_{i=1}^{n} = x_i \mathbf{e}_i$. ♣

Lemma 3.1.8. *Every vector space has a Hamel basis. Any two bases have the same cardinality, called the* ***dimension*** *of V.*

Note that the Hamel basis of infinite-dimensional vector spaces can be unwieldy: remember that each vector must be a linear combination of *finitely* many vectors in the basis. In these notes, we will avoid lengthy discussion of infinite-dimensional vector spaces. In these cases, it can be useful to define an alternative notion of a basis in which vectors may be written as *infinite* sums of basis vectors (e.g., Schauder bases). Occasionally, we will do examples using infinite-dimensional vector spaces (e.g., the space of all functions on $[0, 1]$), but approach these with caution: some of the intuitions from finite-dimensional linear algebra do not carry directly over to infinite-dimensional spaces.

Theorem 3.1.9. *In an n-dimensional space, every set of more than n vectors is dependent. Consequently, any independent set of n-vectors is a basis.*

Exercise 3.1. *Show that the following set is a basis for* $\mathbb{R}^3$*:*

$$\left\{\begin{pmatrix}1\\1\\0\end{pmatrix},\begin{pmatrix}-1\\0\\1\end{pmatrix},\begin{pmatrix}2\\1\\1\end{pmatrix}\right\}.$$

3.2 Inner products and norms

Definition 3.2.1. The **dot product** (or inner product or scalar product) of two vectors **x** and **y** in $\mathbb{R}^n$ is defined by

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

The vector space $\mathbb{R}^n$ equipped with the dot product is called *Euclidean n-space.*

The dot product is an example of an inner product, which may be defined more generally on vector spaces as follows.

Definition 3.2.2. An **inner product space** is a vector space V over field K equipped with an **inner product**

$$\langle\cdot,\cdot\rangle : V \times V \longrightarrow K$$

which is a function satisfying:

- Linearity in first argument: $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$.
- Positive definiteness: $\langle x, x\rangle \geq 0$ with equality only for $x = 0$.

- Conjugate symmetry: $\langle x, y\rangle = \overline{\langle y, x\rangle}$. In particular, if K is the real field, then $\langle x, y\rangle = \langle y, x\rangle$.

Note that conjugate symmetry implies that $\langle x, x\rangle$ is always a real number, so that the formulation of positive definiteness is correct (remember, the order $\geq$ implies we must be working with real numbers since the complex field is not ordered).

Inner products allow us to formalize definition of geometric notions, like lengths, angles and orthogonality of vectors. Let us start with the first notion.

Definition 3.2.3. The **norm induced by the inner product** is given by $\|v\| = \sqrt{\langle v, v\rangle}$.

Sometimes it is possible to define length without an inner product. This gives rise to normed vector spaces.

Definition 3.2.4. A **normed vector space** is a vector space V over a field K equipped with a **norm**

$$\|\cdot\| : V \longrightarrow K,$$

which is a function satisfying

- Nonnegativity: $\|x\| \geq 0$ for all $x \in V$ with equality only for $x = 0$.
- Linearity: $\|\alpha x\| = |\alpha|\|x\|$ for all $\alpha \in K, x \in V$.
- Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$.

A vector in V with $\|x\| = 1$ is called a unit vector. Unit vectors are often notated with hats on top to remind us that their norm is 1.

You should check that the norm induced by the inner product satisfies all these properties. We have seen that inner products induce norms, but not all normed vector spaces have inner products. In fact, only normed spaces satisfying the **parallelogram law**

$$2\|x\|^2 + 2\|y\|^2 = \|x + y\|^2 + \|x - y\|^2$$

may have an inner product. The inner product (consistent with $\|x\| = \sqrt{\langle x, x\rangle}$) is given by the formula

$$\langle x, y\rangle = \frac{\|x + y\|^2 - \|x - y\|^2}{4}.$$

Theorem 3.2.5. *Let $V = \mathbb{R}^m$ and $p \geq 1$. The L^p norm*

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p\right)^{\frac{1}{p}}$$

is a norm, which is not induced by any inner product (except in the case of $n = 2$). When $p = \infty$, the L^∞ or sup norm is defined by

$$\|x\|_\infty = \max_{1 \leq i \leq m} |x_i|,$$

which is also not induced by an inner product.

The triangle inequality in the L^p spaces is of independent interest, and is often known as **Minkowski's Inequality**:

$$\left(\sum_{i=1}^{m} |x_i + y_i|^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{m} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{m} |x_i|^p\right)^{\frac{1}{p}}.$$

An even stronger result is **Hölder's Inequality**:

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} |y_k|^q\right)^{\frac{1}{q}} \text{ for } p, q \text{ such that } \frac{1}{p} + \frac{1}{q} = 1.$$

Let us now return to inner product spaces. An important result is the following inequality.

Theorem 3.2.6 (Cauchy-Schwarz-Bunyakovski Inequality)**.** *Let V be an inner product space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$ and $x, y \in V$. Then we have*

$$|\langle x, y \rangle| \leq \|x\| \|y\|.$$

Equality occurs only if x and y are collinear (i.e., $x = \alpha y$ for some $\alpha \in K$).

The Cauchy-Schwarz inequality allows us to define notions related to angle.

Definition 3.2.7. Let V be an inner product space.

(a) Vectors $\mathbf{x}, \mathbf{y} \in V$ are **orthogonal** if $\langle x, y \rangle = 0$. Sometimes we write $x \perp y$.

(b) A set of vectors $E \subset V$ is **orthogonal** if it is pairwise orthogonal.

(c) A set E is **orthonormal** if E is orthogonal and $\|\mathbf{x}\| = 1$ for all $\mathbf{x} \in E$.

(d) The angle between x, y is defined by $\angle(x, y) = \arccos\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right)$.

Lemma 3.2.8. *If a set of nonzero vectors is orthogonal, then the set is independent.*

Proof. Suppose that $\sum_{i=1}^{n} a_i\mathbf{x}_i = 0$, where $\mathbf{x}_i$s are orthogonal. Then for each k,

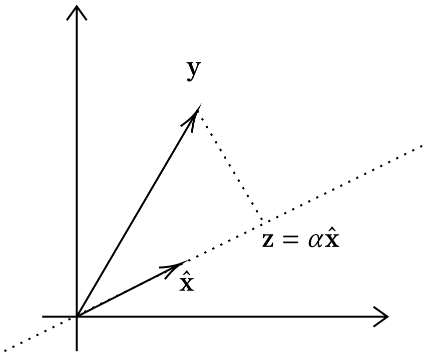
$$0 = \mathbf{x}_k \cdot \mathbf{0} = \mathbf{x}_k \cdot \left(\sum_{i=1}^{n} a_i\mathbf{x}_i\right) = \sum_{i=1}^{n} a_i\mathbf{x}_k \cdot \mathbf{x}_i = a_k\mathbf{x}_k \cdot \mathbf{x}_k$$

which implies that $a_k = 0$. □

Consider an inner product space V and the line passing through the origin in the direction of a unit vector $\hat{\mathbf{x}} \in V$, i.e. the set $\{\alpha\hat{\mathbf{x}} : \alpha \in K\}$. A common problem is to determine the point $\mathbf{z}$ on the line which is closest to another point $\mathbf{y}$, in terms of $\|\mathbf{z} - \mathbf{y}\|$. This corresponds to our first optimization problem

$$\min_{\alpha\in K} \|\alpha\hat{\mathbf{x}} - \mathbf{y}\|.$$

We call the point $\mathbf{z} = \alpha\hat{\mathbf{x}}$ the **vector projection** of $\mathbf{y}$ on $\hat{\mathbf{x}}$ and the associated scalar α the **scalar projection** of $\mathbf{y}$ on $\hat{\mathbf{x}}$.



We have all the tools at hand to solve this problem (without calculus). We begin by squaring the objective, noting that the minimizer of a strictly positive function is the same as the minimizer of that function's square. Thus, the squared objective is

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

The second and third terms in the above sum are constants with respect to α, while the first term is clearly minimized by setting $\alpha^* = \langle \hat{\mathbf{x}}, \mathbf{y} \rangle$. This is the scalar projection of $\mathbf{y}$ on $\hat{\mathbf{x}}$. The vector projection is then just $\alpha^* \hat{\mathbf{x}}$. This leads to an interpretation of the inner product as the projection onto a unit vector. (If you are applying this fact, remember that we have assumed that $\|x\| = 1$, so don't forget to normalize the vector of interest first!)

It turns out it is always possible to choose a basis of a finite-dimensional inner product space V which is orthogonal. This is achieved by projecting basis vectors on one another, maintaining only the orthogonal part. This is the so-called *Gram-Schmidt procedure*.

Theorem 3.2.9 (Gram-Schmidt Orthogonalization)**.** *Let V be a vector space with an inner product. Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is a basis for V. Let $\mathbf{v}_1 = \mathbf{x}_1$,*

$$\begin{aligned}
\mathbf{v}_2 &= \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 \\
\mathbf{v}_3 &= \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{x}_3, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2, \\
&\ldots \\
\mathbf{v}_n &= \mathbf{x}_n - \frac{\langle \mathbf{x}_n, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \cdots - \frac{\langle \mathbf{x}_n, \mathbf{v}_{n-1} \rangle}{\langle \mathbf{v}_{n-1}, \mathbf{v}_{n-1} \rangle} \mathbf{v}_{n-1}.
\end{aligned}$$

Then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is an orthogonal basis for V.

An orthogonal basis is **orthonormal** if moreover the norm of each vector in the basis is 1. It is easy to take the output of the Gram-Schmidt Orthogonalization and normalize it by dividing each vector by its length in order to obtain an orthonormal basis.

Exercise 3.2. *Is the basis in Exercise 3.1 orthonormal? If not, orthogonalize it.*

Note that in infinite-dimensional inner product spaces, the above is *not* the correct definition of an orthonormal basis: instead, an orthonormal basis is a (possibly uncountable) set of vectors $\{x_\alpha\}_{\alpha \in A}$ such that $\langle x_\alpha, x_\alpha \rangle = 1$, $\langle x_\alpha, x_\beta \rangle = 0$ for $\alpha \neq \beta$ and $\langle y, x_\alpha \rangle = 0$ for all $\alpha \in A$ iff $y = 0$. Hopefully it should be clear that these definitions are equivalent in finite-dimensional spaces. But in infinite-dimensional spaces, these definitions are not equivalent because the orthonormal basis as just defined may not be a Hamel basis.

Exercise 3.3. *Let $\{x_1, ..., x_n\}$ be an orthonormal basis for V. Show that*

$$v = \sum_{i=1}^{n} \langle v, x_i \rangle x_i.$$

and $\|v\| = \sum_{i=1}^{n} |\langle v, x_i \rangle|^2$.

In finite dimensions, a consequence of the preceding exercise is the following important theorem (which is also true in *complete* infinite-dimensional vector spaces, see Section 5.3).

Theorem 3.2.10 (Riesz Representation Theorem)**.** *Let V be a (complete) inner product space and $T : V \longrightarrow \mathbb{R}$ a linear transformation. Then there exists a unique $z \in V$ such that*

$$T(x) = \langle z, x \rangle \text{ for all } x \in V.$$

3.3 Matrices

Definition 3.3.1. A **matrix** over field K is a rectangular array of scalars from the field K, or in other words, a doubly indexed ordered list of scalars. An $m \times n$ matrix $\mathbf{A}$ has m rows and n columns. It is a function $\mathbf{A} : \{1, ..., m\} \times \{1, ..., n\} \longrightarrow K$. The set of $m \times n$ matrices (over a fixed field) is denoted $\mathcal{M}(m, n)$. We define the following special matrices:

(a) A **real matrix** is a matrix over the field of real numbers ($K = \mathbb{R}$).

(b) A **square matrix** is a matrix with $m = n$.

(c) A matrix is called a **diagonal matrix** if it is a square matrix and all its nonzero entries are on the main diagonal, the set of a_{ij} with $i = j$.

(d) The **n** $\times$ n **identity matrix I** is the $n \times n$ diagonal matrix whose diagonal entries are all 1.

(e) A square matrix is **upper triangular** if the only nonzero elements are on, or above the main diagonal (i.e. $a_{ij} = 0$ for $i > j$). It is **lower triangular** if $a_{ij} = 0$ for $i < j$.

(f) The **zero matrix 0** has all its entries equal to zero.

Note vectors in K^m may be thought of as matrices with $n = 1$.

Here is a generic matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ . & & . \\ . & & . \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

Definition 3.3.2. Let $\mathbf{a}_i$ define the i-th row of $\mathbf{A}$ and $\mathbf{a}^j$ denote the j-th column. Then matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ . \\ \mathbf{a}_m \end{bmatrix} = \begin{bmatrix} \mathbf{a}^1 & \dots & \mathbf{a}^n \end{bmatrix}$$

The **column space** of $\mathbf{A}$ is the subset of K^m spanned by the n columns of $\mathbf{A}$, and its **row space** is the subspace of K^n spanned by its m rows.

There are two main interests in studying matrices. The first is their connection to linear transformations between finite-dimensional vector spaces.

Definition 3.3.3. A **linear transformation** is a function $T : V \longrightarrow W$ between vector spaces V and W satisfying $T(a\mathbf{x}+b\mathbf{y}) = aT(\mathbf{x})+bT(\mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in V$. The set of linear transformations from the vector space V into the vector space W is denoted $L(V, W)$.

Let T be a linear transformation from V into W, and let $x_1, \dots, x_n$ be an ordered basis for V and $y_1, \dots, y_m$ be an ordered basis for W. Define

$$M(T) = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ . & & . \\ . & & . \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

such that

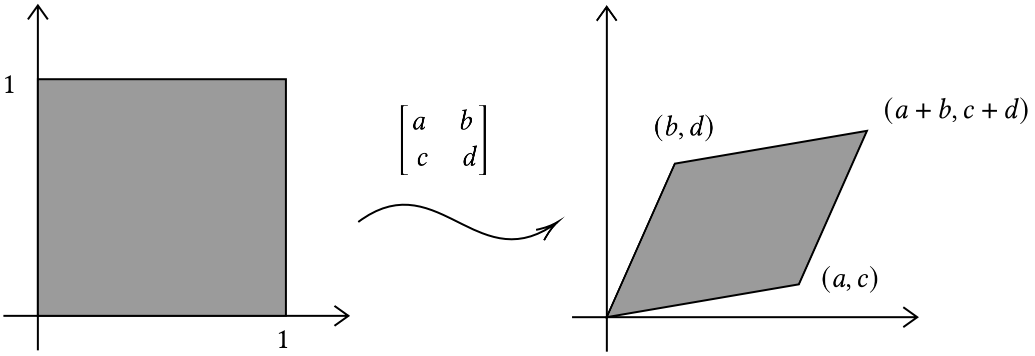
$$T(x_1) = \sum_{i=1}^{m} a_{i1} y_i$$

$$T(x_2) = \sum_{i=1}^{m} a_{i2} y_i$$

$$\vdots$$

$$T(x_n) = \sum_{i=1}^{m} a_{in} y_i$$

Then $M(T)$ is the **matrix representation of T with respect to the ordered bases** (x_i), (y_j). An illustration of this transformation is below.



The matrix representation provides a 1-to-1 mapping between matrices and linear transformations from vector spaces.

Theorem 3.3.4. *Let V be an n-dimensional vector space, and let W be an m-dimensional vector space. Fix an ordered basis for each, and let $M(T)$ be the matrix representation of the linear transformation $T : V \longrightarrow W$. Then the mapping $T \longrightarrow M(T)$ is a linear one-to-one mapping from $L(V, W)$ to $\mathcal{M}(m, n)$.*

Corollary 3.3.5. *There is a one-to-one mapping between linear transformations from $\mathbb{R}^m$ and $\mathbb{R}^n$ and $m \times n$ dimensional real matrices.*

Our second reason for studying matrices is their relationship to linear equations and inequalities. To discuss this motivation, we need to introduce a couple of matrix operations.

Definition 3.3.6. If $\mathbf{A}$ and $\mathbf{B}$ are $m \times n$ matrixes, then the **sum** $\mathbf{A} + \mathbf{B}$ is the $m \times n$ matrix $\mathbf{C}$ defined by $c_{ij} = a_{ij} + b_{ij}$. The scalar multiple of a matrix $\mathbf{A}$ by a scalar c is the $m \times n$ matrix $\mathbf{D}$ defined by $d_{ij} = ca_{ij}$.

Clearly, $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$.

Theorem 3.3.7. *The set $\mathcal{M}(m, n)$ is a vector space under the operation of matrix addition and scalar multiplication. It has dimension mn.*

Definition 3.3.8. If $\mathbf{A}$ is an $m \times p$ matrix and $\mathbf{B}$ is an $p \times n$ matrix, then the **product** of $\mathbf{A}$ and $\mathbf{B}$ is the $m \times n$ matrix $\mathbf{C}$ defined by $c_{ij} = \mathbf{a}_i' \cdot \mathbf{b}^j$, where $\cdot$ is the *dot* product. Two matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be conformable if $\mathbf{AB}$ is well-defined.

Lemma 3.3.9. *Properties of matrix multiplication*

(a) *If* $\mathbf{A}$ *is a square matrix and* $\mathbf{I}$ *is the conformable identity matrix, then* $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$.

(b) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

(c) $\mathbf{AB} \neq \mathbf{BA}$ *(in general)*

(d) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = (\mathbf{AB}) + (\mathbf{AC})$

(e) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

We now introduce the linear equations interpretation of matrices. Consider the problem of finding n scalars $x_1, \ldots, x_n \in \mathbb{R}$ which satisfy the conditions

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= y_2 \\ \ldots &= . \\ \ldots &= . \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= y_m \end{aligned}$$

We call the above system **a system of m linear equations in n unknowns**.
Any n-tuple $(x_1, \ldots, x_n)$ of elements of $\mathbb{R}$ which satisfies each of the equations above is called a solution of the system.
If $y_1 = y_2 = \ldots = y_m = 0$, we say that the system is homogeneous.

We can rewrite the above system of equations as

$$\mathbf{Ax} = \mathbf{y}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ . & & . \\ . & & . \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \mathbf{x} = \begin{bmatrix} x_1 \\ . \\ . \\ x_n \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ . \\ . \\ y_m \end{bmatrix}$$

Three important questions arise:

(i) When does there exist a solution to the above system of equations?

(ii) When is the solution unique?

(iii) How do we go about identifying a solution to the equation?

A somewhat tautological answer to the first question is that **b** needs to be in the column space of **A**. A slightly more satisfying answer involves the rank of the matrix.

Definition 3.3.10. Let **A** be a matrix.

(a) The **column rank** of A is the largest number of linearly independent columns. It is the dimension of the column space of A or the dimension of the range of the associated linear transformation.

(b) The **row rank** of A is the largest number of linearly independent rows. It is the dimension of the row space of A.

(c) The **null space** of A is the set of solutions to the homogeneous system $Ax = 0$, and its dimension is called the **nullity** of A.

Theorem 3.3.11. *The column and row ranks of any matrix are equal.*

Proof. Let **A** be a $m \times n$ matrix. Let the column rank of **A** be r, and let $\mathbf{c}^1, \dots, \mathbf{c}^r$ be any basis for the column space of **A**. Let $\mathbf{C} = [\mathbf{c}^1, \dots, \mathbf{c}^r]$. Every column of **A** can be expressed as a linear combination of the r columns in **C**. This means that there is an $r \times n$ matrix **R** such that $\mathbf{A} = \mathbf{CR}$.

$\mathbf{R}$ is the matrix whose ith column is formed from the coefficients giving the ith column of $\mathbf{A}$ as a linear combination of the r columns of $\mathbf{C}$. But each row of $\mathbf{A}$ is given by a linear combination of the r rows of $\mathbf{R}$. Therefore, the rows of $\mathbf{R}$ form a spanning set of the row space of $\mathbf{A}$ and hence the row rank of $\mathbf{A}$ cannot exceed r. We conclude that the row rank of $\mathbf{A}$ is less than or equal to the column rank of $\mathbf{A}$. Apply the same result to $\mathbf{A}'$, we can that the row rank of $\mathbf{A}'$ is less than or equal to the column rank of $\mathbf{A}'$. But the row rank of $\mathbf{A}'$ is the column rank of $\mathbf{A}$ and the column rank of $\mathbf{A}'$ is the row rank of $\mathbf{A}$. Hence the last statement is equivalent to: the column rank of $\mathbf{A}$ is less than or equal to the row rank of $\mathbf{A}$. Putting the two together, we conclude that the row rank of $\mathbf{A}$ is equal to the column rank of $\mathbf{A}$. □

Theorem 3.3.12 (Rank-Nullity Theorem)**.** *For any $m \times n$ matrix $\mathbf{A}$, the sum of the rank of A and the nullity of A is n.*

Proof. It is easy to show that the nullspace of $\mathbf{A}$ is a linear subspace of K^n. Suppose that its dimension is k (the nullity of A) and let $\mathbf{v_1}, \ldots, \mathbf{v_k}$ be a basis for the nullspace. This basis can be extended by $n - k$ linearly independent vectors to obtain a basis for K^n, and write $\{\mathbf{w_1}, \ldots, \mathbf{w_{n-k}}\}$ for these additional basis vectors.

Now, the rowspace of $\mathbf{A}$ is spanned by $\{\mathbf{Av_1}, \mathbf{Av_2}, \ldots, \mathbf{Av_k}, \mathbf{Aw_1}, \ldots, \mathbf{Aw_{n-k}}\}$. But by construction, these $\mathbf{Av_1}$, ..., $\mathbf{Av_k}$ are all the zero vector. So, the rowspace is spanned by $\mathbf{Aw_1}$, ..., $\mathbf{Aw_{n-k}}$. If we can show that this set is linearly independent, then this will establish that the dimension of the rowspace, the rank, is $n - k$, as required.

To see this, suppose otherwise, and let $\sum_j \alpha_j \mathbf{Aw_j} = 0$ for some α_j. But then $\mathbf{A}\left(\sum_j \alpha_j \mathbf{w_j}\right) = 0$ by the distributive property. But then $\sum_j \alpha_j \mathbf{w_j}$ is in the nullspace of A, which contradicts our assumption that the $\{\mathbf{w_j}\}$ were independent of the $\{v_i\}$.

□

Theorem 3.3.13 (Fundamental Theorem of Linear Algebra)**.** *Let $\mathbf{A}$ be any $m \times n$ matrix. The nullspace of A is equal to the* ***orthogonal complement*** *of the rowspace of A, that is, the set of all vectors that are orthogonal to vectors in the nullspace of A.*

Proof. If x is in the nullspace of A, then $Ax = 0$, so that x is orthogonal to each of the rows of A. □

Our first results cover the simple cases of homogeneous linear equations which are underdetermined and exactly-determined.

Theorem 3.3.14. *If* $\mathbf{A}$ *is an* $m \times n$ *matrix with* $m < n$*, then the homogeneous system of linear equations* $\mathbf{Ax} = \mathbf{0}$ *has a non-trivial solution. We call this an* ***under-determined linear system.***

Proof. Note that $\text{rank}\mathbf{A} \leq \min\{\mathbf{m}, \mathbf{n}\} = \mathbf{m} < \mathbf{n}$. Then the columns of $\mathbf{A}$ are linearly dependent, and the result follows. □

Theorem 3.3.15. *If* $\mathbf{A}$ *is an* $n \times n$ *matrix then* $\mathbf{Ax} = \mathbf{0}$ *has only the trivial solution if and only if* $\mathbf{A}$ *has rank* n*. We call this an* ***exactly-determined linear system.***

In general, we have the following result.

Theorem 3.3.16 (Rouché-Capelli Theorem)**.** *A system of linear equations with* n *variables has a solution if and only if the rank of its coefficient matrix* $\mathbf{A}$ *is equal to the rank of its augmented matrix* $[\mathbf{A}|\mathbf{b}]$*, obtained by appending the columns* $\mathbf{b}$ *to the right side of the matrix. Moreover, the solutions form a linear subspace of* K^n *of dimension* $n - \text{rank}(\mathbf{A})$.

Now, let us return to the question of identifying the solutions to the equation. One approach would be to use elementary algebra techniques to manipulate the equations in order to identify a solution. This underpins the method of solution called Gaussian elimination.

Definition 3.3.17. The three elementary row operations on an $m \times n$ matrix $\mathbf{A}$ over the field $\mathbb{R}$ are:

(a) multiplication of one row of $\mathbf{A}$ by a non-zero scalar c;

(b) replacement of the rth row $\mathbf{a_r}$ of $\mathbf{A}$ by row $\mathbf{a_r}$ plus c times row $\mathbf{a_s}$;

(c) interchange of two rows of $\mathbf{A}$.

Definition 3.3.18. If $\mathbf{A}$ and $\mathbf{B}$ are $m \times n$ matrices, we say that $\mathbf{B}$ is **row-equivalent to A** if $\mathbf{B}$ can be obtained from $\mathbf{A}$ by a finite sequence of elementary row operations.

Theorem 3.3.19. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are row-equivalent* $m \times n$ *matrices, the homogeneous systems of linear equations* $\mathbf{Ax} = \mathbf{0}$ *and* $\mathbf{Bx} = \mathbf{0}$ *have the exact same solutions.*

Proof. Suppose we pass from $\mathbf{A}$ to $\mathbf{B}$ by a finite sequence of elementary row operations:

$$\mathbf{A} = \mathbf{A_0} \to \mathbf{A_1} \to \ldots \to \mathbf{A_k} = \mathbf{B}$$

It suffices to prove that the systems $\mathbf{A_j x} = \mathbf{0}$ and $\mathbf{A_{j+1} x} = \mathbf{0}$ have the same solutions, i.e. that one elementary row operation does not disturb the set of solutions. Observe that no matter which of the three types of the operation is, each equation in the system $\mathbf{A_{j+1} x} = \mathbf{0}$ will be a linear combination of the equations in the system $\mathbf{A_j x} = \mathbf{0}$. Also, an inverse of an elementary row operation is an elementary row operation, so each equation in $\mathbf{A_j x} = \mathbf{0}$ is a linear combination of the equations in $\mathbf{A_{j+1} x} = \mathbf{0}$. Hence these two systems have the same solutions. □

Example. Consider the following system of equations:

$$3x_1 + 2x_2 = 8$$
$$2x_1 + 3x_2 = 7$$

The above system can be written as

$$\begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \end{bmatrix}$$

One way to proceed is to write the augmented matrix of the system:

$$\left[\begin{array}{cc|c} 3 & 2 & 8 \\ 2 & 3 & 7 \end{array}\right]$$

We can use elementary row operations to transform the augmented coefficient matrix until we obtain the identity matrix on the left. Gaussian elimination proceeds left to right, first normalizing an element on the main diagonal to 1, then eliminating all coefficients above and below that main diagonal entry before moving to the next main diagonal entry. The final result should take the form $[I|x]$. We show the sequence of elementary row operations below.

$$\left[\begin{array}{cc|c} 3 & 2 & 8 \\ 2 & 3 & 7 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 2/3 & 8/3 \\ 2 & 3 & 7 \end{array}\right]$$
$$\sim \left[\begin{array}{cc|c} 1 & 2/3 & 8/3 \\ 0 & 5/3 & 5/3 \end{array}\right]$$
$$\sim \left[\begin{array}{cc|c} 1 & 2/3 & 8/3 \\ 0 & 1 & 1 \end{array}\right]$$
$$\sim \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & 1 \end{array}\right]$$

which implies the solution $x_1 = 2$ and $x_2 = 1$. ♣

Exercise 3.4. *Try to solve*

$$\begin{aligned} 2x + 5y &= 9 \\ x + 2y - z &= 3 \\ -3x - 4y + 7z &= 0. \end{aligned}$$

What goes wrong, and why?

Note that the sequence of row operations did not depend on the coefficient matrix. This means that for **square matrices**, if there is a solution to the equation $Ax = b$ for some b, then there is a solution to $Ax = b$ for *all* b. The mapping from b to this solution is itself a linear transformation (check this!), so it must have a matrix representation. This matrix representation is called the inverse matrix of A.

Definition 3.3.20. The **left** and **right inverses** of the $n \times n$ matrix $\mathbf{A}$ are, respectively $n \times n$ matrices $\mathbf{L}$ and $\mathbf{R}$ such that $\mathbf{LA} = \mathbf{I}$ and $\mathbf{AR} = \mathbf{I}$.

Theorem 3.3.21. *If* $\mathbf{A}$ *has both a left and right inverse, then they are unique and equal.*

It is easy to see that invertibility is key to ensuring that the equation $\mathbf{Ax} = \mathbf{b}$ has a unique solution, as it may then be calculated as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Proof. Let $\mathbf{L}$ and $\mathbf{R}$ be a left and right inverse for $\mathbf{A}$, respectively, so $\mathbf{LA} = \mathbf{I} = \mathbf{AR}$. Then

$$\mathbf{L} = \mathbf{LI} = \mathbf{L}(\mathbf{AR}) = (\mathbf{LA})\mathbf{R} = \mathbf{IR} = \mathbf{R}$$

Since every left inverse is a right inverse and vice versa, we have a uniqueness. □

Definition 3.3.22. A square matrix is **invertible** if it has an inverse. A square matrix is **singular** or **degenerate** is it is not invertible.

If matrix $\mathbf{A}$ is invertible, you can use Gaussian elimination on the augmented matrix $[A|I]$ to obtain $[I|A^{-1}]$.

Lemma 3.3.23. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $n \times n$ *invertible matrixes, then* $\mathbf{AB}$ *is invertible with inverse* $\mathbf{B}^{-1}\mathbf{A}^{-1}$.

Definition 3.3.24. Let $\mathbf{A}$ be an $m \times n$ dimensional matrix. The **transpose** $\mathbf{B}$ of matrix $\mathbf{A}$ is the $n \times m$ dimensional matrix such that $a_{ij} = b_{ji}$. We denote the transpose of $\mathbf{A}$ as $\mathbf{A}'$ or $\mathbf{A}'$.

Lemma 3.3.25. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, $(\mathbf{A}')' = \mathbf{A}$, $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$, $(c\mathbf{A})' = \mathbf{cA}'$.

Lemma 3.3.26. *The transpose* $\mathbf{A}'$ *of* $\mathbf{A} \in \mathbb{R}^{\mathbf{m}\times\mathbf{m}}$ *is the* adjoint *of* $\mathbf{A}$ *under the inner product, that is*

$$\langle \mathbf{Ax}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{A}'\mathbf{y} \rangle.$$

If $\mathbf{A} \in \mathbb{C}^{\mathbf{m}\times\mathbf{m}}$, *then the adjoint is the* conjugate transpose *obtained by transposing the matrix and taking the complex conjugate of each entry (i.e.,* $a + ib \mapsto a - ib$*).*

Lemma 3.3.27. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$, *provided that* $\mathbf{A}$ *has an inverse.*

Proof. $\mathbf{AA}^{-1} = \mathbf{I} \Rightarrow (\mathbf{AA}^{-1})' = \mathbf{I} \Rightarrow (\mathbf{A}^{-1})'\mathbf{A}' = \mathbf{I}$. □

Theorem 3.3.28. *An* $n \times n$ *(real) matrix has an inverse if and only if it has rank* n.

Proof. Let $\mathbf{A}$ be an $n \times n$ matrix. Suppose $\mathbf{A}$ has rank n. Then the columns of $\mathbf{A}$ span $\mathbb{R}^n$, and hence for any $\mathbf{i}^j = \mathbf{e}_j$, $j = 1, \ldots, n$ there is a unique vector $\mathbf{x}^j$ such that $\mathbf{Ax}^j = \mathbf{i}^j$. But then $\mathbf{A}^{-1} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$. If $\mathbf{A}$ has rank strictly less than n, then $\text{span}\{\mathbf{a}^1, \ldots, \mathbf{a}^n\} \subset \text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_\mathbf{n}\}$, so there exists $j \in \{1, \ldots, n\}$ such that $\mathbf{Ax} \neq \mathbf{e}_\mathrm{j}$ for any $\mathbf{x} \in \mathbb{R}^\mathbf{n}$. Therefore, $\mathbf{A}$ does not have an inverse. □

Let us now return to the undetermined case of a non-square matrix, A with dimensions $m \times n$ with $m < n$. It is clear that it is not possible to identify an inverse matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$, merely since a single matrix cannot be conformable for both left and right multiplication if $m \neq n$.

However, if A has full row rank, i.e. rank m, then there by the Rouché-Capelli Theorem, $Ax = b$ must have an $(n - m)$-dimensional subspace of solutions and A must be row-reducible to an augmented matrix with an $m \times m$ identity matrix on the left (and other columns to the right). But then, we can obtain an $n \times m$ matrix $A^\dagger$ with $AA^\dagger = I_m$. We call such a matrix a **right pseudoinverse**. Pseudoinverses are typically not unique.

We will show that the $m \times n$ matrix $A^\dagger = A'(AA')^{-1}$ works as a pseudoinverse. First, we show that AA' is invertible. To see this, we note that A has linearly independent rows, so that $A'x = 0$ has only the trivial solution $x = 0$. But then suppose that AA' was non-invertible. Then there exists a non-trivial solution to $AA'x = 0$. But this implies $x'AA'x = 0$, which implies $(A'x)'(A'x) = 0$ or $||A'x||^2 = 0$, which implies that $A'x = 0$, a contradiction. Finally, we note that $AA^\dagger = AA'(AA')^{-1} = I_m$.

Similarly, if A has full column rank, i.e. rank n, then $A^\dagger = (A'A)^{-1}A'$ is a **left pseudoinverse**.

Those of you who have taken econometrics before may recognize this expression as the **ordinary least squares** projection operator. We will derive this now.

Let X be a matrix of k observations of n covariates (plus a constant term), which we assume has *full column rank* (here rank $n + 1$):

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{bmatrix},$$

and Y be a vector of outcome variables,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}.$$

The ordinary least squares problem is to identify a set of weights

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

which solves

$$\min_{\beta \in \mathbb{R}^n} \|Y - X\beta\|^2.$$

Let β^* be the minimizer of this expression, the OLS estimator, and let $Z^* = X\beta^*$. Note that Z^* is in the column space of X.

The claim now is that $Y - Z^*$ (the vector of residuals) must be orthogonal to every column of X. To see this, note that by the fundamental theorem of linear algebra, we may write all vectors x in $\mathbb{R}^k$ as the sum of a component x_n in the column space of X and a component x_o orthogonal to the column space of X. So let, $Y = Y_n + Y_0$. Then for any point Z in the column space of X, we have that

$$\|Y - Z\|^2 = \|Y - Y_n\|^2 + \|Y_n - Z\|^2.$$

Since $\|Y_n - Z\|^2 \geq 0$, we see that Y_n is the unique point in the column space of X such that the OLS objective is minimized, and thus $Y - Z^* = Y_0$, and so is orthogonal to the column space of X.

We now must solve the equation $Y_n = X\beta$ to obtain the OLS estimator. If X has full column rank, there is a unique solution β^*. Since $Y - Z^* = Y - X\beta^*$ is orthogonal to the column space of X, we must have $X'(Y - X\beta^*) = 0$. But this implies that $X'Y = X'X\beta^*$, so that $\beta^* = (X'X)^{-1}X'Y = X^\dagger Y$.

Note that if X does not have full column rank, all of the steps above are valid, except that any β solving $Y_n = X\beta$ now constitutes **an** OLS estimator and $\beta^* = (X'X)^{-1}X'Y$ is the OLS estimator with the lowest norm (all other β are equal to β^* plus a component in the nullspace of X).

Now, before I said that the pseudoinverse was a “projection” operator. Hopefully you can see the analogy of the above discussion with scalar and vector projection interpretations of the inner product (think of Y_n as the vector projection and β^* as the scalar projection). Here is what we mean formally:

Definition 3.3.29. Consider a vector space V over field K.

(a) A **projection** on vector space V is a linear operator $T : V \to V$ such that $T(T(x)) = T(x)$ for all $x \in V$.

(b) A $n \times n$ matrix M is **idempotent** if $\mathbf{M} = \mathbf{M}^2$. Matrices representing a projection are idempotent.

You can check that $X(X'X)^{-1}X'$ is a projection operator, so that Y_n can be thought of as an "OLS" projection of Y on to X.

Exercise 3.5. *Write P_X for the OLS projection operator and $M_X = I - P_X$.*

(a) Check that P_X and M_X are projection operators.

(b) Show that $P_X y$ is in the column space of X and $M_X y$ is in its orthogonal complement (that is, the set of vectors satisfying $X'y = 0$).

(c) Show Pythagoras' Theorem:

$$\|y\|_2^2 = \|P_X y\|_2^2 + \|M_X y\|_2^2.$$

Argue that P_X and M_X shorten vectors.

(d) Let v be in the column space of X. Show that $P_X v = v$ and $M_X v = 0$. Let w be in its orthogonal complement. Show that $P_X w = 0$ and $M_X w = w$.

(e) Suppose there are two groups of regressors, so you partition X and $\hat{\beta}$ as follows:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \quad \text{and} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

Observe that you can now write

$$y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{u}.$$

Suppose X_1 is a group of regressors that you don't care much about. You would like to obtain a closed-form expression for $\hat{\beta}_2$ so that you can analyze it more carefully. How would you do it?

(f) Now consider the following problem:

$$\min_{\gamma} \|M_1 y - M_1 X_2 \gamma\|_2\,,$$

where $M_1 = X_1\left(X_1'X_1\right)^{-1}X_1'$. Derive an expression for the solution $\hat{\gamma}$ and the vector of *residuals. What is the significance of what you find? This result is called the* ***Frisch-Waugh-Lovell Theorem****.*

Exercise 3.6. *Check that the space of real-valued matrices* $\mathbb{R}^{m\times n}$ *satisfies the definition of a vector space (under the usual definitions of matrix addition and scalar multiplication). Give a basis and verify its dimension is mn.*

Check that the Frobenius inner product

$$\langle A, B\rangle_F = \mathrm{tr}(A'B),$$

defines an inner product on the space, resulting in the ***Frobenius norm***

$$\|A\|_F = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2.$$

As an aside, there are many more norms that may be chosen for this vector space, e.g. the operator norm, $\sup_{x:x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$*, and various norms based on the matrix's eigenvalues.*

4

Eigendecompositions and SVD

Contents

4.1 Eigenvalues

Definition 4.1.1. Let M be an $n \times n$ real matrix. A real number λ is an **eigenvalue** of M if there is a nonzero vector $x \in V$ such that $Mx = \lambda x$. The vector x is called a **(right) eigenvector** of M associated with λ. (Note that the vector **0** is by definition not an eigenvector of M).

Here are some observations:

- If M has an eigenvalue λ with eigenvector x, the transformation "stretches" the space by a factor λ in the direction x.
- While the vector **0** is never an eigenvector, the scalar 0 may be an eigenvalue.
- There are real matrices with no real eigenvalues.
- The identity matrix has an eigenvalue 1 associated with every nonzero vector.
- There is a unique eigenvalue associated with each eigenvector: If $M\mathbf{x} = \lambda\mathbf{x}$ and $M\mathbf{x} = \alpha\mathbf{x}$, then $\alpha\mathbf{x} = \lambda\mathbf{x}$, so $\alpha = \lambda$.
- On the other hand, one eigenvalue can be associated with many eigenvectors. The span of

the set of eigenvectors associated with the eigenvalue λ is called the **eigenspace** of M corresponding to λ. The dimension of the eigenspace is called the **geometric multiplicity** of λ.

Theorem 4.1.2. *Let* $\mathbf{x_1}, ..., \mathbf{x_n}$ *be eigenvectors associated with distinct eigenvalues* $\lambda_1, ..., \lambda_n$*. Then the vectors* $\mathbf{x_1}, ..., \mathbf{x_n}$ *are independent.*

Proof. The proof is by induction on n. The case $n = 1$ is trivial, since by definition eigenvectors are nonzero. Now consider $n > 1$ and suppose that the result is true for $n - 1$. Now let

$$\sum_{i=1}^{n} a_i \mathbf{x_i} = \mathbf{0}$$

Applying the matrix M on both sides gives

$$\sum_{i=1}^{n} a_i \lambda_i \mathbf{x_i} = \mathbf{0}$$

Since $\sum_{i=1}^{n} a_i \mathbf{x_i} = \mathbf{0}$, so is $\sum_{i=1}^{n} a_i \lambda_n \mathbf{x_i} = \mathbf{0}$, which combined with the equation above gives

$$\sum_{i=1}^{n} a_i(\lambda_i - \lambda_n)\mathbf{x_i} = \sum_{\mathbf{i=1}}^{\mathbf{n-1}} \mathbf{a_i}(\lambda_\mathbf{i} - \lambda_\mathbf{n})\mathbf{x_i} = \mathbf{0}$$

Since $\mathbf{x_1}, ..., \mathbf{x_{n-1}}$ are independent by the induction hypothesis, it follows that $a_i(\lambda_i - \lambda_n) = 0$ for $i = 1, ..., n - 1$. Since the eigenvalues are distinct, this implies that $a_i = 0$ for $i = 1, ..., n - 1$. Thus, the the above equation reduces to $a_n \mathbf{x_n} = \mathbf{0}$, which implies $a_n = 0$. This shows that $\mathbf{x_1}, ..., \mathbf{x_n}$ are independent. □

Corollary 4.1.3. *A* $n \times n$ *matrix* M *has at most* n *distinct eigenvalues. If it has* n*, then the space has a basis made up of eigenvectors.*

Theorem 4.1.4. *If* M *is idempotent, then each of its eigenvalues is either* 0 *or* 1*.*

Proof. Suppose $M\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{x} \neq 0$. Since M is idempotent, we have $\lambda\mathbf{x} = M\mathbf{x} = M^2\mathbf{x} = M(M\mathbf{x}) = M\lambda\mathbf{x} = \lambda^2\mathbf{x}$. Since $\mathbf{x} \neq 0$, this implies that $\lambda = \lambda^2$, so $\lambda = 0$ or $\lambda = 1$. □

Theorem 4.1.5. *Let M be a symmetric $n \times n$ real matrix. Then $\mathbb{R}^n$ has an orthonormal basis consisting of eigenvectors of M.*

Exercise 4.1. *A* ***row-stochastic*** *matrix is a square matrix with nonnegative entries such that the sum of each row is one.*

(a) Show that 1 is always an eigenvalue of the matrix.

(b) Show that the product of two row-stochastic matrices is row-stochastic.

(c) Use this fact to show that all other eigenvalues of the matrix are smaller in modulus than 1.

For any matrix A, show that A and A' have the same eigenvalues, so that the above results on eigenvalues also hold for column-stochastic matrices.

Exercise 4.2 (Markov Chain: running example)**.** *Consider two cities A and B. Suppose that in each year, each resident of A has a 10% chance of moving to B (and a 90% chance of remaining in A), while each resident of B has a 30% chance of moving to A.*

If initially the cities have the same population, write a matrix equation for the populations in the following year (in terms of the proportion of the populations). The matrix in this equation is called the ***transition matrix****. Check that it is row-stochastic.*

Calculate its eigenvalues and eigenvectors.

4.2 Diagonalization

Definition 4.2.1. Two square matrices $\mathbf{A}$ and $\mathbf{B}$ are called **similar** if there is some nonsingular matrix $\mathbf{C}$ such that $\mathbf{A} = \mathbf{C}\mathbf{B}\mathbf{C}^{-1}$, and equivalently $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$.

Theorem 4.2.2. *Two matrices are similar if and only if they represent the same linear transformation.*

Theorem 4.2.3. *If $\mathbf{A}$ and $\mathbf{B}$ are similar, with $\mathbf{A} = \mathbf{C}\mathbf{B}\mathbf{C}^{-1}$, then λ is an eigenvalue of $\mathbf{A}$ if it is an eigenvalue of $\mathbf{B}$. If x is an eigenvector of $\mathbf{A}$, then $\mathbf{C}^{-1}\mathbf{x}$ is an eigenvector of $\mathbf{B}$.*

Proof. Suppose $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, so $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Since $\mathbf{A} = \mathbf{C}\mathbf{B}\mathbf{C}^{-1}$,

$$\lambda\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{C}\mathbf{B}\mathbf{C}^{-1}\mathbf{x} = \mathbf{C}\mathbf{B}\mathbf{y}$$

Premultiplying by $\mathbf{C}^{-1}$, we have that $\lambda\mathbf{y} = \mathbf{B}\mathbf{y}$. □

Theorem 4.2.4. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are similar, then rank*$\mathbf{A}$ = *rank*$\mathbf{B}$.

Proof. We prove rank$\mathbf{B} \geq$ rank$\mathbf{A}$. Symmetry completes the argument. Let $\mathbf{z_1}, \ldots, \mathbf{z_k}$ be a basis for range of $\mathbf{A}$, $\mathbf{z_i} = \mathbf{A}\mathbf{y_i}$. Put $\mathbf{w_i} = \mathbf{C}^{-1}\mathbf{y_i}$. Then $\mathbf{B}\mathbf{w_i}$'s are independent. To see this, suppose that $\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \alpha_{\mathbf{i}}(\mathbf{B}\mathbf{w_i})$. It needs to be shown that $\alpha_i = 0$ for $i = 1, \ldots, k$. We have that

$$\mathbf{0} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \alpha_{\mathbf{i}}(\mathbf{B}\mathbf{w_i}) = \sum_{\mathbf{i=1}}^{\mathbf{k}} \alpha_{\mathbf{i}}\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\mathbf{C}^{-1}\mathbf{y_i} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \alpha_{\mathbf{i}}\mathbf{C}^{-1}\mathbf{z_i} = \mathbf{C}^{-1}\left(\sum_{\mathbf{i=1}}^{\mathbf{k}} \alpha_{\mathbf{i}}\mathbf{z_i}\right)$$

Since $\mathbf{C}^{-1}$ is nonsingular, this implies that $\sum_{i=1}^{k} \alpha_i\mathbf{z_i} = \mathbf{0}$. Since $\mathbf{z_i}$'s are linearly independent, this implies that $\alpha_i = 0$ for $i = 1, \ldots, k$. □

Definition 4.2.5. A square matrix $\mathbf{X}$ is *orthogonal* if $\mathbf{X}'\mathbf{X} = \mathbf{I}$, or equivalently $\mathbf{X}' = \mathbf{X}^{-1}$.

Theorem 4.2.6. ***(Principal Axis Theorem)*** *Let* $\mathbf{A}$ *be a symmetric* $m \times m$ *real matrix. Let* $\mathbf{x_1}, \ldots, \mathbf{x_m}$ *be an orthonormal basis for* $\mathbb{R}^m$ *made up of eigenvectors of* $\mathbf{A}$*, with corresponding eigenvalues* $\lambda_1, \ldots, \lambda_m$*. Let* $\mathbf{\Lambda} = diag(\lambda_{\mathbf{i}})$ *and let* $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_m}]$*. Then*

$$\begin{aligned} \mathbf{A} &= \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} \\ \mathbf{\Lambda} &= \mathbf{X}^{-1}\mathbf{A}\mathbf{X} \end{aligned}$$

and $\mathbf{X}$ *is orthogonal, that is* $\mathbf{X}^{-1} = \mathbf{X}'$.

Proof. $\mathbf{X}'\mathbf{X} = \mathbf{I}$ by orthonormality, so $\mathbf{X}^{-1} = \mathbf{X}'$. Pick any $\mathbf{z}$ and set $\mathbf{y} = \mathbf{X}^{-1}\mathbf{z}$, so $\mathbf{z} = \mathbf{X}\mathbf{y} = \sum_{\mathbf{j=1}}^{\mathbf{m}} \mathbf{y_j}\mathbf{x_j}$. Then

$$\mathbf{Az} = \sum_{\mathrm{j=1}}^{\mathbf{m}} \mathbf{y_j A x_j} = \sum_{j=1}^{m} y_j(\lambda_j \mathbf{x_j})$$
$$= \mathbf{X\Lambda y}$$
$$= \mathbf{X\Lambda X^{-1} z}$$

Since $\mathbf{z}$ is arbitrary, $\mathbf{A} = \mathbf{X\Lambda X^{-1}}$. □

Theorem 4.2.7 (Rayleigh Quotients). *Let A be a symmetric matrix and $\lambda_{\min}$ and $\lambda_{\max}$ the smallest and largest eigenvalues. Then, we have that*

$$\lambda_{\min}(A) = \min_x \{x'Ax : x'x = 1\}$$
$$\lambda_{\max}(A) = \max_x \{x'Ax : x'x = 1\}\,.$$

Proof. We use the Principal Axis Theorem to prove this result. We show this only for the largest eigenvalue; the proof for the expression for the smallest eigenvalue follows similar lines.

Since $A = X\Lambda X^{-1}$, we have

$$\max_x \{x'Ax : x'x = 1\} = \max_x \left\{x'X\Lambda X^{-1}x : x'x = 1\right\}.$$

Now we can define the new variable $\widetilde{x} = X'x$, so that $x = X\widetilde{x}$, and express the problem as

$$\max_x \{x'Ax : x'x = 1\} = \max_{\widetilde{x}} \{\widetilde{x}'\Lambda\widetilde{x} : \widetilde{x}'\widetilde{x} = 1\}$$
$$= \max_{\widetilde{x}} \left\{\sum_{i=1}^{n} \lambda_i \widetilde{x}_i^2 : \sum_{i=1}^{n} \widetilde{x}_i^2 = 1\right\}.$$

Clearly, the maximum is less than $\lambda_{\max}$. That upper bound is attained, with $\widetilde{x}_i = 1$ for an index i such that $\lambda_i = \lambda_{\max}$, and $\widetilde{x}_j = 0$ for $j \neq i$. This proves the result. This corresponds to setting $x = U\widetilde{x} = u_i$, where u_i is the eigenvector corresponding to $\lambda_i = \lambda_{\max}$. □

Definition 4.2.8. Let $\mathbf{A}$ be an $m \times m$ matrix. The **trace** of $\mathbf{A}$, denoted $\mathrm{tr}\mathbf{A}$ is defined by

$$\mathrm{tr}\mathbf{A} = \sum_{\mathrm{i=1}}^{\mathbf{m}} \mathbf{a_{ii}}$$

Lemma 4.2.9. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $m \times m$ *matrixes. Then*

$$tr(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha tr\mathbf{A} + \beta tr\mathbf{B}$$

$$tr(\mathbf{AB}) = tr(\mathbf{BA})$$

Proof. The linearity is straightforward. For the second claim, observe that

$$\mathrm{tr}\mathbf{AB} = \sum_{i=1}^{m}\left(\sum_{j=1}^{m} a_{ij}b_{ji}\right) = \sum_{j=1}^{m}\left(\sum_{i=1}^{m} b_{ji}a_{ij}\right) = \mathrm{tr}\mathbf{BA}$$

□

Corollary 4.2.10. *If* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{AC}$, *then tr*$\mathbf{B}$ = *tr*$\mathbf{A}$.

Proof.

$$\mathrm{tr}\mathbf{B} = \mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{AC}\right) = \mathrm{tr}\left(\mathbf{ACC}^{-1}\right) = \mathrm{tr}\left(\mathbf{AI}\right) = \mathrm{tr}\mathbf{A}$$

□

Theorem 4.2.11. *The trace is equal to the sum of its eigenvalues.*

Theorem 4.2.12. *If* $\mathbf{A}$ *is symmetric and idempotent, then tr*$\mathbf{A}$ = *rank*$\mathbf{A}$.

Proof. Since $\mathbf{A}$ is symmetric, $\mathbf{A} = \mathbf{XBX}^{-1}$, where $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_m}]$ us an orthogonal matrix whose columns are eigenvectors of $\mathbf{A}$, and $\mathbf{B}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$, which are either 0 or 1. Therefore, tr$\mathbf{B}$ is the number of nonzero eigenvalues of $\mathbf{A}$. Also, rank$\mathbf{B}$ is the number of nonzero diagonal elements. Thus tr$\mathbf{B}$ = rank$\mathbf{B}$. Since $\mathbf{A}$ and $\mathbf{B}$ are similar, tr$\mathbf{A}$ = tr$\mathbf{B}$ = rank$\mathbf{B}$ = rank$\mathbf{A}$. □

Exercise 4.3. *Diagonalize the transition matrix in Problem 4.2, and calculate the proportions of population living in each city after fifty years. As* $t \longrightarrow \infty$, *what is the limiting proportions of population in each city?*

4.3 Principal components analysis

In this section, let X be an $k \times n$ data set, where each row corresponds to an observation of n covariates.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \dots & x_{kn} \end{bmatrix},$$

One question we might ask is what column—or linear combination of columns—explains most of the variance in the data. This might be thought of as the most important dimension of variation in the data.

Definition 4.3.1. The **variance maximization problem** for X solves

$$\max_{u \in \mathbb{R}^n : \|u\|=1} u' \Sigma u,$$

where

$$\Sigma = \frac{1}{n} \sum_{i=1}^{k} (x_i - \overline{x})(x_i - \overline{x})'$$

is the covariance matrix of the data and $\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i$. The solution to the variance maximization problem u^* is the **first principal component** of X.

We have encountered the variance maximization problem before: it was the problem that defined the Rayleigh quotient. We thus have the following theorem.

Theorem 4.3.2. *The first principal component of X is the normalized eigenvector corresponding to the largest eigenvalue of its covariance matrix Σ.*

There is no need to stop at the first principal component. The data can then be projected onto the hyperplane orthogonal to the first principal component, and the first principal component of this projected data is the **second principal component**, and so on. It may not surprise you that the second principal component is just the normalized eigenvector corresponding to the second-largest eigenvalue and so on.

Principal components analysis is used to identify a lower-dimensional approximation of high-dimensional data. A natural question is how much of the variance is explained by the first j principal components. We now show how this question relates to the eigenvalue decomposition of the matrix Σ.

Note that the total variance of the data set X is just the sum of the entries on the diagonals of the covariance matrix Σ. But this is just $\mathrm{tr}(\Sigma)$. But this is just the eigenvalues of Σ.

On the other hand, if we project the data on a two-dimensional plane corresponding to the first two eigenvectors v_1 and v_2, we obtain a new covariance matrix $P\Sigma P'$, where $P = [v_1 v_2]'$, and the total variance of this matrix is just $\lambda_1 + \lambda_2$.

Thus, the ratio of the variance explained by the first two principal components is $\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+...+\lambda_n}$

4.4 Quadratic forms

Definition 4.4.1. Let $\mathbf{A}$ be an $n \times n$ symmetric matrix, and let $\mathbf{x}$ be an n-vector. Then $\mathbf{x}'\mathbf{A}\mathbf{x}$ is scalar, and

$$\mathbf{x'Ax} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{a_{ij}x_ix_j}$$

The mapping $Q : \mathbf{x} \rightarrow \mathbf{x}'\mathbf{A}\mathbf{x}$ is the **quadratic form** defined by $\mathbf{A}$.

Definition 4.4.2. A symmetric matrix $\mathbf{A}$ is called:

(a) **positive definite** if $\mathbf{x}'\mathbf{A}\mathbf{x} > \mathbf{0}$ for all nonzero $\mathbf{x}$;

(b) **negative definite** if $\mathbf{x}'\mathbf{A}\mathbf{x} < \mathbf{0}$ for all nonzero $\mathbf{x}$;

(c) **positive semidefinite** if $\mathbf{x}'\mathbf{A}\mathbf{x} \geq \mathbf{0}$ for all $\mathbf{x}$;

(d) **negative semidefinite** if $\mathbf{x}'\mathbf{A}\mathbf{x} \leq \mathbf{0}$ for all $\mathbf{x}$.

Theorem 4.4.3. *The symmetric matrix* $\mathbf{A}$ *is*
(a) positive definite if and only if all its eigenvalues are strictly positive;
(b) negative definite if and only if all its eigenvalues are strictly negative;
(c) positive semidefinite if and only if all its eigenvalues are nonnegative;
(d) negative semidefinite if and only if all its eigenvalues are nonpositive.

Proof. By the Principal Axis Theorem, $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}'$, where $\mathbf{X}$ is an orthogonal matrix with columns that are eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{A}$. Then the quadratic form can be written in term of the diagonal matrix $\mathbf{\Lambda}$:

$$\mathbf{x'Ax = x'X\Lambda X'x = y'\Lambda y} = \sum_{\mathbf{i=1}}^{\mathbf{n}} \lambda_{\mathbf{i}} \mathbf{y}_{\mathbf{i}}^{2}$$

where $y = \mathbf{X'x}$. □

.

Exercise 4.4. *Let P be an invertible linear transformation, and let y be a change of variables related to x by $y = Px$. Show that the quadratic form defined by A on y has the same nature as the quadratic form defined by A on x.*

4.5 Determinants

Definition 4.5.1. Let $F : \mathcal{M}(n, n) \to \mathbb{R}$. We say that F is **multilinear** if

$$F\left(\mathbf{a^1}, \dots, \alpha\mathbf{a^k} + \beta\mathbf{b^k}, \dots, \mathbf{a^n}\right) = \alpha F\left(\mathbf{a^1}, \dots, \mathbf{a^k}, \dots, \mathbf{a^n}\right) + \beta F\left(\mathbf{a^1}, \dots, \mathbf{b^k}, \dots, \mathbf{a^n}\right)$$

We say that F is **alternating** if whenever $\mathbf{a^i} = \mathbf{a^j}$ for some $i, j = 1, \dots, n$, $i \neq j$, we have

$$F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^j}, \dots, \mathbf{a^n}\right) = 0$$

Lemma 4.5.2. *The multilinear function F is alternating if and only if interchanging $\mathbf{a^i}$ and $\mathbf{a^j}$ changes the sign of F, that is*

$$F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^j}, \dots, \mathbf{a^n}\right) = -F\left(\mathbf{a^1}, \dots, \mathbf{a^j}, \dots, \mathbf{a^i} \dots, \mathbf{a^n}\right)$$

Proof. Suppose first that F is alternating. Then

$$
\begin{aligned}
0 &= F\left(\mathbf{a^1}, \dots, \mathbf{a^i} + \mathbf{a^j}, \dots, \mathbf{a^j} + \mathbf{a^i}, \dots, \mathbf{a^n}\right) \\
&= F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^j}, \dots, \mathbf{a^n}\right) + F\left(\mathbf{a^1}, \dots, \mathbf{a^j}, \dots, \mathbf{a^i}, \dots, \mathbf{a^n}\right) \\
&+ F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^i}, \dots, \mathbf{a^n}\right) + F\left(\mathbf{a^1}, \dots, \mathbf{a^j}, \dots, \mathbf{a^j}, \dots, \mathbf{a^n}\right) \\
&= F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^j}, \dots, \mathbf{a^n}\right) + F\left(\mathbf{a^1}, \dots, \mathbf{a^j}, \dots, \mathbf{a^i}, \dots, \mathbf{a^n}\right)
\end{aligned}
$$

Now suppose that interchanging $\mathbf{a^i}$ and $\mathbf{a^j}$ changes the sign of F. Then if $\mathbf{a^i} = \mathbf{a^j}$, so

$$F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^i}, \dots, \mathbf{a^n}\right) = -F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^i}, \dots, \mathbf{a^n}\right)$$

so

$$F\left(\mathbf{a^1}, \dots, \mathbf{a^i}, \dots, \mathbf{a^i}, \dots, \mathbf{a^n}\right) = 0$$

Therefore, F is alternating. □

Definition 4.5.3. A **permutation i** is an ordered list $\mathbf{i} = (\mathbf{i_1}, \dots, \mathbf{i_n})$ of the numbers $1, 2, \dots, n$. The **signature** $\operatorname{sgn}(\mathbf{i})$ of $\mathbf{i}$ is 1 if $\mathbf{i}$ can be obtained from $(1, 2, \dots, n)$ by switching terms an even number of times, and -1 if $\mathbf{i}$ can be obtained from $(1, 2, \dots, n)$ by switching terms an odd number of times.

Theorem 4.5.4. *For any matrix $\mathbf{A} \in \mathcal{M}(\mathbf{n}, \mathbf{n})$, there is a number $det(\mathbf{A})$ such that for any alternating multilinear $F : \mathcal{M}(n, n) \to \mathbb{R}$,*

$$F(\mathbf{A}) = det(\mathbf{A}) \cdot \mathbf{F}(\mathbf{I})$$

Proof. Write

$$F(\mathbf{A}) = \mathbf{F}\left(\sum_{\mathbf{i_1=1}}^{\mathbf{n}} \mathbf{a_{i_1 1} e_{i_1}}, \sum_{\mathbf{i_2=1}}^{\mathbf{n}} \mathbf{a_{i_2 2} e_{i_2}}, \dots, \sum_{\mathbf{i_n=1}}^{\mathbf{n}} \mathbf{a_{i_n n} e_{i_n}}\right)$$

Using the linearity in the first component

$$F(\mathbf{A}) = \sum_{\mathbf{i_1=1}}^{\mathbf{n}} \mathbf{a_{i_1 1} F} \left(\mathbf{e_{i_1}}, \sum_{\mathbf{i_2=1}}^{\mathbf{n}} \mathbf{a_{i_2 2} e_{i_2}}, \ldots, \sum_{\mathbf{i_n=1}}^{\mathbf{n}} \mathbf{a_{i_n n} e_{i_n}} \right)$$

and repeating this for other components leads to

$$F(\mathbf{A}) = \sum_{\mathbf{i_1=1}}^{\mathbf{n}} \sum_{\mathbf{i_2=1}}^{\mathbf{n}} \cdots \sum_{\mathbf{i_n=1}}^{\mathbf{n}} \mathbf{a_{i_1 1} a_{i_2 2}} \cdots \mathbf{a_{i_n n} F} \left(\mathbf{e_{i_1}}, \mathbf{e_{i_2}}, \ldots, \mathbf{e_{i_n}} \right)$$

Now consider $F\left(\mathbf{e_{i_1}}, \mathbf{e_{i_2}}, \ldots, \mathbf{e_{i_n}}\right)$. Since F is alternating, this term is zero unless $i_1, \ldots, i_n$ are distinct.

Since $F(\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}) = \mathbf{F(I)}$, it follows that

$$F(\mathbf{A}) = \det(\mathbf{A})\mathbf{F(I)}$$

where

$$\det(\mathbf{A}) = \sum_{\mathbf{i}} sgn(\mathbf{i}) \cdot \mathbf{a_{i_1 1} a_{i_2 2}} \cdots \mathbf{a_{i_n n}}$$

□

Theorem 4.5.5. *The function*

$$det(\mathbf{A}) = \sum_{\mathbf{i}} sgn(\mathbf{i}) \cdot \mathbf{a_{i_1 1} a_{i_2 2}} \cdots \mathbf{a_{i_n n}}$$

is an alternating multilinear form.

Proof. Observe that in each product $sgn(\mathbf{i}) \cdot a_{i_1 1} a_{i_2 2} \cdots a_{i_n n}$ in the sum there is exactly one element from each row and each column of $\mathbf{A}$. Then it is obvious that $F\left(\mathbf{a^1}, \ldots, \alpha\mathbf{a^i}, \ldots, \mathbf{a^n}\right) = \alpha F\left(\mathbf{a^1}, \ldots, \mathbf{a^i}, \ldots, \mathbf{a^n}\right)$. It is straightforward to verify that

$$F\left(\mathbf{a^1}, \ldots, \mathbf{a^i} + \mathbf{b^i}, \ldots, \mathbf{a^n}\right) = F\left(\mathbf{a^1}, \ldots, \mathbf{a^i}, \ldots, \mathbf{a^n}\right) + F\left(\mathbf{a^1}, \ldots, \mathbf{b^i}, \ldots, \mathbf{a^n}\right)$$

To prove that det is alternating, suppose $\mathbf{a^i} = \mathbf{a^j}$ for $i \neq j$. For any permutation $\mathbf{i}$ with $i_p \neq i_q$ for $p \neq q$, there is exactly one permutation $\mathbf{i'}$ satisfying $i_p = i'_p$ for $p \notin \{i, j\}$, and $i_i = i'_j$ and $i_j = i'_i$.

Observe that $\operatorname{sgn}(\mathbf{i}) = -\operatorname{sgn}(\mathbf{i}')$ as it requires an odd number of interchanges to swat two elements in a list. Hence, we can rewrite

$$\det A = \sum_{\mathbf{i}:\operatorname{sgn}(\mathbf{i})=1} a_{i_1 1} a_{i_2 2} \cdots a_{i_n n} - a_{i'_1 1} a_{i'_2 2} \cdots a_{i'_n n}$$

Since each $a_{i_1 1} a_{i_2 2} \cdots a_{i_n n} - a_{i'_1 1} a_{i'_2 2} \cdots a_{i'_n n} = 0$, it follows that det$\mathbf{A}$, so det is alternating. □

Summing things up:

Corollary 4.5.6. *An alternating multilinear function $F : \mathcal{M}(n,n) \to \mathbb{R}$ is identically zero if and only if $F(\mathbf{I}) = \mathbf{0}$. The determinant is the unique alternating multilinear function $F : \mathcal{M}(n,n) \to \mathbb{R}$ satisfying $F(\mathbf{I}) = \mathbf{1}$. Any other alternating multilinear function $F : \mathcal{M}(n,n) \to \mathbb{R}$ is of the form $F = F(\mathbf{I}) \cdot \det$.*

This characterization of the determinant as the unique alternating multilinear form with $F(I) = 1$ allows us to offer the following geometric interpretation of the determinant.

Definition 4.5.7. A **parallelepiped** of vectors $v_1, v_2, ..., v_n$ is the set

$$\left\{ \sum_{i=1}^{n} \alpha_i v_i \mid \text{for all } i, 0 \le \alpha_i \le 1 \right\}.$$

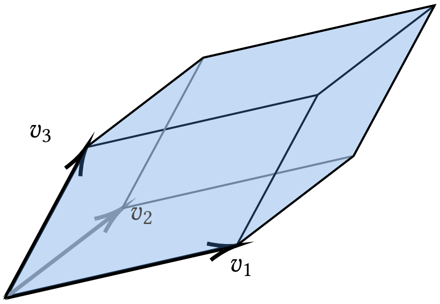


Figure 4.1. Parallelepiped of v_1, v_2 and v_3

Let A be the $n \times n$ matrix with rows $v_1, v_2, ..., v_n$. We argue that $|\det(A)|$ is the volume of the parallelepiped formed by $v_1, v_2, ..., v_n$. To do so, we note that the volume operator on matrices shares the same properties as the determinant. First, $\det(I) = 1$ corresponds to the parallelepiped formed by the standard unit normal vectors, which is a unit cube of volume 1. The alternating property of the determinant corresponds to the parallelepiped having two collinear sides, making it $(n-1)$–dimensional, therefore having zero volume (e.g., think about a parallelogram in three dimensions). The multilinearity of the determinant is more complicated: if one side of the parallelepiped is $\alpha w_i + \beta w_2$, we can consider the two parallelepipeds with the same base (other sides) and with sides w_i and w_2 respectively. Since the volume of a parallelepiped is (base) times (perpendicular height), and the perpendicular height is the scalar projection of the height vector onto the orthogonal complement of the subspace corresponding to the other sides, the multilinearity of the volume follows from the linearity of the inner product.

Here are some other properties of the determinant:

Theorem 4.5.8. *If* $\mathbf{A}'$ *is the transpose of* $\mathbf{A}$*, then* $det\mathbf{A} = det\mathbf{A}'$*.*

Theorem 4.5.9. *Adding a scalar multiple of one column of* $\mathbf{A}$ *to a different column leaves the determinant unchanged. Likewise for rows.*

Proof.

$$\begin{aligned}\det\left(\mathbf{a^1}, \dots, \mathbf{a^j} + \alpha\mathbf{a^k}, \dots, \mathbf{a^k}, \dots, \mathbf{a^n}\right) &= \det\left(\mathbf{a^1}, \dots, \mathbf{a^j}, \dots, \mathbf{a^k}, \dots, \mathbf{a^n}\right) + \alpha\det\left(\mathbf{a^1}, \dots, \mathbf{a^k}, \dots, \mathbf{a^k}, \dots, \mathbf{a^n}\right)\\ &= \det\left(\mathbf{a^1}, \dots, \mathbf{a^j}, \dots, \mathbf{a^k}, \dots, \mathbf{a^n}\right)\end{aligned}$$

The result for rows follows from the previous theorem. □

Theorem 4.5.10. *The determinant of an upper triangular matrix is the product of the diagonal entries.*

Proof. Recall that an upper triangular matrix is one for which $i > j$ implies $a_{ij} = 0$. Observe that the only summand that is nonzero comes from the permutation $(1, 2, \dots, n)$, since for any other permutation there is some j satisfying $i_j > j$. □

Theorem 4.5.11. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *matrices. Then*

$$det\mathbf{AB} = det\mathbf{A} \cdot det\mathbf{B}$$

Proof. If $\mathbf{B}$ is an $n \times n$ matrix and F is an alternating multilinear function, so is $F_{\mathbf{B}}$ defined by

$$F_{\mathbf{B}}(\mathbf{a^1}, \dots, \mathbf{a^n}) = \mathbf{F}(\mathbf{Ba^1}, \dots, \mathbf{Ba^n})$$

Then $F_{\mathbf{B}}(\mathbf{I}) = \mathbf{F}(\mathbf{B}) = \det\mathbf{B} \cdot \mathbf{F}(\mathbf{I})$. Then

$$F_{\mathbf{AB}}(\mathbf{I}) = \mathbf{F}(\mathbf{AB}) = \det\mathbf{AB} \cdot \mathbf{F}(\mathbf{I})$$

On the other hand

$$F_{\mathbf{AB}}(\mathbf{I}) = \mathbf{F_A}(\mathbf{B}) = \det\mathbf{B}\mathbf{F_A}(\mathbf{I}) = \det\mathbf{B}\cdot\mathbf{F}(\mathrm{A}) = \det\mathbf{B}\cdot\det\mathbf{A}\cdot\mathbf{F}(\mathbf{I})$$

Therefore, $\det\mathbf{AB} = \det\mathbf{A}\cdot\det\mathbf{B}$. □

Corollary 4.5.12. *If* $det\mathbf{A} = \mathbf{0}$*, then* $\mathbf{A}$ *has no inverse.*

Proof. Observe that if $\mathbf{A}$ has an inverse, then

$$1 = \det\mathbf{I} = \det\mathbf{A}\cdot\det(\mathbf{A}^{-1})$$

so $\det\mathbf{A} \neq \mathbf{0}$. □

Corollary 4.5.13. *The determinant of an orthogonal matrix is* ±1.

Proof. Recall that $\mathbf{A}$ is orthogonal is $\mathbf{A}'\mathbf{A} = \mathbf{I}$. By the above theorems, we have that $\det(\mathbf{A}')\det(\mathbf{A}) = \mathbf{1}$, and $\det(\mathbf{A}') = \det(\mathbf{A})$, which imply that $(\det\mathbf{A})^2 = \mathbf{1}$. □

Definition 4.5.14. Let $\mathbf{A}$ be an $m \times n$ matrix, and let $0 < k \leq \min\{m, n\}$. A $k \times k$ *minor* of $\mathbf{A}$ is the determinant of a $k \times k$ matrix obtained from $\mathbf{A}$ by deleting $m - k$ rows and $n - k$ columns. If $\mathbf{A}$ is an $n \times n$ matrix, then the minor obtained by deleting the same k rows and columns is called a **principal minor of order k**. Let $m_{i,j}$ denote the minor of a square matrix $\mathbf{A}$ obtained by deleting the i-th row and j-the column from $\mathbf{A}$.
A *cofactor* $c_{i,j}$ of a square $n \times n$ matrix $\mathbf{A}$ is the determinant obtained by replacing the j-th column of $\mathbf{A}$ with i-th unit coordinate vector $\mathbf{e_i}$:

$$c_{i,j} = \det\left(\mathbf{a^1}, ..., \mathbf{a^{j-1}}, \mathbf{e_i}, \mathbf{a^{j+1}}, ..., \mathbf{a^n}\right)$$

Lemma 4.5.15. *For any* $n \times n$ *matrix* $\mathbf{A}$*,*

$$c_{i,j} = (-1)^{i+j} m_{i,j}$$

Consequently

$$det\mathbf{A} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (-\mathbf{1})^{\mathbf{i+j}} \mathbf{a_{ij}} \mathbf{m_{i,j}}$$

Similarly

$$\mathit{det}\mathbf{A} = \sum_{\mathbf{j=1}}^{\mathbf{n}} (\mathbf{-1})^{\mathbf{i+j}} \mathbf{a_{ij}} \mathbf{m_{i,j}}$$

Theorem 4.5.16. *For a square matrix* A*, we have that*

$$\mathbf{A} \begin{bmatrix} c_{1,1} & \cdots & c_{n,1} \\ . & & . \\ . & & . \\ c_{1,n} & \cdots & c_{n,n} \end{bmatrix} = (\mathit{det}\mathbf{A})\mathbf{I}$$

*Consequently, if det*A $\neq$ **0***, then*

$$\mathbf{A^{-1}} = \frac{\mathbf{1}}{\mathit{det}\mathbf{A}} \begin{bmatrix} c_{1,1} & \cdots & c_{n,1} \\ . & & . \\ . & & . \\ c_{1,n} & \cdots & c_{n,n} \end{bmatrix}$$

Combined with a theorem above we obtained the following:

Corollary 4.5.17. *A square matrix is invertible if and only if its determinant is nonzero.*

Definition 4.5.18. The characteristic polynomial f of a square matrix **A** is defined by $f(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A})$. Roots of this polynomial are called *characteristic roots* of **A**.

Lemma 4.5.19. *Every eigenvalue of a matrix is a characteristic root, and every real characteristic root is an eigenvalue.*

Proof. To see this, note that if λ is an eigenvalue with eigenvector $\mathbf{x} \neq \mathbf{0}$, then $(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \lambda\mathbf{x} - \mathbf{A}\mathbf{x} = \mathbf{0}$, so $(\lambda\mathbf{I} - \mathbf{A})$ is singular, so $\det(\lambda\mathbf{I} - \mathbf{A}) = \mathbf{0}$. Therefore, λ is a characteristic root of **A**.

Conversely, if $\det(\lambda\mathbf{I} - \mathbf{A}) = \mathbf{0}$, then there is some nonzero **x** with $(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$, or $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. □

□

Lemma 4.5.20. *The determinant of a square matrix is the product of its characteristic roots.*

Proof. Let $\mathbf{A}$ be an $n \times n$ square matrix and let f be its characteristic polynomial. Then $f(0) = \det(-\mathbf{A}) = (-1)^{\mathbf{n}}\det\mathbf{A}$. On the other hand, we can factor f as

$$f(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n)$$

where $\lambda_1, ..., \lambda_n$ are its characteristic roots. Thus, $f(0) = (-1)^n \lambda_1 \cdots \lambda_n$. □

The **algebraic multiplicity** of an eigenvalue λ_0 is the number of times that $(\lambda - \lambda_0)$ divides into the characteristic polynomial.

Exercise 4.5. *(For this problem, I am assuming you know the basics of differentiation, but if you are rusty, don't worry: we will review differentiation in a week or two.)*

Let $f, g : \mathbb{R} \to \mathbb{R}$ be differentiable functions. Show that f and g are linearly independent in $C([0,1])$ if the ***Wronskian determinant***

$$W := \begin{vmatrix} f_1 & f_2 \\ f_1' & f_2' \end{vmatrix},$$

is not identically zero (note that $W : \mathbb{R} \to \mathbb{R}$ is a function, so by "not identically zero", it is okay for the function to take the value of zero at some points in its domain).

4.6 Generalizing Diagonalization

We have the following result related to the Principal Axis Theorem.

Theorem 4.6.1 (Spectral Theorem). *Let A be an $n \times n$ matrix. Then A is diagonalizable—that is, similar to a diagonal matrix—if and only if the algebraic multiplicity of each eigenvalue equals its geometric multiplicity. This is the case if and only if there exists a basis for F^n consisting of eigenvectors of A, in which case the diagonalization may be written $A = PDP^{-1}$, where P is the matrix having these basis vectors as columns and D is the diagonal matrix containing the corresponding eigenvalues.*

There are, however, real matrices that are not diagonalizable. For example, the projection matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is not diagonalizable.

We now discuss two decompositions of matrices that generalize the diagonalization. The first works for all square matrices and is called the Jordan normal form. The second works more generally for $m \times n$ matrices and is called the singular value decomposition.

Definition 4.6.2. A **Jordan block** is a square matrix of the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$$

Theorem 4.6.3 (Jordan normal form). *Every square matrix is similar to a block diagonal matrix*

$$J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{bmatrix}$$

where each block J_i is a Jordan block.

The Jordan normal form of a matrix A is found as follows: first, calculate the characteristic polynomial and the eigenvalues of A. The number of Jordan blocks of eigenvalue λ is the geometric multiplicity of λ. The number of Jordan blocks of eigenvalue λ of size k is equal to

$$2 \times \text{nullity}((A-\lambda I)^k) - \text{nullity}((A-\lambda I)^{k-1}) - \text{nullity}((A-\lambda I)^{k+1}).$$

The columns of the similarity matrix P (so that $J = P^{-1}AP$) are the **generalized eigenvectors**, which solve $(A-\lambda I)^k x = 0$.

Definition 4.6.4. Given an eigenvalue λ, we say that $v_1, v_2, \ldots, v_r$ form a **chain of generalized eigenvectors** of length r if $v_1 \neq 0$ and

$$\begin{aligned} v_{r-1} &= (A-\lambda I)v_r \\ v_{r-2} &= (A-\lambda I)v_{r-1} \\ &\vdots \\ v_1 &= (A-\lambda I)v_2 \\ 0 &= (A-\lambda I)v_1 \end{aligned}$$

The vectors in a chain of generalized eigenvectors are linearly independent. Moreover, for any eigenvalue with algebraic multiplicity k, there exists k generalized eigenvectors associated with eigenvalue λ.

A nice result that can be proven using the Spectral Decomposition (but we will not) is the Cayley-Hamilton Theorem: every square matrix A satisfies its own characteristic polynomial.

We now turn to non-square matrices, and identify a decomposition similar to the eigendecomposition of square matrices, exploiting the fact that $A'A$ is always square and symmetric.

Theorem 4.6.5 (Singular Value Decomposition). *Let $A \in \mathbf{R}^{m\times n}$ be a rank r matrix. There exists orthogonal matrices $U \in \mathbf{R}^{m\times m}, V \in \mathbf{R}^{n\times n}$ and a diagonal matrix $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, such that*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i' = U\widetilde{S}V', \quad \widetilde{S} := \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}.$$

*The positive numbers $\sigma_1 \geq \ldots \geq \sigma_r > 0$ are unique and called the **singular values** of A.*

*The first r columns of U : $u_i, i = 1, \ldots, r$ (resp. V : $v_i, i = 1, \ldots, r$) are called left (resp. right) **singular vectors** of A, and satisfy*

$$Av_i = \sigma_i u_i, \quad u_i'A = \sigma_i v_i, \quad i = 1, \ldots, r.$$

Proof. First note that the $n \times n$ matrix $A'A$ is real and symmetric, and so by the Principal Axis Theorem, it may be diagonalized as $A'A = V\Lambda V'$, with V a $n \times n$ matrix whose columns form an orthonormal basis (that is, $V'V = VV' = I_n$), and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$. Here, r is the rank of $A'A$ (if $r = n$ then there are no trailing zeros in Λ).

Since $A'A$ is positive semi-definite, the λ_j 's are non-negative, and we can define the non-zero quantities $\sigma_j := \sqrt{\lambda_j}, j = 1, \ldots, r$. Note that when $j > r, Av_j = 0$, since then $\|Av_j\|_2^2 = v_j'A'Av_j = \lambda_j v_j'v_j = 0$. The matrix U is constructed by setting

$$u_i = \frac{1}{\sigma_i} Av_i, \quad i = 1, \ldots, r.$$

These m-vectors are unit vectors, and mutually orthogonal, since the v_j 's are eigenvectors of $A'A$.

If $r < m$, we can complete this set of vectors using the Gram-Schmidt procedure to get $u_{r+1}, \ldots, u_m$ in order to form an orthogonal matrix $U := (u_1, \ldots, u_m) \in \mathbf{R}^m$. Let us check that U, V satisfy the conditions of the theorem, by showing that $U'AV' = \widetilde{S} := \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$. We have

$$(U'AV)_{ij} = u_i'Av_j = \begin{cases} \sigma_j u_i'u_j & \text{if } j \leq r \\ 0 & \text{otherwise,} \end{cases}$$

where the second line follows by $Av_j = 0$ when $j > r$. Thus, $U'AV = \widetilde{S}$, as claimed. □

Exercise 4.6. *Calculate the Jordan form of the matrix*

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Calculate an expression for the nth power of the matrix.

Part III

Analysis

5

Metric Spaces, Sequences and Compactness

Contents

5.1 Metric spaces

Definition 5.1.1. A set X together with a real-valued function d is called a **metric space** if for all $x, y, z \in X$, the following properties are satisfied:

(a) $d(x, x) = 0$;

(b) $d(x, y) > 0$ if $x \neq y$;

(c) $d(x, y) = d(y, x)$;

(d) $d(x, y) \leq d(x, z) + d(z, y)$.

The last property is often referred to as the **triangle inequality**. Any function that satisfies these properties is called a **distance function**, or a **metric**. The elements of X are often called **points**.

In normed spaces, metrics can always be defined in such a way that the distance between two points is equal to the distance of their difference from the zero vector, so that

$$d(x, y) = \|x - y\|$$

is the **metric induced by norm** $\| \cdot \|$.

Example. Here are some examples:

(i) Euclidean space $\mathbb{R}^n$. Recall the p of $x \in \mathbb{R}^n$, defined by $||x||_p = \left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}$. Each p-norm induces a metric on $\mathbb{R}^n$. The distance induced by the 2-norm is called the **Euclidean metric** is given by $d(x, y) = ||x - y|| = \left(\sum_{i=1}^n (x_i - y_i)^2\right)^{\frac{1}{2}}$.

(ii) Let $f : X \longrightarrow \mathbb{R}$ be a continuous, bounded function. Let $CB(X)$ denote the space of such functions. The *sup norm* of f is given by $||f|| = \sup_{x \in X} |f(x)|$. The distance function $d(f, g) = ||f - g||$ satisfies properties (a)-(d), turning the space of continuous bounded functions $CB(X)$ into a metric space.

(iii) Any nonempty set X can be trivially made a metric space using the discrete metric, in which $d(x, y) = 1$ if $x \neq y$ while $d(x, x) = 0$.

♣

Note that as the examples suggest, it is possible to define many different metrics on the same space. The 'right' metric to use is sometimes clear, but sometimes proofs proceed more easily under one metric than another. It's worth having this in the back of your mind.

Exercise 5.1. *Let* $S^1 = \{(x_1, x_2) \in \mathbb{R}^2 | x_1^2 + x_2^2 = 1\}$ *be the unit circle and define* $d(x, y)$ *as the length of the shortest arc between* x *and* y*. Check that* (S^1, d) *is a metric space.*

Definition 5.1.2. Let X be a metric space, $S \subseteq X$, and $x, y \in X$.

(a) The r-**neighborhood** of a point x is a set $N_r(x)$ consisting of points y such that $d(x, y) < r$, where $r > 0$ is called the **radius** of $N_r(x)$. A set is a **neighborhood** of x if it contains an r neighborhood of x for some $r > 0$.

(b) A point x is a **limit point** of S if every neighborhood of x contains a point $y \neq x$ such that $y \in S$.

(c) If $x \in S$ and x is not a limit point of S, then x is said to be an **isolated point** of S.

(d) S is **closed** if every limit point of S is a point of S.

(e) A point x is an **interior point** of S if there exists a neighborhood N of x such that $N \subseteq S$. The set of interior points of S is denoted $\operatorname{int}(S)$.

(f) S is **open** if every point of S is an interior point of S.

(g) S is **bounded** if there exists $r > 0$ and a point x such that $S \subseteq N_r(x)$.

(h) Set S is dense in set E if for all $x \in E$, every neighborhood of x intersects with S.

Theorem 5.1.3. *Every r-neighborhood in a metric space X is open.*

Proof. Let $x \in X$ and let $N_r(x)$ be a neighborhood of x. For any point $y \in N_r(x)$, let $h = r - d(x, y)$. If $z \in N_h(y)$, then we have

$$d(x, z) \leq d(x, y) + d(y, z) < r - h + h = r,$$

which implies $z \in N_r(x)$. Hence, $N_h(y) \subseteq N_r(x)$. ■

Exercise 5.2. *Sketch a representative ε-neighborhood, $N_\varepsilon(x)$, in $\mathbb{R}^2$ under the distances induced by the p-norm for $p = 1, 2, 3, \infty$. Describe a ε-neighborhood under the discrete metric.*

Theorem 5.1.4. *Let S be a subset of a metric space X. If x is a limit point of S, then every neighborhood of x contains infinitely many points of S.*

Proof. Suppose there is a neighborhood N of x which contains a finite number of points $y_1, y_2, \ldots, y_n$ of S that are distinct from x. Let $r = \min\{d(x, y_m)\}$ for $1 \leq m \leq n$. Clearly, $r > 0$. The neighborhood $N_r(x)$ contains no point of S that are distinct from x. This implies that x is not a limit point of S. ■

Corollary 5.1.5. *A finite set in a metric space has no limit points.*

Example. Here are some examples and observations:

(i) (a, b) is open subset of $\mathbb{R}^1$, but not of $\mathbb{R}^2$.

(ii) $[a, b]$ is a closed subset of $\mathbb{R}^1$.

(iii) Any finite set is closed.

(iv) Set of all integers is a closed subset of $\mathbb{R}^1$.

(v) $(a, b]$ is neither open nor closed.

♣

Theorem 5.1.6. *A subset S of a metric space is open if and only if its complement is closed.*

Proof. Suppose S is open and let x be a limit point of S^c. Then every neighborhood of x contains a point of S^c, so x is not an interior point of S. Hence, x cannot be a point of S. This implies $x \in S^c$, so S^c is closed.

Conversely, suppose S^c is closed and let $x \in S$. Then x is not in S^c, so x cannot be a limit point of S^c. Thus, there exists a neighborhood N of x that contains no point of S^c. This implies $N \subseteq S$, so S is open. ■

Theorem 5.1.7. *All sets mentioned below are understood to be subsets of a metric space.*

(a) For any collection $\{G_\alpha\}$ of open sets, $\bigcup_\alpha G_\alpha$ is open.

(b) For any collection $\{F_\alpha\}$ of closed sets, $\bigcap_\alpha F_\alpha$ is closed.

(c) For any finite collection $G_1, \ldots, G_n$ of open sets, $\bigcap_{i=1}^n G_i$ is open.

(d) For any finite collection $F_1, \ldots, F_n$ of closed sets, $\bigcup_{i=1}^n F_i$ is closed.

Proof. Let $x \in \bigcup_\alpha G_\alpha$. Then x is an interior point of G_α for some α, and hence of the union. This proves (a).

By De Morgan's law, $(\bigcap_\alpha F_\alpha)^c = \bigcup_\alpha F_\alpha^c$. By Theorem 5.1.6, each F_α^c is open. Hence, (b) follows from (a).

Let $x \in \bigcap_{i=1}^n G_i$. Then there exists r_i such that $N_{r_i}(x) \subseteq G_i$ for $i = 1, 2, \ldots, n$. Let $r = \min(r_1, r_2, \ldots, r_n)$. Then $N_r(x) \subseteq G_i$ for all $i = 1, 2, \ldots, n$, so $N_r(x) \subseteq \bigcap_{i=1}^n G_i$. This proves (c).

Taking complements, $\left(\bigcup_{i=1}^n F_i\right)^c = \bigcap_{i=1}^n (F_i^c)$. Thus, (d) follows from (c). ■

The finiteness in (c) and (d) is essential. To see that, let $G_n = \left(-\frac{1}{n}, \frac{1}{n}\right)$. Then $G = \cap_{n=1}^\infty G_n = \{0\}$ which is not an open subset of $\mathbb{R}^1$. Let $F_n = \left[\frac{1}{n}, 2 - \frac{1}{n}\right]$. Then $F = \cup_{n=1}^\infty F_n = (0, 2)$, which is not closed.

Definition 5.1.8. Let X be a metric space and let $E \subseteq X$. If E' denotes the set of limit points of E in X, then the **closure** of E is the set $\bar{E} = E \cup E'$.

Theorem 5.1.9. *Let E be a subset of a metric space X. Then*

(a) $\bar{E}$ *is closed;*

(b) $E = \bar{E}$ *if and only if E is closed;*

(c) $\bar{E} \subseteq F$ *for every closed set $F \subseteq X$ such that $E \subseteq F$.*

Proof. (a) We show that $\bar{E}^c$ is open. Let $x \in \bar{E}^c$. Then x is neither a point of E nor a limit point of E. Thus x has a neighborhood N such that $N \subseteq E^c$. But since N is open, every point of N is also an interior point of N, and therefore an interior point of E^c, and consequently cannot be a limit point of E. This means that $N \subseteq (E')^c$, so $N \subseteq \bar{E}^c$. Thus, $\bar{E}^c$ is open.

(b) If $E = \bar{E}$, then E is closed by (a). If E is closed, then $E' \subseteq E$, so $\bar{E} \subseteq E$. Hence $E = \bar{E}$.

(c) Let $F \subseteq X$ be a closed set such that $E \subseteq F$. If $x \in E'$, then every neighbourhood of x contains some point of E, and hence of F. Thus, $x \in F'$. But F is closed, so $F' \subseteq F$, and hence $x \in F$. Therefore, $E' \subseteq F$, and $\bar{E} \subseteq F$.

■

Theorem 5.1.10. *Let E be a nonempty set of real numbers which is bounded above. Let $y = \sup E$. Then $y \in \bar{E}$. Hence $y \in E$ if E is closed.*

Proof. If $y \in E$, then $y \in \bar{E}$, so assume $y \notin E$. Since y is the supremum, for every $\varepsilon > 0$ there exists some $x \in E$ such that $y - \varepsilon < x < y$. But this means y is a limit point of E. Hence $y \in \bar{E}$. ■

Exercise 5.3. *Let S and E be subsets of a metric space X and $S \subseteq E$. Establish the following equivalences:*

(i) *S is dense in E.*

(ii) $\operatorname{int}(X \setminus S) = \emptyset$.

(iii) $\overline{S} = E$.

5.2 Sequences

Definition 5.2.1. A **sequence** is a function defined on the set of positive integers $\mathbb{N}$. If $f(n) = p_n$ for $n \in \mathbb{N}$, we often denote the sequence by $\{p_n\}$ or by $p_1, p_2, p_3, \ldots$. The elements p_n are called the terms of the sequence. If S is a set and $p_n \in S$ for all n, then we say that $\{p_n\}$ is a sequence in S. A sequence is said to be **bounded** if its range is bounded.

5.2.1 Convergent sequences

Definition 5.2.2. A sequence $\{p_n\}$ in a metric space X is said to **converge** if there exists a point $p \in X$ such that, for every $\varepsilon > 0$, there exists a positive integer N such that $n \geq N$ implies $d(p_n, p) < \varepsilon$. In this case, we may also say that $\{p_n\}$ converges to p, or that p is the limit of $\{p_n\}$, and write $p_n \longrightarrow p$, or $\lim_{n\to\infty} p_n = p$. If a sequence does not converge, it is said to **diverge**.

Whether a sequence converges depends not only on $\{p_n\}$ but also on X. For instance, $\{1/n\}$ converges in $\mathbb{R}$ to 0, but fails to converges in the set of all positive real numbers, $\mathbb{R}_{++} = (0, \infty)$

Example. Here are some examples in $\mathbb{R}$:

(i) If $p_n = \frac{1}{n}$, the sequence is convergent with $\lim_{n\to\infty} p_n = 0$. The range is infinite, and the sequence is bounded.

(ii) If $p_n = n^2$, the sequence $\{p_n\}$ is unbounded, divergent, and has infinite range.

(iii) If $p_n = 1$, then $\{p_n\}$ converges to 1, is bounded, and has finite range.

♣

Theorem 5.2.3. *Let $\{p_n\}$ be a sequence in a metric space X. Then we have:*

(a) $\{p_n\}$ converges to p if and only if every neighborhood of p contains all but finitely many of the terms of $\{p_n\}$.

(b) If $p_n \longrightarrow p$ and $p_n \longrightarrow p'$, then $p = p'$.

(c) If $\{p_n\}$ converges, then $\{p_n\}$ is bounded.

Proof. (a) Suppose $p_n \longrightarrow p$ and let N_ε be a neighborhood of p with given radius $\varepsilon > 0$. For any $y \in X$, $d(y, p) < \varepsilon$ implies $y \in N_\varepsilon$. Corresponding to this ε, there exists N such that $n \geq N$ implies $d(p_n, p) < \varepsilon$. Thus, $p_n \in N_\varepsilon$ whenever $n \geq N$.

Conversely, suppose every neighborhood of x contains all but finitely many of the terms of $\{p_n\}$. Given $\varepsilon > 0$, let N_ε be a neighborhood of p with radius ε. Since N_ε contains all but finitely many of the p_n, there exists N such that $n \geq N$ implies $p_n \in N_\varepsilon$. Hence, we have $d(p_n, p) < \varepsilon$ whenever $n \geq N$, so $p_n \longrightarrow p$.

(b) Suppose $p_n \longrightarrow p$ and $p_n \longrightarrow p'$. Then for any $\varepsilon > 0$, there exists N and N' such that $n \geq N$ implies $d(p_n, p) < \varepsilon/2$ and $n \geq N'$ implies $d(p_n, p') < \varepsilon/2$. Take $n \geq \max\{N, N'\}$. Then

$$0 \leq d(p, p') \leq d(p, p_n) + d(p_n, p') < \varepsilon.$$

Since ε was arbitrary, $d(p, p') = 0$.

(c) Suppose $p_n \longrightarrow p$. Then there exists N such that $n \geq N$ implies $d(p_n, p) < 1$. Let

$$r = 1 + \max\{d(p_1, p), d(p_2, p), \ldots, d(p_{N-1}, p)\}.$$

Then $d(p_n, p) < r$ for all n. Thus, $p_n \in N_r(p)$ for all n.

■

Theorem 5.2.4. *Let S be a subset of a metric space X. Then $p \in X$ is a limit point of S if and only if there exists a sequence $\{p_n\}$ in S with $p_n \neq p$ that converges to p.*

Proof. Suppose p is a limit point of S. Then for every n, there exists a point $p_n \in S$ distinct from p such that $d(p_n, p) < 1/n$. Given $\varepsilon > 0$, there exists a positive integer N such that $N\varepsilon > 1$. It follows that $d(p_n, p) < \varepsilon$ whenever $n \geq N$. Hence, $p_n \longrightarrow p$.

Conversely, let $\{p_n\}$ be a sequence in S such that $p_n \longrightarrow p$ and $p_n \neq p$. Then for any $\varepsilon > 0$, there is N such that $n \geq N$ implies $d(p_n, p) < \varepsilon$. Thus, $p_n \in N_\varepsilon(p)$ whenever $n \geq N$. ■

Theorem 5.2.5. *A subset S of a metric space X is closed if and only if every convergent sequence in S has its limit in S.*

Proof. Suppose S is closed and let $\{p_n\}$ be a sequence in S with limit p. For any $\varepsilon > 0$, there is N such that $n \geq N$ implies $d(p_n, p) < \varepsilon$. If $p_n \neq p$, then p is a limit point of S and hence is a point of S. If $p_n = p$ for some n, $p \in S$ since $\{p_n\}$ is a sequence in S.

Conversely, suppose every convergent sequence in S has its limit in S. Let p be a limit point of S. By Theorem 5.2.4, there exists a sequence $\{p_n\}$ in S that converges to p. This implies p is a point of S, so S is closed. ■

Definition 5.2.6. Let $\{p_n\}$ be a sequence in a metric space X and let $\{n_i\}$ be a sequence in $\mathbb{N}$ with the property that $n_1 < n_2 < n_3 < \cdots$. Then the sequence $\{p_{n_i}\}$ is said to be a **subsequence** of $\{p_n\}$. If $\{p_{n_i}\}$ converges, its limit is called a subsequential limit of $\{p_n\}$.

Theorem 5.2.7. *Let $\{p_n\}$ be a sequence in a metric space X. Then $\{p_n\}$ converges to p if and only if every subsequence of $\{p_n\}$ converges to p.*

Proof. Suppose $p_n \longrightarrow p$ and let $\{p_{n_i}\}$ be a subsequence of $\{p_n\}$. Given $\varepsilon > 0$, there exists N such that $n \geq N$ implies $d(p_n, p) < \varepsilon$. Let k be the smallest integer such that $n_k \geq N$. Then we have $d(p_{n_i}, p) < \varepsilon$ whenever $i \geq k$. Thus, $\{p_{n_i}\}$ converges. The converse holds trivially since $\{p_n\}$ is a subsequence of itself. ■

Exercise 5.4. *Let $f_n \in B([0, 1])$ be defined as $f_n(t) = t^n$ for $n = 1, 2, \ldots$.*

(a) Let $x \in [0, 1]$. Does the sequence $(f_n(x))$ converge in the Euclidean metric?

(b) Does the sequence (f_n) converge in the sup-metric on $B([0, 1])$?

5.2.2 Cauchy sequences

Definition 5.2.8. A sequence $\{p_n\}$ in a metric space X is said to be a **Cauchy sequence** if for every $\varepsilon > 0$, there exists a positive integer N such that $d(p_n, p_m) < \varepsilon$ whenever $n \geq N$ and $m \geq N$.

Theorem 5.2.9. *Let $\{p_n\}$ be a sequence in a metric space X. Then we have:*

(a) If $\{p_n\}$ converges, then $\{p_n\}$ is a Cauchy sequence.

(b) If $\{p_n\}$ is a Cauchy sequence and a subsequence $\{p_{n_i}\}$ of $\{p_n\}$ converges to p, then $\{p_n\}$ also converges to p.

(c) If $\{p_n\}$ is a Cauchy sequence, then $\{p_n\}$ is bounded.

Proof. (a) Suppose $p_n \longrightarrow p$ and let $\varepsilon > 0$. There exists N such that $n \geq N$ implies $d(p_n, p) < \varepsilon/2$. It follows that

$$d(p_n, p_m) \leq d(p_n, p) + d(p, p_m) < \varepsilon$$

whenever $n \geq N$ and $m \geq N$. Hence, $\{p_n\}$ is a Cauchy sequence.

(b) Since $\{p_n\}$ is a Cauchy sequence, there is, for any $\varepsilon > 0$, an integer N_1 such that $n \geq N_1$ and $m \geq N_1$ imply $d(p_n, p_m) < \varepsilon/2$. Since $p_{n_i} \to p$, there is N_2 such that $n_i \geq N_2$ implies $d(p_{n_i}, p) < \varepsilon/2$. Let $N = \max\{N_1, N_2\}$. Then for $n \geq N$, we have

$$d(p_n, p) \leq d(p_n, p_{n_i}) + d(p_{n_i}, p) < \varepsilon,$$

for any $n_i \geq N$. Thus, $\{p_n\}$ converges to p.

(c) Since $\{p_n\}$ is a Cauchy sequence, there exists N such that $d(p_n, p_m) < 1$ for $n \geq N$ and $m \geq N$. Let

$$r = 1 + \max\{d(p_1, p_N), d(p_2, p_N), \ldots, d(p_{N-1}, p_N)\}.$$

Then $d(p_n, p_N) < r$ for all n. Hence, $\{p_n\}$ is bounded.

■

Exercise 5.5. *True or false: if for each $\varepsilon > 0$, there is an N such that for all $m \geq N$,* successive terms *satisfy $d(x_m, x_{m+1}) \leq \varepsilon$, then $\{x_n\}$ is Cauchy.*

5.2.3 Real sequences

Definition 5.2.10. A sequence $\{x_n\}$ of real numbers is said to be

(a) **increasing** if $x_n \leq x_{n+1}$ for all n;

(b) **decreasing** if $x_n \geq x_{n+1}$ for all n.

If $\{x_n\}$ is increasing or decreasing, then $\{x_n\}$ is said to be **monotonic**.

Theorem 5.2.11. *A monotonic sequence converges if and only if it is bounded.*

Proof. Let $\{x_n\}$ be a real sequence that is increasing and bounded; the proof for decreasing sequences is analogous. Let A be the range of $\{x_n\}$ and let $x = \sup A$. Given $\varepsilon > 0$, there exists N such that $x_N \in A$ and $x - \varepsilon < x_N \leq x$. Since $\{x_n\}$ increases, $n \geq N$ implies $x - \varepsilon < x_n \leq x$. Hence, $\{x_n\}$ converges to x. The converse follows from Theorem 5.2.3c. ■

Theorem 5.2.12. *Every real sequence contains a monotonic subsequence.*

Proof. Let $\{x_n\}$ be a sequence of real numbers. Let K be the set of positive integers k such that $x_k > x_n$ for all $n > k$. If K is infinite, then the subsequence $\{x_{n_i}\}$ with $n_i \in K$ is decreasing.

Suppose K is finite. If K is nonempty, $\sup K$ exists. Let $N = \sup K$ and let $n_1 = K + 1$. If K is empty, let $n_1 = 1$. In either case, $n_1 \notin K$, so there exists $n_2 > n_1$ such that $x_{n_1} \leq x_{n_2}$. Similarly, there is $n_3 > n_2$ such that $x_{n_2} \leq x_{n_3}$. Continuing the process, we obtain a sequence $\{n_i\}$ such that $\{x_{n_i}\}$ is increasing. ■

Theorem 5.2.13 (Bolzano-Weierstrass). *Every bounded real sequence has a convergent subsequence.*

Proof. Let $\{x_n\}$ be a bounded sequence of real numbers. Then $\{x_n\}$ contains a monotonic subsequence $\{x_{n_i}\}$ that is also bounded. Hence, by Theorem 5.2.11, $\{x_{n_i}\}$ converges. ■

Theorem 5.2.14. *Let $\{\boldsymbol{x}_n\}$ be a sequence in $\mathbb{R}^k$ and $\boldsymbol{x}_n = (x_{1,n}, x_{2,n}, \dots, x_{k,n})$. Then $\{\boldsymbol{x}_n\}$ converges to $\boldsymbol{x} = (x_1, x_2, \dots, x_k)$ if and only if $x_{j,n} \to x_j$ for all $j = 1, 2, \dots, k$.*

Proof. Suppose $\boldsymbol{x}_n \to \boldsymbol{x}$. Since

$$0 \leq \left\| x_{j,n} - x_j \right\| \leq \| \boldsymbol{x}_n - \boldsymbol{x} \|,$$

we have $x_{j,n} \to x_j$ for all $j = 1, 2, \dots, k$.

Conversely, suppose $x_{j,n} \to x_j$ for all $j = 1, 2, \dots, k$. Then for each $\varepsilon > 0$ there is N_j such that $n \geq N_j$ implies

$$\left\| x_{j,n} - x_j \right\| < \frac{\varepsilon}{\sqrt{k}},$$

for $j = 1, 2, \dots, k$. Let $N = \max\{N_1, N_2, \dots, N_k\}$. Then $n \geq N$ implies

$$\| \boldsymbol{x}_n - \boldsymbol{x} \| = \left(\sum_{j=1}^{k} \left\| x_{j,n} - x_j \right\|^2 \right)^{1/2} < \varepsilon.$$

Thus, $\{\boldsymbol{x}_n\}$ converges to $\boldsymbol{x}$. ■

Theorem 5.2.15. *Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Proof. We prove by induction. Theorem 5.2.13 establishes the case for $k = 1$. Next, suppose every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence. Let $\{\boldsymbol{x}_n\}$ be a bounded sequence in $\mathbb{R}^{k+1}$. Write $\boldsymbol{x}_n = (\boldsymbol{y}_n, z_n)$, where $\{\boldsymbol{y}_n\}$ is a sequence in $\mathbb{R}^k$ and $\{z_n\}$ is a sequence in $\mathbb{R}$. By the induction hypothesis, $\{\boldsymbol{y}_n\}$ has a convergent subsequence $\{\boldsymbol{y}_{n_i}\}$. The sequence $\{n_i\}$ so obtained gives a subsequence $\{z_{n_i}\}$ of $\{z_n\}$, which itself contains a convergent subsequence $\{z_{n_{i_j}}\}$. Since $\{\boldsymbol{y}_{n_i}\}$ converges, the subsequence $\{\boldsymbol{y}_{n_{i_j}}\}$ also converges. Hence, by Theorem 5.2.14, the subsequence $\{\boldsymbol{x}_{n_{i_j}}\}$ converges. ■

Theorem 5.2.16. *Suppose $\{p_n\}$ and $\{s_n\}$ are real sequences, and* $\lim_{n\to\infty} p_n = p$, $\lim_{n\to\infty} s_n = s$. *Then*

(a) $\lim_{n\to\infty}(p_n + s_n) = p + s$;

(b) $\lim_{n\to\infty} cp_n = cp$, $\lim_{n\to\infty}(c + p_n) = c + p$ *for any real number c;*

(c) $\lim_{n\to\infty} p_n s_n = ps$;

(d) $\lim_{n\to\infty} \frac{1}{p_n} = \frac{1}{p}$ *provided that* $p_n \neq 0$ $(n = 1, .2, ...)$ *and* $p \neq 0$.

Proof. (a) Given $\varepsilon > 0$, there exist integers N_1, N_2 such that

$$n \geq N_1 \longrightarrow |p_n - p| < \frac{\varepsilon}{2}$$
$$n \geq N_2 \longrightarrow |s_n - s| < \frac{\varepsilon}{2}$$

Let $N = \max(N_1, N_2)$. Then $n \geq N$ implies

$$|(p_n + s_n) - (p + s)| \leq |p_n - p| + |s_n - s| < \varepsilon$$

(b) Trivial.

(c) Observe that

$$p_n s_n - ps = (p_n - p)(s_n - s) + p(s_n - s) + s(p_n - p)$$

Given $\varepsilon > 0$, there exist integers N_1, N_2 such that

$$n \geq N_1 \longrightarrow |p_n - p| < \sqrt{\varepsilon}$$
$$n \geq N_2 \longrightarrow |s_n - s| < \sqrt{\varepsilon}$$

Let $N = \max(N_1, N_2)$. Then $n \geq N$ implies

$$|(p_n - p)(s_n - s)| < \varepsilon$$

and so $\lim_{n\to\infty}(p_n - p)(s_n - s) = 0$. By (a) and (b), $\lim_{n\to\infty} p(s_n - s) = 0$ and $\lim_{n\to\infty} s(p_n - p) = 0$, and hence by (a) $\lim_{n\to\infty}(p_n - p)(s_n - s) + p(s_n - s) + s(p_n - p) = 0$. Therefore $\lim_{n\to\infty}(p_n s_n - ps) = 0$.

(d) Choose m such that $|p_n - p| < \frac{1}{2}|p|$ if $n \geq m$. Then $|p_n - p| \geq |p| - |p_n|$, which implies $|p_n| > \frac{1}{2}|p|$ for $n \geq m$.
Given $\varepsilon > 0$, there is an integer $N > m$ such that $n \geq N$ implies

$$|p_n - p| < \frac{1}{2}|p|^2\varepsilon$$

Hence, for $n \geq N$,

$$\left|\frac{1}{p_n} - \frac{1}{p}\right| = \left|\frac{p_n - p}{p_n p}\right| < \frac{2}{|p|^2}|p_n - p| < \varepsilon$$

■

Definition 5.2.17. Let $\{p_n\}$ be a sequence of real numbers with the following property: For every real M there is an integer N such that $n \geq N$ implies $p_n \geq M$. Then we write $p_n \to \infty$. Similarly, if for every real M there is an integer N such that $n \geq N$ implies $p_n \leq M$, we write $p_n \to -\infty$.

Definition 5.2.18. Let $\{p_n\}$ be a sequence of real numbers. Let E be the set of numbers x (in the extended real number system $\mathbb{R} \cup \{-\infty, +\infty\}$) such that $p_{n_k} \to x$ for some subsequence $\{p_{n_k}\}$. Let $p^* = \sup E$ and $p_* = \inf E$. Equivalently, $\limsup_{n\to\infty} p_n = p^*$ and $\liminf_{n\to\infty} p_n = p_*$.

Corollary 5.2.19. *A real sequence $\{p_n\}$ converges if and only if $\limsup_{n\to\infty} p_n = \liminf_{n\to\infty} p_n$.*

Exercise 5.6 (Squeeze Principle). *Let (x_n), (y_n) and (z_n) be real sequences with $x_n \leq y_n \leq z_n$. Show that if $\lim x_n = \lim z_n = L$, then $\lim y_n = L$.*

5.3 Completeness

Definition 5.3.1. A subset S of a metric space X is said to be **complete** if every Cauchy sequence in S converges in S. A complete normed space is called a **Banach space**, while a complete inner product space is called a **Hilbert space**.

Theorem 5.3.2. *Let S be a subset of a metric space X. Then we have:*

(a) If S is complete, then S is closed.

(b) If S is closed and X is complete, then S is complete.

Proof. (a) Let $\{p_n\}$ be a convergent sequence in S with limit $p \in X$. Since $\{p_n\}$ is also Cauchy and S is complete, $\{p_n\}$ converges in S. Hence, by Theorem 5.2.5, S is closed.

(b) Let $\{p_n\}$ be a Cauchy sequence in S. Since X is complete, $\{p_n\}$ converges to a point $p \in X$. It follows that $p \in S$ since S is closed. Thus, S is complete.

■

Corollary 5.3.3. *A subset S of a complete metric space is complete if and only if it is closed.*

Theorem 5.3.4. *$\mathbb{R}$ is complete with respect to the Euclidean metric.*

Proof. Let $\{x_n\}$ be a Cauchy sequence in $\mathbb{R}$. By Theorem 5.2.9c, $\{x_n\}$ is bounded. Hence, Theorem 5.2.13 implies that there is a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ that converges to some real number x. By Theorem 5.2.9b, $\{x_n\}$ also converges to x. Hence, $\mathbb{R}$ is complete. ■

In fact, this would have been an equivalent way to construct $\mathbb{R}$: as the completion of the rational numbers (the smallest superset of $\mathbb{Q}$ such that all Cauchy sequences converge).

Theorem 5.3.5. *$\mathbb{R}^k$ is complete with respect to the Euclidean metric.*

Proof. Let $\{\boldsymbol{x}_n\}$ be a Cauchy sequence in $\mathbb{R}^k$ and $\boldsymbol{x}_n = (x_{1,n}, x_{2,n}, \ldots, x_{k,n})$. Given $\varepsilon > 0$, there is N such that $n \geq N$ and $m \geq N$ imply

$$\left\|x_{j,n} - x_{j,m}\right\| \leq \|\boldsymbol{x}_n - \boldsymbol{x}_m\| < \varepsilon,$$

for $j = 1, 2, \ldots, k$. Hence, each real sequence $\{x_{j,n}\}$ is a Cauchy sequence. Since $\mathbb{R}$ is complete, $\{x_{j,n}\}$ converges to some real number x_j for all $j = 1, 2, \ldots, k$. Hence, by Theorem 5.2.14, $\{\boldsymbol{x}_n\}$ converges to $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$. ■

5.4 Total boundedness

Definition 5.4.1. A subset S of a metric space X is said to be **totally bounded** if for every $\varepsilon > 0$, there exists a finite subset T of S such that $S \subseteq \bigcup_{x \in T} N_\varepsilon(x)$.

Theorem 5.4.2. *A subset S of a metric space X is totally bounded if and only if every sequence in S has a Cauchy subsequence.*

Proof. Suppose S is totally bounded. Let T_k be the finite set of points of S such that $S \subseteq \bigcup_{x \in T_k} N_{1/k}(x)$ and let $\{p_n\}$ be a sequence in S. Since T_1 is finite, there exists at least one point $x \in T_1$ such that $N_1(x)$ contains a subsequence $\{q_{1,n}\}$ of $\{p_n\}$. Similarly, at least one point $x \in T_2$ has the property that $N_{1/2}(x)$ contains a subsequence $\{q_{2,n}\}$ of $\{q_{1,n}\}$. Continuing the process, we obtain sequences $\{q_{1,n}\}, \{q_{2,n}\}, \ldots$ such that $d(q_{k,i}, q_{k,j}) < 1/k$ for all k and for all i, j. Then the sequence $\{q_{k,k}\}$ has the property that for every $\varepsilon > 0$, there exists N such that $i \geq N$ and $j \geq N$ imply $d(q_{i,i}, q_{j,j}) < 1/N < \varepsilon$, since both $q_{i,i}$ and $q_{j,j}$ are elements of the sequence $\{q_{N,n}\}$. Thus, $\{q_{k,k}\}$ is a Cauchy subsequence of $\{p_n\}$.

Conversely, suppose S is not totally bounded. This implies that there is $\varepsilon > 0$ such that for every finite subset T of S, $\bigcup_{x \in T} N_\varepsilon(x)$ cannot contain S. Let p_1 be a point in S. Then $N_\varepsilon(p_1)$ cannot contain S, so there must exists a point $p_2 \in S$ such that $d(p_1, p_2) \geq \varepsilon$. But $N_\varepsilon(p_1) \cup N_\varepsilon(p_2)$ also cannot contain S, so there is a point $p_3 \in S$ such that $d(p_k, p_3) \geq \varepsilon$ for $k = 1, 2$. Continuing the process, we obtain a sequence $\{p_n\}$ with the property that $d(p_m, p_n) \geq \varepsilon$ for all distinct m, n. Clearly, no subsequence of $\{p_n\}$ can be a Cauchy sequence. ■

Theorem 5.4.3. *Let S be a subset of a metric space X. If S is totally bounded, then S is bounded.*

Proof. Suppose S is totally bounded. Then there is a finite number of points $x_1, x_2, \ldots, x_k$ of S such that $S \subseteq \bigcup_{j=1}^{k} N_1(x_j)$. Choose x_m and x_n such that $d(x_m, x_n) \geq d(x_i, x_j)$ for all $i, j = 1, 2, \ldots, k$, and let $r = 1 + d(x_m, x_n)$. For any $x \in S$, $x \in N_1(x_j)$ for some j. Thus we have

$$d(x, x_m) \leq d(x, x_j) + d(x_j, x_m) < r.$$

Hence, $x \in N_r(x_m)$ for all $x \in S$, so S is bounded. ∎

Theorem 5.4.4. *A subset S of $\mathbb{R}^k$ is bounded if and only if it is totally bounded.*

Proof. Suppose S is bounded and let $\{\boldsymbol{x}_n\}$ be a sequence in S. By Theorem 5.2.15, $\{\boldsymbol{x}_n\}$ has a convergent subsequence $\{\boldsymbol{x}_{n_i}\}$. Since every convergent sequence is a Cauchy sequence, $\{\boldsymbol{x}_{n_i}\}$ is a Cauchy subsequence of $\{\boldsymbol{x}_n\}$. Hence, by Theorem 5.4.2, S is totally bounded. The converse follows from Theorem 5.4.3. ∎

5.5 Compactness

Definition 5.5.1. Let S be a subset of a metric space X and let $\{G_\alpha\}$ be a collection of open subsets of X. If $S \subseteq \bigcup_\alpha G_\alpha$, then we say that $\{G_\alpha\}$ covers S or that $\{G_\alpha\}$ is an **open cover** of S. A subset of $\{G_\alpha\}$ that also covers E is called a subcover.

Definition 5.5.2. A subset S of a metric space X is **compact** if every open cover of S contains a finite subcover.

Definition 5.5.3. Let S be a subset of a metric space X.

(a) S is **limit point compact** if every infinite subset of S has a limit point in S.

(b) S is **sequentially compact** if every sequence in S has a convergent subsequence whose limit lies in S.

Theorem 5.5.4. *Let X be a metric space. Then the following statements are equivalent:*

(a) X is compact.

(b) X is limit point compact.

(c) X is sequentially compact.

(d) X is complete and totally bounded.

Proof. First, suppose X is compact. Let S be an infinite subset of X and suppose S has no limit point in X. Then for each $x \in X$, there exists a neighborhood V_x that contains at most one element of S. It is clearly that no finite subcollection of $\{V_x\}$ can cover S, and hence X. This contradicts the compactness of X. Hence, (a) implies (b).

Next, suppose X is limit point compact and $\{p_n\}$ be a sequence in X. Let S be the range of $\{p_n\}$. If S is finite, then there exists $p \in S$ and a sequence $\{n_i\}$ with $n_1 < n_2 < n_3 < \cdots$ such that $p_{n_1} = p_{n_2} = p_{n_3} = \cdots = p$. A subsequence $\{p_{n_i}\}$ thus obtained converges to p. If S is infinite, S has a limit point $p \in X$. Hence for any positive integer i, there is p_{n_i} such that $d(p_{n_i}, p) < 1/i$. Given $\varepsilon > 0$, let k be a positive integer such that $1/k < \varepsilon$. Then, if $i \geq k$, we have $d(p_{n_i}, p) < \varepsilon$. Thus, $\{p_{n_i}\}$ converges to p, so (b) implies (c).

Now suppose X is sequentially compact. First, let $\{p_n\}$ be a Cauchy sequence in X. There is a subsequence $\{p_{n_i}\}$ of $\{p_n\}$ that converges to some $p \in X$. By Theorem 5.2.9b, $\{p_n\}$ also converges to p. Hence, X is complete. Next, let $\{q_n\}$ be a sequence in X. By sequental compactness, $\{q_n\}$ has a convergent subsequence $\{q_{n_i}\}$. Since every convergent sequence is a Cauchy sequence, $\{q_{n_i}\}$ is a Cauchy subsequence of $\{q_n\}$. Thus, by Theorem 5.4.2, X is totally bounded. Hence, (c) implies (d).

Finally, suppose X is complete and totally bounded but not compact. Then there exists an open cover $\{G_\alpha\}$ of X which contains no finite subcollection that also covers X. For any $x \in X$, let $V_r(x)$ be the set of points y such that $d(x, y) \leq r$. Then $V_r(x)$ is closed and $N_r(x) \subseteq V_r(x)$. Since X is totally bounded, there is a finite set $S \subseteq X$ such that $X \subseteq \bigcup_{x \in S} V_{1/2}(x)$. Because X is not compact, we can find $T_1 = V_{1/2}(x)$ for some $x \in S$ such that T_1 is not covered by a finite subcollection of $\{G_\alpha\}$. Since T_1 is also totally bounded, we can continue the process to find a sequence $\{T_n\}$ of subsets of X with the following properties:

(i) $T_1 \supseteq T_2 \supseteq T_3 \supseteq \cdots$;

(ii) T_n is closed;

(iii) T_n is not covered by any finite subcollection of $\{G_\alpha\}$;

(iv) For any $x, y \in T_n$, $d(x, y) \leq 1/n$.

Let $p_n \in T_n$. By (i) and (iv), $\{p_n\}$ is a Cauchy sequence. Since X is complete, $p_n \to p$ for some $p \in X$. For any fixed positive integer k, (i) implies that $\{p_{n+k-1}\}$ is a subsequence of $\{p_n\}$ in T_k, which also converges to p. Hence, by (ii), $p \in T_n$ for all n. For some α, $p \in G_\alpha$. Since G_α is open, there is $\varepsilon > 0$ such that $N_\varepsilon(p) \subseteq G_\alpha$. If n is sufficiently large that $1/n < \varepsilon$, we obtain, by (iv), $T_n \subseteq N_\varepsilon(p)$, which contradicts (iii). Thus, (d) implies (a). ∎

Theorem 5.5.5. *Every closed subset of a compact metric space X is compact.*

Proof. Let S be a closed subset of X. Since X is totally bounded, S is also totally bounded. By Theorem 5.3.2b, S is complete since X is complete. Thus, S is compact. ∎

Theorem 5.5.6. *Every compact subset of a metric space X is closed and bounded.*

Proof. Let S be a compact subset of X. Since S is complete and totally bounded, it is closed and bounded by Theorems 5.3.2a and 5.4.3. ∎

Theorem 5.5.7 (Heine-Borel)**.** *Every subset S of $\mathbb{R}^k$ is compact if and only if it is closed and bounded.*

Proof. Suppose S is closed and bounded. Since $\mathbb{R}^k$ is complete, S is complete by Theorem 5.3.2b. That S is totally bounded follows from Theorem 5.4.4. Hence, S is compact. The converse follows from Theorem 5.5.6. ∎

5.6 Series

Definition 5.6.1. Given a sequence $\{a_n\}_{n=1}^{\infty}$, the sum

$$\sum_{n=1}^{\infty} a_n,$$

is called an **infinite series**.

The sequence

$$s_n = \sum_{i=1}^{n} a_i,$$

is called the **partial sum** of the series.

If $\lim_{n\to\infty} s_n < 0$, the series is said to **converge**, else it **diverges**. If $\sum a_n$ converges but $\sum |a_n|$ does not, we say that a_n *conditionally converges*, otherwise it is *absolutely convergent.*

An important example of a series is the **geometric series** $\sum_{i=1}^{\infty} ar^i$. We have the following theorem.

Theorem 5.6.2. *The geometric series converges if and only if* $|r| < 1$, *in which case it converges to* $\frac{a}{1-r}$.

Proof. We note that for $r \neq 1$,

$$\begin{aligned} s_n(1-r) &= a\left(1+r+r^2+r^3+\cdots+r^n\right)(1-r) \\ &= a\left(1+r+r^2+r^3+\cdots+r^n\right)1 - a\left(1+r+r^2+r^3+\cdots+r^{n-1}+r^n\right)r \\ &= a\left(1+r+r^2+r^3+\cdots+r^n-r-r^2-r^3-\cdots-r^n-r^{n+1}\right) \\ &= a\left(1-r^{n+1}\right) \end{aligned}$$

SO

$$\begin{aligned} s_n(1-r) &= a\left(1-r^{n+1}\right) \\ s_n &= a\frac{1-r^{n+1}}{1-r} \end{aligned}$$

If $|r| < 1, \lim_{n\to\infty} r^n = 0$ so

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} a\frac{1-r^{n+1}}{1-r} = a\frac{1}{1-r}.$$

Thus, when $|r| < 1$ the geometric series converges to $a/(1-r)$.

If $|r| > 1$, then

$$\lim_{n\to\infty} r^n = \infty$$

and so the series diverges.

If $|r| = 1$, then

$$|s_n| = \sum_{i=0}^{n} 1 = n+1.$$

Therefore,

$$\lim_{n\to\infty} |s_n| = \lim_{n\to\infty} (n+1) = \infty$$

and so the series diverges. ∎

We have the following facts that follow from the linearity of addition.

Theorem 5.6.3. *Suppose $\sum a_n$ and $\sum b_n$ are convergent series and $c \in \mathbb{R}$, then $\sum a_n + cb_n = \sum a_n + c \sum b_n$.*

We have the following tests for convergence and divergence. The proofs are omitted.

Theorem 5.6.4. *Let $\sum a_n$ and $\sum b_n$ be a series.*

1. *(Divergence Test) If $\sum a_n$ converges, then $\lim_{n\to\infty} a_n = 0$.*
2. *(Integral Test) If f is a continuous, positive, decreasing function such that $a_n = f(n)$ then $\sum a_n$ and $\int_1^\infty f(x)dx$ both converge or diverge.*
3. *(p-Series Test) If $a_n = \frac{1}{n^p}$, then $\sum a_n$ converges if and only if $p > 1$.*
4. *(Alternating Test) If $a_n = (-1)^n b_n$ with b_n decreasing and $b_n \geq 0$, $\sum a_n$ converges if and only if $b_n \longrightarrow 0$.*
5. *(Comparison Test) Suppose $a_n \leq b_n$. If $\sum b_n$ converges, so does $\sum a_n$. If $\sum a_n$ diverges, so does $\sum b_n$.*
6. *(Absolute Convergence Test) If $\sum |a_n|$ converges, so does $\sum a_n$.*
7. *(Ratio Test) If $a_n \geq 0$ and $\frac{a_{n+1}}{a_n} \longrightarrow L$, then $\sum a_n$ converges if $L < 1$ and diverges if $L > 1$.*
8. *(Root Test) If $a_n \geq 0$ with $(a_n)^{\frac{1}{n}} \longrightarrow L$, then $\sum a_n$ converges if $L < 1$ and diverges if $L > 1$.*
9. *(Dirichlet's Test) If $s_n = \sum_{i=1}^n a_i$ is a bounded sequence, and b_n is a decreasing positive sequence limiting to zero, then $\sum a_n b_n$ converges.*
10. *(Abel's Test) If $\sum a_n$ is convergent and b_n is a monotone convergence sequence of numbers, then $\sum a_n b_n$ converges.*
11. *(Cauchy Product) If $\sum a_n$ converges absolutely, and $\sum b_n$ converges, then $\sum a_n b_n = (\sum a_n)(\sum b_n)$.*

Be careful about conditionally convergent series.

Theorem 5.6.5 (Riemann's Rearrangement Theorem). *If $\sum a_n$ is absolutely convergent, then for any bijection $p : \mathbb{N} \longrightarrow \mathbb{N}$, $\sum a_{p(n)}$ converges absolutely and their sums coincide.*

If $\sum a_n$ is conditionally convergent, for all $t \in \mathbb{R}$, there exists a bijection $p : \mathbb{N} \longrightarrow \mathbb{N}$ such that $\sum a_{p(n)} = t$.

Proof. We only prove the second part (the first part is mostly definition-chasing). Let $t > 0$ (the proof for negative t is much the same).

Let $b_n = \max\{a_n, 0\}$ be the sequence of positive terms and $c_n = \min\{a_n, 0\}$ be the sequence of negative terms. Note that the first tends to $+\infty$ and the latter to $-\infty$. (Hopefully this is intuitive, but to see it formally, note that $b_n = \frac{1}{2}(|a_n| + a_n)$ and note that if $\sum b_n$ converges, then $\sum a_n$ must - a contradiction.)

Since the series $\sum b_n$ of positive terms in $\sum a_n$ is divergent, there exists first index k such that

$$t_1 = b_1 + b_2 + \ldots + b_k > t.$$

Since the series $\sum c_n$ of negative terms in $\sum a_n$ is divergent, there is first index m such that

$$t_2 = t_1 + c_1 + c_2 + \ldots + c_m < t.$$

We continue in the same manner to claim the existence of smallest indices u and v such that

$$t_3 = t_2 + b_{k+1} + b_{k+2} + \ldots + b_u > a,$$

and

$$t_4 = t_3 + c_{m+1} + c_{m+2} + \ldots + c_v < t,$$

and so on. The sequence t_i so obtained converges to t since the terms a_n (and thus also b_n and c_n) tend to 0, implying that, with every step in the construction, the difference between t_i and t (which is no larger than the last term in b_n or c_n added to the sequence) tends to zero. ■

At some point in the first year, you will likely use the following helpful formula.

Theorem 5.6.6 (Abel's Partial Summation Formula). *Let $n > 1$, and $s_n = \sum_{i=1}^n a_i$*

$$\sum_{i=1}^{n} a_i b_i = b_{n+1} s_n + \sum_{i=1}^{n} s_i (b_m - b_{m+1}).$$

This is analogous to the integration by parts formula.

Exercise 5.7. *Write and prove an expression for $\sum_{i=1}^m i$, $\sum_{i=1}^m i^2$, and*

$$\sum_{k=1}^{\infty} kr^k \text{ for } 0 < r < 1.$$

6

Continuity and Fixed Point Theorems

Contents

6.1 Continuity

Definition 6.1.1. Let X and Y be metric spaces, E be a subset of X, p be a limit point of E, q be a point of Y, and $f : E \to Y$ be a function from E into Y. Then the notation

$$\lim_{x \to p} f(x) = q.$$

means for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $x \in E$ and $0 < d_X(x, p) < \delta$ imply $d_Y(f(x), q) < \varepsilon$. In this case we may also write $f(x) \to q$ as $x \to p$.

Note that p need not be a point in E, and even if $p \in E$, it may be the case that $f(p) \neq \lim_{x \to p} f(x)$. For the latter remark, consider $f(x) = 1$ if $x = 1$, and $f(x) = 0$ otherwise.

Theorem 6.1.2. *Let X, Y, E, p, q and f be defined as in Definition 6.1.1. Then*

$$\lim_{x \to p} f(x) = q$$

if and only if

$$\lim_{n \to \infty} f(p_n) = q$$

for every sequence $\{p_n\}$ in E such that $p_n \neq p$ and $p_n \to p$.

Proof. Suppose $f(x) \to q$ as $x \to p$ and let $\{p_n\}$ be a sequence such that $p_n \neq p$ and $p_n \to p$. Hence, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $0 < d_X(x, p) < \delta$ implies $d_Y(f(x), q) < \varepsilon$.

Since $p_n \longrightarrow p$, there exists an integer N such that $n \geq N$ implies $0 < d_X(p_n, p) < \delta$. Thus, $d_Y(f(p_n), q) < \varepsilon$ as soon as $n \geq N$.

Conversely, suppose $f(p_n) \longrightarrow q$ for every sequence $\{p_n\}$ in E such that $p_n \neq p$ and $p_n \longrightarrow p$, but $f(x)$ does not converge to q as $x \longrightarrow p$. This means that there exists, for all $\delta > 0$ and some $\varepsilon > 0$, a point x (which may depend on δ) in E such that $0 < d_X(x, p) < \delta$ but $d_Y(f(x), q) \geq \varepsilon$. However, setting $\delta = 1/n$, $n = 1, 2, 3, \ldots$, we have $0 < d_X(x_n, p) < 1/n$ but $d_Y(f(x_n), q) \geq \varepsilon$. This is a contradiction because we have found a sequence $\{x_n\}$ that converges to p but $f(x_n)$ does not converge to q. ■

Corollary 6.1.3. *If f has a limit at p, this limit is unique.*

Proof. This follows from Theorems 5.2.3b and 6.1.2. ■

Theorem 6.1.4. *Suppose $E \subset X$, a metric space, p is a limit point of E, f and g are real functions on E, and*

$$\lim_{x\to p} f(x) = q \qquad \lim_{x\to p} g(x) = r$$

Then

$$\lim_{x\to p} (f+g)(x) = q + r$$
$$\lim_{x\to p} (fg)(x) = qr$$
$$\lim_{x\to p} (\frac{f}{g})(x) = \frac{q}{r} \qquad \text{if } r \neq 0$$

Proof. Follows from Theorems 6.1.2 and 5.2.16. ■

Definition 6.1.5. Let X and Y be metric spaces and $f : X \longrightarrow Y$ be a function from X into Y. f is said to be **continuous at a point** $p \in X$ if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_X(x, p) < \delta$, $x \in X$, implies $d_Y(f(x), f(p)) < \varepsilon$. If f is continuous at every point of a subset E of X, then f is said to be **continuous on** E.

Note that f has to be defined at the point p in order to be continuous at p. If p is an isolated point of X, then any function f which has E as its domain is continuous at p.

Theorem 6.1.6. *In the situation given in Definition 6.1.5, assume further that p is also a limit point. Then f is continuous at p if and only if*

$$\lim_{x\to p} f(x) = f(p).$$

Proof. This is clear if we compare Definitions 6.1.1 and 6.1.5. ■

Example. Here are some examples:

(i) $f(x) = 1$ if $x = 1$, and $f(x) = 0$ otherwise. f has a limit at $x = 1$, but the limit is not equal to $f(x)$, so $f(x)$ is not continuous at $x = 1$.

(ii) $f(x) = 0$ for $x < 1$, and $f(x) = 1$ otherwise. Then f does not have a limit at $x = 1$. Hence, f is not continuous at $x = 1$

♣

If $f : \mathbb{R} \to R$, roughly speaking f is continuous if the graph is a single unbroken curve with no "holes" or "jumps".

Theorem 6.1.7. *Let X and Y be metric spaces and $f : X \to Y$ be a function from X into Y. Then f is continuous on X if and only if for every open set V in Y, $f^{-1}(V)$ is open in X.*

Proof. Suppose f is continuous on X. Let V be an open set in Y and $p \in f^{-1}(V)$. Since V is open, there exists an $\varepsilon > 0$ such that $d_Y(y, f(p)) < \varepsilon$ implies $y \in V$. Since f is continuous at p, there exists a $\delta > 0$ such that $d_X(x, p) < \delta$ implies $d_Y(f(x), f(p)) < \varepsilon$. This means $f(x) \in V$, so $x \in f^{-1}(V)$. Thus, p is an interior point of $f^{-1}(V)$.

Conversely, suppose $f^{-1}(V)$ is open in X for every open set V in Y. Let $p \in X$ and $\varepsilon > 0$. Let V be the set of points y such that $d_Y(y, f(p)) < \varepsilon$. Hence, V is open, so $f^{-1}(V)$ is open. Thus, there exists a $\delta > 0$ such that $d_X(x, p) < \delta$ implies $x \in f^{-1}(V)$. This means $f(x) \in V$, so $d_Y(f(x), f(p)) < \varepsilon$. ■

Corollary 6.1.8. *Let X and Y be metric spaces and f be a function from X into Y. Then f is continuous on X if and only if for every closed set V in Y, $f^{-1}(V)$ is closed in X.*

Proof. If V is closed in Y, V^c is open in Y. Since $f^{-1}(V^c) = (f^{-1}(V))^c$, the latter set is open by Theorem 6.1.7. ■

Theorem 6.1.9. *Let f and g be real continuous functions on a metric space X. Then $f+g$, fg and $\frac{f}{g}$ are continuous on X (assuming $g(x) \neq 0$ in the last case).*

Proof. At isolated points of X there is nothing to prove. At limit points, it follows from Theorem 6.1.4 and 6.1.6. ∎

Theorem 6.1.10. *(a) Let $f_1, \ldots, f_k$ be real functions on a metric space and let $\mathbf{f}$ be the mapping of X into $\mathbb{R}^k$ defined by*

$$\mathbf{f}(x) = (f_1(x), \ldots, f_k(x))(x \in X)$$

then $\mathbf{f}$ is continuous if and only if each of the functions $f_1, \ldots, f_k$ is continuous.

(b) If $\mathbf{f}$ and $\mathbf{g}$ are continuous mappings of X into $\mathbb{R}^k$, then $\mathbf{f}+\mathbf{g}$ and $\mathbf{f}\cdot\mathbf{g}$ are continuous.

Proof. (a) follows from

$$|f_j(x) - f_j(y)| \leq |\mathbf{f}(x) - \mathbf{f}(y)| = \left(\sum_{i=1}^{k} |f_i(x) - f_i(y)|\right)^{\frac{1}{2}}$$

for $j = 1, \ldots, k$.
(b) follows from (a) and Theorem 6.1.9. ∎

Example. Let $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$, and define ϕ_i for $i = 1, \ldots, k$ as

$$\phi_i(\mathbf{x}) = x_i \qquad (\mathbf{x} \in \mathbb{R}^k)$$

Then ϕ_i is continuous on $\mathbb{R}^k$ since the inequality

$$|\phi_i(\mathbf{x}) - \phi_i(\mathbf{y})| \leq |\mathbf{x} - \mathbf{y}|$$

shows that we may take $\delta = \varepsilon$.

Repeated application of Theorem 6.1.9 shows that every monomial

$$x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$$

where $n_1, \ldots, n_k$ are nonnegative integers, is continuous on $\mathbb{R}^k$. Since constant functions are continuous, applying Theorem 6.1.9 again we have that every polynomial P, given by

$$P(\mathbf{x}) = \sum_{i=1}^{n} c_i x_1^{n_{1,i}} x_2^{n_{2,i}} \cdots x_k^{n_{k,i}}$$

is continuous on $\mathbb{R}^k$. ♣

Theorem 6.1.11. *If f is a continuous mapping of a compact metric space X into a metric space Y, then $f(X)$ is compact.*

Proof. Let $\{V_\alpha\}$ be an open cover of $f(X)$. By Theorem 6.1.7, $f^{-1}(V_\alpha)$ is open for every α. Since X is compact, there exists $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$X \subseteq f^{-1}(V_{\alpha_1}) \cup f^{-1}(V_{\alpha_2}) \cup \cdots \cup f^{-1}(V_{\alpha_n}).$$

Hence,

$$f(X) \subseteq V_{\alpha_1} \cup V_{\alpha_2} \cup \cdots \cup V_{\alpha_n}.$$

■

Theorem 6.1.12. *Let f be a continuous real function on a compact metric space X. Then there exists points p and q in X such that $f(q) \leq f(x) \leq f(p)$ for all $x \in X$. That is, the function f* attains *its maximum and minimum on any compact set.*

Proof. Let $M = \sup_{p \in X} f(p)$ and $m = \inf_{q \in X} f(q)$. Since $f(X)$ is compact, it is also closed and bounded, by the Heine-Borel Theorem. Hence $M, m \in f(X)$, by Theorem 5.1.10. ■

Theorem 6.1.13 (Intermediate Value Theorem)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function and let $f(a) < k < f(b)$. Then there exists some $c \in (a, b)$ such that $f(c) = k$.*

Proof. Let S be the set of all $x \in [a, b]$ such that $f(x) \leq k$. Clearly $a \in S$, so that S is non-empty. It is also bounded above, so that S is compact and $c = \sup S$ exists. We will show that $f(c) = k$.

Since f is continuous, for any $\varepsilon > 0$, there is a $\delta > 0$ such that for all $x \in (c-\delta, c+\delta)$, we have

$$f(x) - \varepsilon < f(c) < f(x) + \varepsilon.$$

Because $c = \sup S$, there is some $a^* \in (c - \delta, c]]$ that is contained in S, and so

$$f(c) < f(a^*) + \varepsilon \leq k + \varepsilon.$$

Meanwhile, for any $a' \in (c, c + \delta)$, since $a' \notin S$, we have

$$f(c) > f(a') - \varepsilon > k - \varepsilon.$$

Combining these inequalities, we obtain

$$k - \varepsilon < f(c) < k + \varepsilon$$

for any $\varepsilon > 0$, so that $f(c) = k$ as required.

■

Definition 6.1.14. Let f be a mapping of a metric space X into a metric space Y. We say that f is **uniformly continuous** on X if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$d_Y(f(p), f(q)) < \varepsilon$$

for all p and q in X for which $d_X(p, q) < \delta$.

Uniform continuity is a property of a function on a set, while continuity can be defined at a single point. Clearly, every uniformly continuous function is continuous. However, a continuous function may not be uniformly continuous. To see that, let $f(x) = \frac{1}{x}$ defined on $(0, \infty)$. Continuous but not uniformly continuous.

Theorem 6.1.15. *Let f be a continuous mapping of a compact metric space X into a metric space Y. Then f is uniformly continuous on X.*

Proof. Let $\varepsilon > 0$ be given. Since f is continuous, we can associate to each point $p \in X$, $\delta_p > 0$ such that

$$q \in X,\ d_X(p, q) < \delta_p \Rightarrow d_Y(f(p), f(q)) < \frac{\varepsilon}{2}$$

Let J_p be the set of all $q \in X$ for which $d_X(p,q) < \frac{1}{2}\delta_p$.
Since $p \in J_p$, the collection of all sets J_p is an open cover of X, and since X is compact, there is a finite set of points $p_1, ..., p_n \in X$ such that

$$X \subset J_{p_1} \cup .... \cup J_{p_n}$$

Let $\delta = \frac{1}{2}\min\{\delta_{p_1}, ..., \delta_{p_n}\}$. Then $\delta > 0$ (for this, the finiteness of the covering is crucial).
Now take $q, p \in X$ such that $d_X(p,q) < \delta$. Then there is $m \in \{1, ..., n\}$ such that $p \in J_{p_m}$, and hence $d_X(p, p_m) < \frac{1}{2}\delta_{p_m}$. Also

$$d_X(q, p_m) \leq d_X(p,q) + d_X(p, p_m) < \delta + \frac{1}{2}\delta_{p_m} \leq \delta_{p_m}$$

But then

$$d_Y(f(p), f(q)) \leq d_Y(f(p), f(p_m)) + d_Y(f(q), f(p_m)) < \varepsilon$$

■

6.2 Correspondences

Recall that a *function* is a mathematical object that associates to every point in the *domain* a single point in the *range*. A correspondence generalizes the idea of a function, allowing a point from the domain to be associated with more than one point in the range.

Definition 6.2.1. Let X and Y be sets. A **correspondence** ϕ between X and Y is a nonempty relation $\phi \subset X \times Y$. That is, for every $x \in \text{domain}(\phi)$, $\phi(x)$ is a *subset* of Y. We write $\phi : X \rightrightarrows Y, x \rightrightarrows \phi(x)$ or $\phi : X \longrightarrow 2^Y$ to denote a correspondence from X to Y.

Definition 6.2.2. If the correspondence ϕ never maps a point in the domain into the empty set, we say that ϕ is **nonempty valued**.
If the correspondence ϕ maps every point in the domain into a set containing a single element,

we say that ϕ is **singleton valued**. If the correspondence maps every point in the domain into a closed (compact) set, we say the correspondence is **closed-valued (compact-valued)**.

Definition 6.2.3. The **graph** of the correspondence $\phi : X \rightrightarrows Y$ is the set of points $\operatorname{gr}(\phi) = \{(x, y) \in X \times Y : y \in \phi(x)\}$.

Defining continuity of correspondences is slightly more complicated than defining continuity of functions. Intuitively, a correspondence ϕ is continuous if "small" changes in x produce "small" changes in the set $\phi(x)$. With functions, it is obvious what it means that $f(x)$ and $f(x')$ are similar when x and x' are similar. With correspondences we need to make a comparison between the sets $\phi(x)$ and $\phi(x')$. For that, we need two distinct concepts.

Definition 6.2.4. Let X and Y be metric spaces, and $\phi : X \rightrightarrows Y$ be a correspondence.

(a) ϕ is **upper hemicontinuous (uhc)** at $x_0 \in X$ if, for every open set $V \supseteq \phi\left(x_0\right)$, there is an open set U with $x_0 \in U$ such that

$$\phi(x) \subseteq V \text{ for every } x \in U \cap X$$

(b) ϕ is **lower hemicontinuous (lhc)** at $x_0 \in X$ if, for every open set V such that $\phi\left(x_0\right) \cap V \neq \emptyset$, there is an open set U with $x_0 \in U$ such that $\phi(x) \cap V \neq \emptyset$ for every $x \in U \cap X$

(c) ϕ is **continuous** at $x_0 \in X$ if it is both uhc and lhc at x_0.

(d) ϕ is upper hemicontinuous (respectively lower hemicontinuous, continuous) if it is uhc (respectively lower hemicontinuous, continuous) at every $x \in X$.

Upper hemicontinuity captures the idea that $\phi(x)$ will not "suddenly" contain new points just as we move past some point x, while lower hemicontinuity captures the idea that $\phi(x)$ will not "suddenly" lose points just as we move past some point x.

Just as continuity can be defined in terms of open sets or in terms of sequences, hemicontinuity also admits a sequence definition. However, the sequence definition of upper hemicontinuity *only applies when the space is compact.*

Theorem 6.2.5. *Let $\phi : X \rightrightarrows Y$ be a non-empty-valued correspondence.*

1. *Suppose ϕ satisfies the following: whenever $\{x_n\}$ is a sequence in X with limit $x \in X$ and $\{y_n\}$ is a sequence in Y with $y_n \in \phi(x_n)$, then $\{y_n\}$ has a convergent subsequence with limit in $\phi(x)$. Then ϕ is uhc.*

2. *Suppose that ϕ is uhc and compact-valued. Then for any $x \in X$ and sequence $\{x_n\}$ converging to x, and any sequence $\{y_n\}$ in Y with $y_n \in f(x_n)$, there is a $y \in \phi(x)$ such that a subsequence of y_n converges to y.*

3. *The correspondence ϕ is lhc if and only if for any $x \in X$, $y \in \phi(x)$ and sequence $\{x_n\}$ with limit x, there exists a sequence $\{y_n\}$ in Y and $N \in \mathbb{N}$ such that for all $n > N$, $y_n \in \phi(x_n)$ and $\lim_{n\to\infty} y_n = y$.*

Proof. (1) Suppose that ϕ satisfied the property but were not uhc. Then there is an x and open set V containing $f(x)$ such that for any open set U containing x, there is an $x' \in U$ with $f(x') \notin V$. By taking smaller U, we may obtain a sequence $\{x_n\}$ limiting to x with $y_n \in f(x_n)$ but $y_n \notin V$. Since V^c is closed, and y_n is a sequence in V^c, the limit of any convergent subsequence is in V^c as well. But then we have constructed an $\{x_n\}$ with no subsequence converging to a point in $\phi(x)$, a contradiction.

(2) Let ϕ be uhc and compact-valued. Consider any $\{x_n\}$ and $\{y_n\}$ as in the statement of the theorem.

We first show that $\{y_n\}$ is bounded. This will imply it has a convergent subsequence, and then show that the limit of this subsequence is in $\phi(x)$.

By assumption $\phi(x)$ is compact and thus bounded, so that there is some open B containing $\phi(x)$. By uhc, there is a neighborhood U of x with $\phi(z) \subseteq B$ for all $z \in U$. Since $x_n \to x$, $x_n \in U$ for sufficiently large n, so that $\phi(x_n) \subseteq B$ for sufficiently large n. But then $\{y_n\} \subseteq \{\phi(x_n)\}$ is bounded.

Let y be the limit of a convergent subsequence of y_n. Suppose that $y \notin \phi(x)$. Since $\phi(x)$ is compact and therefore closed, the distance between y and $\phi(x)$ is strictly positive. Hence we can identify a closed $\varepsilon-$ball around $\phi(x)$ that does not contain y. Because ϕ is uhc, $\phi(x_n)$ will be contained in this closed ball $B_\varepsilon(\phi(x))$ for sufficiently large n. Therefore, so will the convergent subsequence. The limit of the subsequence must also be in $B_\varepsilon(\phi(x))$. But this contradicts our choice of ε.

(3) First, suppose that ϕ is lhc and define $\{x_n\}$, y as in the theorem. For each integer k, consider $N_{1/k}(y)$. Clearly $N_{1/k}(y) \cap \phi(x)$ is nonempty, because it at least contains y. Because ϕ is lhc, for each k there exists a neighborhood U_k of x such that for each $z \in U_k$, $\phi(z_k) \cap N_{1/k}(y) \neq \emptyset$. Because $x_n \to x$, x_n is eventually in U_k for each k and sufficiently large n. Choose a subsequence x_{n_k} of x_n with successive terms in U_k. The associated $y_{n_k} \in \phi(x_{n_k}) \cap N_{1/k}(y)$. As k tends to infinity, clearly we must have $y_{n_k} \to y$.

Now suppose that the property holds but that ϕ were not lhc. Then there exists an open V with $\phi(x) \cap V \neq \emptyset$ such that every neighborhood U of x contains a point z_U with $\phi(z_U) \cap V = \emptyset$. Taking a sequence of such neighborhoods $U_n = N_{1/n}(x)$ and an appropriate $x_n \in U_n$, we obtain

x_n converging to x with $\phi(x_n) \cap V = \emptyset$ for all n. Then any sequence of $y_n \in \phi(x_n)$ is contained in V^c, so that if it converges, it converges in V^c (which is closed). If we let y be a point in $\phi(x) \cap V$, it is clear that no $\{y_n\}$ can converge to y. This is a contradiction.

■

Definition 6.2.6. A correspondence ϕ has the **closed graph** property if its graph is a closed subset of $X \times Y$, that is, for any $x_n \longrightarrow x \in X$ and $y_n \longrightarrow y \in Y$ with $y_n \in \phi(x_n)$, then $y \in \phi(x)$.

Check that the correspondence

$$\phi(x) = \begin{cases} \{1/x\} & \text{if } x > 0 \\ \{0\} & \text{if } x = 0, \end{cases}$$

has the closed graph property but is not uhc and that

$$\phi'(x) = \begin{cases} \{1/x\} & \text{if } x > 0 \\ \mathbb{R} & \text{if } x = 0, \end{cases}$$

has the closed graph property but is uhc.

We have the following important theorem that makes identifying upper hemicontinuity easier in some examples.

Theorem 6.2.7. *Suppose $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$, and $\phi : X \longrightarrow 2^Y$.*

(i) If ϕ is closed-valued and upper hemicontinuous, then φ has closed graph.

(ii) If Y is compact, then φ has closed graph $\iff \phi$ is closed-valued and upper hemicontinuous.

Proof. Proof of (i): Suppose that ϕ is closed-valued and upper hemicontinuous. If ϕ does not have closed graph, there is some sequence $(x_n, y_n) \longrightarrow (x, y)$ where $y_n \in \phi(x_n)$ but $y \notin \phi(x)$. Since ϕ is closed-valued, $\phi(x)$ is closed, and we can identify disjoint open sets G, V such that $y \in G$ and $\phi(x) \subseteq V$. Since ϕ is uhc, there is an open set U with $x \in U$ such that $x \in U \cap X$ implies $\phi(x) \subseteq V$.

Since $(x_n, y_n) \longrightarrow (x, y)$, $x_n \in U$ for sufficiently large n, so that $y_n \in \phi(x_n) \subseteq V$. Thus for sufficiently large n, $\|y_n - y_0\| \geq \varepsilon$ for some $\varepsilon > 0$, so that $y_n \nrightarrow y$, a contradiction. Thus, ϕ has closed graph.

Proof of (ii): We only need to prove the forward implication. Suppose that the graph of ϕ is a closed set, and that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$, and $y_n \in \phi(x_n)$ for all n. Then $\{(x_n, y_n)\}$ is a sequence from graph of ϕ with limit point (x, y), and since the graph of ϕ is closed this means that (x, y) is in the graph, which is to say that $y \in \phi(x)$. Therefore, ϕ is upper hemicontinuous. ■

Theorem 6.2.8. *A singleton-valued correspondence ϕ is lower hemicontinuous if and only if it describes a continuous function. Similarly, a singleton-valued correspondence is upper hemicontinuous if and only if it describes a continuous function.*

Proof. Suppose ϕ is lower hemicontinuous. Take any $x \in X$ and sequence $\{x_n\}$ with limit x. Since $f(x) \in \phi(x)$, lower hemicontinuity means that for some sequences $\{y_n\}$ with $y_n \in \phi(x_n)$, $\lim_n y_n = f(x)$. But the only choice for y_n is $f(x_n)$, so $\lim f(x_n) = f(x)$. Therefore, f is continuous. Conversely, for $x \in X$ and $\{x_n\}$ with limit x, continuity of f ensures that $\lim_{n\to\infty} f(x_n) = f(x)$. Therefore, for all $y \in \phi(x)$ (and there is only one such y, namely $f(x)$), we can find $y_n \in \phi(x_n)$, namely $y_n = f(x_n)$ with limit $y = f(x)$. This is lower hemicontinuity.

Now suppose ϕ is upper hemicontinuous. Fix V open in Y, and consider the set $\phi^u(V) = \{x \in X : \phi(x) \subseteq V$. Upper hemicontinuity implies that this set is open. But $\phi^u(V) = f^{-1}(V)$, so f is continuous by the open mapping definition of continuity. The same logic gives that f continuous implies ϕ uhc. ■

Example. Consider the following correspondence ϕ defined for $X = Y = \mathbb{R}$ as follows:

$$\phi(x) = \begin{cases} \{2-x, 4-x\} & \text{for } x < 2 \\ [2-x, 4-x] & \text{for } 2 \le x \le 3 \\ \{x-3\} & \text{for } x > 3 \end{cases}$$

This correspondence is upper hemicontinuous. The easiest way to see this is to note that its graph is a closed set. It is not lower hemicontinuous. It fails at $x = 2$ and $x = 3$. For instance $1 \in \phi(2)$, but as you approach $x = 2$ from below, you cannot get "close" to value 1; there are only sequences approaching to value $y = 2$ and $y = 0$ (provided that they converge). ♣

Example. Consider the correspondence $\phi : \mathbb{R} \rightrightarrows \mathbb{R}$ defined by

$$\phi(x) = \begin{cases} \{2-x, 4-x\} & \text{for } x < 2 \\ [3-x, 5-x] & \text{for } 2 \le x \le 3 \\ \{x-3\} & \text{for } x > 3 \end{cases}$$

It is neither upper nor lower hemicontinuous. For lower hemicontinuity, it fails in a similar way as the previous correspondence at $x = 2$ and $x = 3$. For upper hemicontinuity, observe that the limit point $(2, 0)$ is not in the graph of ϕ (so the graph of ϕ is not closed). ♣

Example. Consider the correspondence $\phi : \mathbb{R} \rightrightarrows \mathbb{R}$ defined by

$$\phi(x) = \begin{cases} \{2-x, 4-x\} & \text{for } x < 2 \\ \emptyset & \text{for } x = 2 \\ [3-x, 5-x] & \text{for } 2 < x < 3 \\ \{0\} & \text{for } x = 3 \\ \{x-3\} & \text{for } x > 3 \end{cases}$$

The correspondence is lower hemicontinuous, but not upper hemicontinuous. ♣

6.3 Fixed Points

Definition 6.3.1. A **fixed point** of function $f : X \to X$ is a point $x \in X$ such that $f(x) = x$.

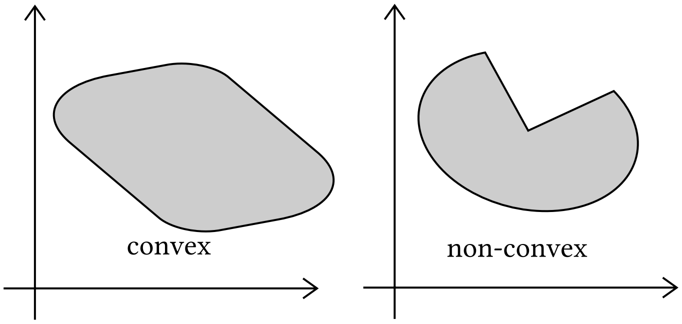
Fixed points are useful in economics because they characterize equilibria of carefully formulated economic problems. Fixed point theorems are conditions under which fixed points are guaranteed to exist. These come in many flavors, and we will consider only a couple of the most well-known and useful ones here. The simplest fixed point theorem is as follows.

Theorem 6.3.2. *Let $f : [0,1] \to [0,1]$ be continuous. Then there exists a fixed point of f.*

Proof. Consider the function $\phi : [0,1] \to \mathbb{R}$ given by $\phi(x) = f(x) - x$. Now $\phi(0) = f(0) \in [0,1]$, while $\phi(1) = f(1) - 1 \in [-1,0]$. Since $\phi(0) \geq 0$ and $\phi(1) \leq 0$, the intermediate value theorem implies that there is an $x \in [0,1]$ such that $\phi(x) = 0$ so that $f(x) = x$. ■

This fixed point theorem is the simplest example of a much more general and deep result called Brouwer's Fixed Point Theorem. In order to state the theorem, we need to introduce one piece of mathematical machinery, called convexity. We will study convexity a lot more in later chapters, so we will only include the very basic facts about convexity now.

Definition 6.3.3. A set $C \subseteq X$ is **convex** if for any $\mathbf{x}, \mathbf{y} \in C$, we have $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in C$ for all $\lambda \in (0,1)$.



Theorem 6.3.4 (Brouwer's Fixed Point Theorem). *Let $X \subset \mathbb{R}^k$ be nonempty, compact and convex, and let $f : X \to X$ be continuous function. Then there exists a fixed point of f.*

Brouwer's Fixed Point Theorem is more complicated to prove. We will not do it here. The classic proofs of Brouwer's Fixed Point Theorem are *non-constructive*, that is, they do not show you how to actually compute the fixed point. On the other hand, there are constructive proofs (using a combinatorial technique called Sperner's Lemma) that describe how to approximate fixed points to any desired level of approximation. However, a very famous result of modern computer science is that this problem of approximation is NP-hard, which means (approximately) that such approximation can take an exponentially long time (in the desired level of approximation). This result has had a significant influence in microeconomic theory and has led to a philosophical debate about the concept of equilibrium in economics (summarized roughly by 'if equilibria can't be computed efficiently by a computer, how do people behave according to them').

Most fixed point theorems have a version for correspondences, which are particularly helpful in economics (where correspondences arise naturally in situations of indifference). The following is the correspondence version of Brouwer's Fixed Point Theorem.

Theorem 6.3.5 (Kakutani's Fixed Point Theorem). *Let $F : X \rightrightarrows X$ be a nonempty, convex-valued and upper hemicontinuous correspondence on X, a nonempty, compact and convex subset of $\mathbb{R}^k$. Then there exists a fixed point of F, so that $x \in F(x)$.*

Another branch of fixed point theorems make more restrictive assumptions than continuity, but with the added benefit of tractable computation of fixed points. We begin with the following definition.

Definition 6.3.6. Let X be a metric space and $T : X \to X$ be a function on X.

(a) T is **Lipschitz continuous** with Lipschitz constant β if there exist $\beta > 0$ such that $d(T(x), T(y)) \leq \beta d(x, y)$.

(b) T is a **contraction mapping** with **modulus** β if T is Lipschitz continuous with a Lipschitz constant $\beta < 1$.

Lipschitz continuous mappings are very well-behaved, and contraction mappings especially so.

Theorem 6.3.7. *If T is Lipschitz continuous, then T is uniformly continuous on X.*

Theorem 6.3.8 (Banach's Fixed Point Theorem *or* **Contraction Mapping Theorem).** *If X is a complete metric space and T is a contraction mapping with modulus β on X, then T has a unique fixed point x^* and for any $y \in X$, $d(T^n(y), x^*) \leq \beta^n d(x, y)$.*

Proof. Let $y \in X$ be arbitrary. We define a sequence (x_n) in X by $x_n = T^n y$.

First, we show (x_n) is a Cauchy sequence. If $n \geq m \geq 1$, then since T is a contraction, we have

$$\begin{aligned} d\left(x_n, x_m\right) &= d\left(T^n x_0, T^m x_0\right) \\ &\leq \beta^m d\left(T^{n-m} x_0, x_0\right) \\ &\leq \beta^m \left[d\left(T^{n-m} x_0, T^{n-m-1} x_0\right) + d\left(T^{n-m-1} x_0, T^{n-m-2} x_0\right) \quad + \cdots + d\left(T x_0, x_0\right)\right] \\ &\leq \beta^m \left[\sum_{k=0}^{n-m-1} \beta^k\right] d\left(x_1, x_0\right) \\ &\leq \beta^m \left[\sum_{k=0}^{\infty} \beta^k\right] d\left(x_1, x_0\right) \\ &\leq \left(\frac{\beta^m}{1-\beta}\right) d\left(x_1, x_0\right) \end{aligned}$$

which implies that (x_n) is Cauchy.

Since X is complete, (x_n) converges to a limit $x^* \in X$. The fact that the limit x^* is a fixed point of T follows from the continuity of T :

$$T x = T \lim_{n \to \infty} x_n = \lim_{n \to \infty} T x_n = \lim_{n \to \infty} x_{n+1} = x.$$

Finally, to show uniqueness, if x^* and y^* were two fixed points then

$$0 \leq d(x^*, y^*) = d(Tx^*, Ty^*) \leq \beta d(x^*, y^*),$$

which is impossible since $\beta < 1$. ∎

Our final fixed-point theorem is more abstract and applies in situations where metrics cannot be defined. Instead, we consider a simpler algebraic structure called a **lattice**. Lattices will come up again in ECON 202 and 203.

Definition 6.3.9. Let X be a set and $\geq$ a partial order defined on X (i.e., a reflexive, transitive and anti-symmetric relation). Set X is called a **lattice** if every pair $x, x' \in X$ has a greatest lower bound and a least upper bound in X, while X is a **complete lattice** if every subset of X has both a greatest lower bound and a least upper bound.

In lattices, the greatest lower bound of x and y is usually called the **meet** and denoted $x \wedge y$ while the least upper bound of x and y is usually called the **join** and denoted $x \vee y$.

Some examples of lattices include any box space $[a, b]^k$ (with the usual order $\leq$) or the power set of any set (ordered by inclusion $\subseteq$).

Theorem 6.3.10 (Tarski's Fixed Point Theorem). *Let $X, \leq$ be a complete lattice and let $f : X \to X$ be a nondecreasing function (i.e. $x \leq x'$ implies $f(x) \leq f(x')$). Then f has a fixed point, and the set of all fixed points is a lattice. This implies there is a least fixed point and a largest fixed point (with respect to $\leq$).*

7

Differentiation

Contents

7.1 Derivatives

You are probably already aware of the definition of the derivative in one dimension as an instantaneous rate of change. In this section, we will generalize this concepts to abstract metric spaces.

Definition 7.1.1. Let $f : \mathbb{R} \to \mathbb{R}$. The **derivative** of f at $x \in \mathbb{R}$ is defined as

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

provided that this limit exists. If f' is defined at every point of a set $E \subseteq \mathbb{R}$, we say that f is **differentiable** on E.

Before we seek to generalize this definition to other metric spaces, let us note some properties of the derivative of single-variable functions.

Theorem 7.1.2. *Let f be defined on $[a, b]$. If f is differentiable at a point $x \in [a, b]$, then f is continuous at x.*

Proof. For any $t \neq x$, rewrite $f(t) - f(x) = \frac{f(t)-f(x)}{t-x} \cdot (t - x)$. Then $\phi(t) = \frac{f(t)-f(x)}{t-x} \to f'(x)$ (since f is differentiable at x), and $t - x \to 0$. Then by Theorem 6.1.4, $f(t) - f(x) \to 0$, and hence by Theorem 6.1.4 again $f(t) \to f(x)$. Thus, f is continuous at x. ■

If f is continuous at x, it may not be differentiable at x. To see that, consider $f(x) = x$ for $x < 1$ and $f(x) = 1$ for $x \geq 1$. f is continuous but not differentiable at $x = 1$.

Definition 7.1.3. If f has a derivative f' on an interval, and f' is itself differentiable, we denote the derivative of f' by f'' and call f'' the second derivative of f. Similarly, in this manner we obtain functions: f, f', f'', $f^{(3)}, \ldots, f^{(n)}$, each of which is the derivative of the preceding one. $f^{(n)}$ is called the nth derivative, of the derivative of order n, of f.
In order for $f^{(n)}(x)$ to exist at point x, $f^{(n-1)}$ must exist in a neighborhood of x, and $f^{(n-1)}$ must be differentiable at x. Since $f^{(n-1)}$ must exist in a neighborhood of x, $f^{(n-2)}$ must be differentiable in that neighborhood.

Theorem 7.1.4. *Suppose f and g are defined on $[a, b]$ and are differentiable at a point $x \in [a, b]$. Then $f + g$, fg and f/g are differentiable at x, and*
(a) $(f + g)'(x) = f'(x) + g'(x)$
(b) $(fg)'(x) = f'(x)g(x) + f(x)g'(x)$
(c) $\left(\frac{f}{g}\right)(x) = \frac{g(x)f'(x)-g'(x)f(x)}{g^2(x)}$, assuming $g(x) \neq 0$.

Proof. (a) follows from Theorem 6.1.4. For (b), let $h = fg$. Then

$$h(t) - h(x) = f(t)[g(t) - g(x)] + g(x)[f(t) - f(x)]$$

Divide this by $t - x$, and note that $f(t) \longrightarrow f(x)$ as $t \longrightarrow x$ (using Theorem 7.1.2), (b) follows. For (c), let $h = f/g$. Then

$$\frac{h(t) - h(x)}{t - x} = \frac{1}{g(t)g(x)}\left[g(x)\frac{f(t) - f(x)}{t - x} - f(x)\frac{g(t) - g(x)}{t - x}\right]$$

Letting $t \longrightarrow x$, and apply Theorem 6.1.4 and previous Theorem 7.1.2 we get (c). ■

Theorem 7.1.5. *Suppose f is continuous on $[a, b]$, $f'(x)$ exists at some point $x \in [a, b]$, g is defined on an interval I which contains the range of f, and g is differentiable at the point $f(x)$. If $h(t) = g(f(t))$, $(a \leq t \leq b)$, then h is differentiable at x, and $h'(x) = g'(f(x))f'(x)$.*

Proof. Let $y = f(x)$. By the definition of the derivative, we have

$$f(t) - f(x) = (t - x)\,[f'(x) + u(t)]$$
$$g(s) - g(y) = (s - y)\,[g'(y) + v(s)]$$

where $t \in [a, b]$, $s \in I$ and $u(t) \to 0$ as $t \to x$, $v(s) \to 0$ as $s \to y$. Let $s = f(t)$. Then

$$\begin{aligned} h(t) - h(x) &= g(f(t)) - g(f(x)) \\ &= [f(t) - f(x)] \cdot [g'(y) + v(s)] \\ &= (t - x)[f'(x) + u(t)] \cdot [g'(y) + v(s)] \end{aligned}$$

or, if $t \neq x$

$$\frac{h(t) - h(x)}{t - x} = [g'(y) + v(s)] \cdot [f'(x) + u(t)]$$

Letting $t \to x$, we have that $s \to y$ by the continuity of f, so that the right-hand side of the above equation tends to $g'(y)f'(x)$, as claimed. ■

We now consider generalizations of the derivative.

Definition 7.1.6. Let V be a Banach space and $f : V \to W$. The **Fréchet derivative** of f at x is a continuous linear[a] mapping $Df[x] : V \to W$ such that

$$\lim_{h \to 0} \frac{\|f(x + h) - f(x) - Df[x](h)\|}{h} = 0.$$

If $V = \mathbb{R}^n$ and $W = \mathbb{R}$, the Fréchet derivative is also called the **differential** or **total derivative**. If $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, the Fréchet derivative is also called the **Jacobian** of f.

[a] In finite-dimensional spaces, linear mappings are always continuous, but in infinite-dimensional spaces, this is not always the case.

Note that the h in the above definition is a vector, which differs from the 1-D definition of the derivative. This might not have been the way you were first taught derivatives in higher dimensions, but we will now see the relationships to the partial derivatives.

Definition 7.1.7. Let $f : V \to W$, $x \in U \subseteq V$ where U is open and $v \in V$. The **Gâteaux derivative** f at x in direction v is

$$df(x; v) = \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h}.$$

Here h is a scalar.

When $V = \mathbb{R}^n$ and $W = \mathbb{R}$, the Gâteaux derivative is also called the **directional**

derivative. If v is one of the coordinate vectors, say x_i, then the Gateaux derivative is just the **partial derivative** with respect to that coordinate, $\frac{\partial f}{\partial x_i}(x)$.

We have the following relationships between the Gâteaux and Fréchet derivatives.

Theorem 7.1.8. *Let $f : V \to W$ be a function.*

(a) If f is Fréchet differentiable at x, then the Gâteaux derivative $df(x;v)$ exists for all $v \in V$ and $df(x;v) = Df[x](v)$.

(b) If f has Gâteaux derivatives that are linear in v and continuous in x in the sense that $\forall \varepsilon > 0$, there exists $\delta > 0$ such that $\|x' - x\| < \delta$ implies

$$\sup_{v \in V} \frac{\|df(x', v) - df(x;v)\|}{\|v\|} < \varepsilon,$$

then f is Fréchet differentiable and the Fréchet derivative satisfies $Df[x](v) = df(x;v)$.

The latter definition implies that Fréchet differentiability is a more stringent requirement than Gâteaux differentiability. A simple example is the 1-D function $f(x) = |x|$ which is Gâteaux differentiable at $x = 0$ but not Frêchet differentiable there.

Definition 7.1.9. Let $f : \mathbb{R}^n \to \mathbb{R}$. The **gradient** of f at x is the vector of partial derivatives

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) & \frac{\partial f}{\partial x_2}(x) & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}'.$$

If f is Fréchet differentiable at x, then $\nabla f(x)$ is a matrix representation of the linear transformation $Df[x]$.

Natural analogies of the sum, product and chain rules hold for the Fréchet derivatives. Here are some convenient examples of Fréchet derivatives.

Example. Here are some examples:

(a) Let $\mathbf{A}$ be an $m \times n$ real matrix and let $f : \mathbb{R}^n \to \mathbb{R}^m$ be $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$. Then $Df[\mathbf{x}] = \mathbf{A}$.

(b) Let $\mathbf{A}$ be an $m \times n$ real matrix and let $f : \mathbb{R}^m \to \mathbb{R}^n$ be $f(\mathbf{x}) = \mathbf{x}'\mathbf{A}$. Then $Df[\mathbf{x}] = \mathbf{A}'$.

(c) Let $\mathbf{A}$ be an $m \times n$ real matrix and let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a differentiable function. Then $D(\mathbf{A}f)[\mathbf{x}] = \mathbf{A}Df[\mathbf{x}]$.

(d) Let $\mathbf{A}$ be an $n \times n$ real matrix and let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$. Then $Df[\mathbf{x}] = \mathbf{x}'(\mathbf{A} + \mathbf{A}')$. If $\mathbf{A}$ is a symmetric matrix, then $Df[\mathbf{x}] = 2\mathbf{x}'\mathbf{A}$.

♣

We now consider higher Fréchet derivatives of a function. Suppose that $f : V \to W$ is differentiable at all points in an open subset $U \subseteq V$. It follows that its derivative, understood now as a function of $x \in U$, is a function from U to the space $L(V, W)$ of all bounded linear operators from V to W, so $Df : U \to L(V, W)$. This function space $L(V, W)$ is a normed vector space (and thus a metric space), with a norm given by

$$\|T\|_{op} = \sup_{v:\|v\|\leq 1} \|Tv\|.$$

The function Df then may also have a derivative, called the **second Fréchet derivative** $D^2 f$ which is a map $D^2 f : U \to L(V, L(V, W))$.

To make it easier to work with second-order derivatives, the space on the right-hand side is identified with the Banach space $L^2(V \times V, W)$ of all continuous bilinear maps from V to W. An element φ in $L(V, L(V, W))$ is thus identified with ψ in $L^2(V \times V, W)$ such that for all $x, y \in V$,

$$\varphi(x)(y) = \psi(x, y).$$

If $f : \mathbb{R}^n \to \mathbb{R}$, the associated matrix of this transformation on $L^2(V \times V, W)$ is called the **Hessian**

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

One may continue on in this fashion, obtaining (as long as the relevant limits exist) the **n-th Fréchet derivatives**, which will be functions

$$D^n f : U \to L^n(V \times V \times \cdots \times V, W)$$

taking values in the Banach space of continuous multilinear maps in n arguments from V to W.

Recursively, a function f is $n + 1$ times differentiable on U if it is n times differentiable on U and for each $x \in U$ there exists a continuous multilinear map A of $n + 1$ arguments such that the limit

$$\lim_{h_{n+1}\to 0} \frac{\|D^n f\,(x + h_{n+1})\,(h_1, h_2, \ldots, h_n) - D^n f(x)\,(h_1, h_2, \ldots, h_n) - A\,(h_1, h_2, \ldots, h_n, h_{n+1})\|}{\|h_{n+1}\|} = 0$$

exists uniformly for $h_1, h_2, \ldots, h_n$ in bounded sets in V. In that case, A is the $(n+1)$st derivative of f at x.

An important result called Schwarz's Theorem (or Clairaut's Theorem or Young's Theorem) tells us that these maps are symmetric.

Theorem 7.1.10. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be twice continuously differentiable. Then*

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

More generally, the nth Fréchet derivatives are symmetric multilinear maps on $L^n(V \times V \times \ldots \times V, W)$, so that permuting the order of inputs has no effect on the output.

Definition 7.1.11. If the k^{th} derivative of function $f : U \to V$ exists and is continuous on U, then we say that f is **k-times continuously differentiable.** The set of k-times continuously differentiable functions from U to V is denoted $C^k(U, V)$. If $U \subseteq \mathbb{R}$ and $V \subseteq \mathbb{R}$, then we just write $f \in C^k$. The set $C^k(U, V)$ is a vector space of functions. A function is called **smooth** if $f \in C^k(U, V)$ for all $k \in \mathbb{N}$.

7.2 Local approximations

Let us return to the real-valued setting. We will introduce some important theorems called the *mean value theorems* which will allow us to derive the important Taylor expansion of a function. To do so, we will need to introduce some machinery regarding local maxima of functions (which will play a very large role later in these notes).

Definition 7.2.1. Let f be a real-valued function defined on a metric space X. We say that f has a **local maximum** at a point $p \in X$ if there exists $\delta > 0$ such that $f(q) \leq f(p)$ for all $q \in X$ with $d(p, q) < \delta$.

Theorem 7.2.2. *Let f be defined on $[a, b]$; if f has a local maximum at a point $x \in (a, b)$, and if $f'(x)$ exists, then $f'(x) = 0$.*

Proof. Pick $\delta > 0$ such that $a < x - \delta < x < x + \delta < b$. If $x - \delta < t < x$, then $\frac{f(t) - f(x)}{t - x} \geq 0$. Letting $t \to x$, we see that $f'(x) \geq 0$. If $x < t < x + \delta$, then $\frac{f(t) - f(x)}{t - x} \leq 0$, which shows that $f'(x) \leq 0$. Hence $f'(x) = 0$. ■

Theorem 7.2.3 (Mean value theorem). *Let f and g be real functions continuous on $[a, b]$ and differentiable in (a, b).*

(a) (Lagrange) There exists an $x \in (a, b)$ such that $f(b) - f(a) = (b - a)f'(x)$.

(b) (Cauchy) There exists an $x' \in (a, b)$ such that $[f(b) - f(a)]g'(x') = [g(b) - g(a)]f'(x')$.

Proof. We prove (b) only, from which (a) follows by setting $g(x) = x$. Let $h(t) = [f(b)-f(a)]g(t)-[g(b) - g(a)]f(t)$, $(a \le t \le b)$. Then h is continuous on $[a, b]$, h is differentiable in (a, b) and $h(a) = f(b)g(a) - f(a)g(b) = h(b)$. It needs to be shown that $h'(x) = 0$ for some $x \in (a, b)$. If h is constant, this holds for every $x \in (a, b)$. If $h(t) > h(a)$ for some $t \in (a, b)$, let x be a point on $[a, b]$ at which h attains its maximum (which exists by Theorem 6.1.12). Observe that $x \in (a, b)$, and then by Theorem 7.2.2, $h'(x) = 0$. If $h(t) < h(a)$ for some $t \in (a, b)$, the same argument applies if we choose for x a point on $[a, b]$ where h attains its minimum. ■

Theorem 7.2.4 (Taylor's theorem). *Suppose f is a real valued function on $[a, b]$, n a positive integer, $f^{(n-1)}$ is continuous on $[a, b]$, $f^{(n)}(t)$ exists for every $t \in (a, b)$. Let $\alpha, \beta \in [a, b]$, $\alpha \neq \beta$, and define*

$$P(t) = \sum_{k=0}^{n-1} \frac{f^{(k)}(\alpha)}{k!}(t - \alpha)^k$$

Then there exists a point $x \in (\alpha, \beta)$ such that

$$f(\beta) = P(\beta) + \frac{f^{(n)}(x)}{n!}(\beta - \alpha)^n$$

When $n = 1$, this is just the mean value theorem. It also gives us the famous interpretation of $f'(x)$ as the (gradient of) the local linear approximation of f at point x. In general, it shows that any $f \in C^n$ can be approximated by a polynomial of degree $n - 1$, and allows us to estimate the error, if we know bounds on $|f^{(n)}(x)|$.

Proof. Let M be the number defined by

$$f(\beta) = P(\beta) + M(\beta - \alpha)^n$$

It needs to be shown that $n!M = f^{(n)}(x)$ for some $x \in (\alpha, \beta)$. Let $g(t) \equiv f(t) - P(t) - M(t - \alpha)^n$ for $a \le t \le b$. Since $P^{(n)}(t) = 0$, we have that

$$g^{(n)}(t) = f^{(n)}(t) - n!M \qquad (a < t < b)$$

Hence, it needs to be shown that $g^{(n)}(x) = 0$ for some $x \in (\alpha, \beta)$.
Observe that $P^{(k)}(\alpha) = f^{(k)}(\alpha)$ for $k = 0, ..., n-1$ and hence

$$g^{(k)}(\alpha) = f^{(k)}(\alpha) - P^{(k)}(\alpha) + \frac{n!}{(n-k)!}M(\alpha - \alpha) = 0 \qquad\qquad k = 0, ..., n-1$$

By the definition of M, $g(\beta) = 0$. By the mean value theorem, there exists $x_1 \in (\alpha, \beta)$ such that $g(\beta) - g(\alpha) = (\beta - \alpha)g'(x_1)$, and hence $g'(x_1) = 0$. Another application of the mean value theorem implies that there exists $x_2 \in (\alpha, x_1)$ such that $g''(x_2) = 0$. Iterating the argument, after n steps we arrive at the conclusion that $g^{(n)}(x_n) = 0$ for some $x_n \in (\alpha, x_{n-1})$. ■

This theorem extends quite generally to Banach spaces.

Theorem 7.2.5 (Taylor's Theorem in Banach Spaces). *If $f \in C^k(U, Y)$, then for all $\mathbf{x} \in U$ and $\mathbf{h} \in X$ such that the line segment $\ell(\mathbf{x}, \mathbf{x}+\mathbf{h}) = \{(1-t)\mathbf{x} + t\,\mathbf{h} : t \in [0,1]\}$ lies in U, there holds*

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + \frac{D^2 f(\mathbf{x})\mathbf{h}^{(2)}}{2!} + \cdots + \frac{D^{k-1}f(\mathbf{x})\mathbf{h}^{(k-1)}}{(k-1)!} + R_k$$

where the remainder R_k satisfies

$$R_k(\mathbf{x}, \mathbf{h}) \leq \frac{\max_{t\in[0,1]} \|D^k f(\mathbf{x}+\mathbf{th})\|}{k!}\|\mathbf{h}\|^{\mathbf{k}}.$$

The proof is by induction using Taylor's Theorem and is omitted.

Using Taylor's Theorem, we can define a few important functions.

Definition 7.2.6. Let $x \in \mathbb{C}$. The **exponential function** e^x or $\exp(x)$ is defined by

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + ... = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

Equivalently (check this!), e^x may be defined by the relation $\frac{d}{dx}e^x = e^x$.

The **sine** function is defined by

$$\sin(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - ... = \sum_{i=0}^{\infty} \frac{(-1)^i x^{2i+1}}{(2i+1)!}$$

while the **cosine** function is defined by

$$\cos(x) = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 - \ldots = \sum_{i=0}^{\infty} \frac{(-1)^i x^{2i}}{(2i)!}.$$

You can check the following facts using these definitions:

$$\begin{aligned} e^{ix} &= \cos(x) + i\sin(x) \\ \frac{d}{dx}\sin(x) &= \cos(x) \\ \frac{d}{dx}\cos(x) &= -\sin(x). \end{aligned}$$

7.3 Implicit and Inverse Function Theorems

Given that the Fréchet derivative gives a local linear approximation to a function at a given point, and linear mappings are easy to invert, knowing the Fréchet derivative should allow us to know a lot about the inverse of a mapping, at least *locally*. That is the idea of two related theorems: the implicit and inverse function theorems.

The implicit function problem asks if $0 = F(x, y)$, can we find a function f such that $y = f(x)$? If F is chosen as $x - g(y)$, then f is just the inverse g^{-1} at x, so that these two problems are closely related.

We have the following theorem

Theorem 7.3.1 (Implicit Function Theorem). *Let X, Y and Z be Banach spaces, and $U \subseteq X \times Y$ an open subset with $F \in C^1(U, Z)$. If $(x_0, y_0) \in U$ is such that $D_y F(x_0, y_0)$ is a bounded invertible map from Y to Z, then there is an open neighborhood V of x_0 and a unique function $f : V \longrightarrow Y$ such that*

$$F(x, f(x)) = F(x_0, y_0), \text{ for all } x \in V.$$

Moreover, $f \in C^1(V, Y)$ with

$$Df(x) = -[D_y f(x, f(x))]^{-1} D_x f(x, f(x)).$$

The inverse function theorem is a corollary of the above where $F(x, y) = x - g(y)$.

Corollary 7.3.2 (Inverse Function Theorem). *Suppose that X and Y are Banach spaces, $U \subseteq X$ an open subset. Let $g \in C^1(U, Y)$ and $x_0 \in X$. If $Dg(x_0)$ has a bounded inverse, then there exists an open neighborhood V of $g(x_0)$ and a unique function $f : U \to V$ such that*

$$g(f(x)) = x, \text{ for all } x \in U.$$

Moreover, f is continuously differentiable, with

$$Df(x) = [Dg(f(x))]^{-1}.$$

8

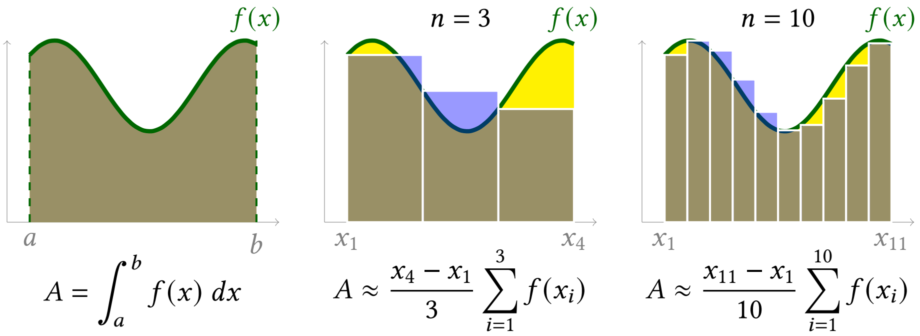
Integration and Measure

Contents

Integration is fundamentally about ways to combine infinitesimal data into a single number like length, area, volume, probability, expectation or other "measures". In this chapter, we begin by reviewing the Riemann Integral—the classic formulation of the integral from undergraduate calculus. We will then talk briefly about the limitations of the Riemann Integral and the mathematical response to these limitations, belonging to a field called *measure theory.* Concepts of measure theory are used all over the place in economics, and yet unfortunately rarely taught as part of the graduate economics core. Our goal here is to give an extremely brief run-down of measure theory, to a level where you will understand, at least, why we might need to worry about problems of measurability occasionally and what someone means when they say they are assuming a function is 'measurable' (you will see this all over the place in economics papers).

8.1 The Riemann Integral

The Riemann integral is the simplest integral to define, and works well for continuous functions and nearly continuous functions. For now consider a positive, bounded function $f : \mathbb{R} \to \mathbb{R}$. The approach of the Riemann integral is to partition the x-axis into a collection of small intervals, and erecting two rectangles above these intervals: the first is the largest rectangle entirely below the graph and one entirely outside the graph area. If by making these intervals sufficiently small, the sum of the areas of the inside rectangles and the outside rectangles approach one another, then that number is called the Riemann Integral of the function.


$f(x)$
a
b
$A = \int_a^b f(x)\, dx$
$n = 3$
$f(x)$
x_1
x_4
$A \approx \frac{x_4 - x_1}{3} \sum_{i=1}^{3} f(x_i)$
$n = 10$
$f(x)$
x_1
x_{11}
$A \approx \frac{x_{11} - x_1}{10} \sum_{i=1}^{10} f(x_i)$

Definition 8.1.1. Let $I = [a, b] \subseteq \mathbb{R}$ be an interval. A **partition** P of I consists of a finite sequences of numbers $(x_i)_{i=1}^n$ such that

$$a = x_0 \leq x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n = b,$$

and we denote the intervals as $I_1 = [x_0, x_1]$, $I_2 = [x_1, x_2]$ and so on. Let $\mathcal{P}$ be the set of all partitions of I.

Write $m_k = \inf_{x \in I_k} f(x)$ and $M_k = \sup_{x \in I_k} f(x)$, which are well-defined since F is bounded.

The **upper and lower Riemann sums** of f with respect to such a partition are then defined as

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

The **upper and lower Riemann integrals** of f on $[a, b]$ are defined as

$$\overline{\int_a^b} f(x) dx = \inf_{P \in \mathcal{P}} U(f; P)$$
$$\underline{\int_a^b} f(x) dx = \sup_{P \in \mathcal{P}} L(f; P)$$

If the upper and lower Riemann integrals are equal, we say that f is **Riemann integrable** and denote the integral as $\int_a^b f(x)dx$ or $\int_I f(x)dx$.

We now determine under what conditions the Riemann integral of a function exists.

Definition 8.1.2. The **oscillation** of a bounded function f on a set A is

$$\operatorname*{osc}_A f = \sup_A f - \inf_A f.$$

If the function $f : [a,b] \longrightarrow \mathbb{R}$ is bounded and $P = \{I_1, I_2, \ldots, I_n\}$ is a partition of $[a,b]$, then

$$U(f;P) - L(f;P) = \sum_{k=1}^{n} \sup_{I_k} f \cdot |I_k| - \sum_{k=1}^{n} \inf_{I_k} f \cdot |I_k| = \sum_{k=1}^{n} \operatorname*{osc}_{I_k} f \cdot |I_k| .$$

So the function$f$ is integrable if the expression on the right can be brought arbitrarily close to zero. This implies that a function is integrable if the oscillation of f on most intervals is nearly zero, and the sum of the lengths of the intervals where the oscilattion is large can be made arbitrarily small. This is sometimes called the **Cauchy criterion** for integrability.

Theorem 8.1.3. *A continuous function on a compact interval is Riemann integrable, as is a monotonic function.*

Here are some simple properties of the Riemann integral.

Theorem 8.1.4. *Let f and g be Riemann integrable function on I. The Riemann integral is*

(a) Linear: $\int_I af(x) + g(x)dx = a\int_I f(x)dx + \int_I g(x)dx$.

(b) Monotone: if $f(x) \leq g(x)$ for all $x \in I$, $\int_I f(x)dx \leq \int_I g(x)dx$.

(c) Additive: if $a < c < b$, then $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$.

(d) Mean Value Theorem for Integrals: there is an $z \in [a,b]$ such that $f(z) = \frac{1}{b-a}\int_a^b f(x)dx$.

We have discussed integration as a measure of area, but another very important property of the integral is that it reverses differentiation.

Theorem 8.1.5 (Fundamental Theorem of Calculus). *Let* $[a, b] \subseteq \mathbb{R}$.

(a) If $F : [a, b] \to \mathbb{R}$ *is continuous on* $[a, b]$ *and differentiable in* (a, b) *with* $F' = f$ *with* $f : [a, b] \to \mathbb{R}$ *Riemann integrable, then*

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

(b) Let $f : [a, b] \to \mathbb{R}$ *be a continuous function and let* $F : [a, b] \to \mathbb{R}$ *be defined by*

$$F(x) = \int_a^x f(t)dt.$$

Then F *is uniformly continuous on* $[a, b]$ *and differentiable on* (a, b) *with* $F'(x) = f(x)$ *for all* $x \in (a, b)$.

Applying the fundamental theorem of calculus gives the famous integration by parts rule (assuming f, g continuous and differentiable, f', g' integrable.

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

So the Riemann Integral has a lot of nice properties. You might ask—why do we need a more complicated version of integration at all? It's a good question, and I will try and give you three possible answers.

The first is that there are functions that cannot be Riemann integrated but for which area "should" be able to be defined. The classic example is the function $f : [0, 1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 0 & \text{if } x \in \mathbb{Q} \\ 1 & \text{otherwise.} \end{cases}$$

We will use the notation $f(x) = 1_{\mathbb{R}} \setminus \mathbb{Q}$ for this function, and more generally write 1_A for the indicator function of set A. It should be clear that f is not Riemann integrable - the lower sums of any partition are always zero and the upper sums are always 1. On the other hand, the rationals are countable and the irrationals are uncountable - there are a *lot* more of them than the rationals, so that the area really "should" be 1.

The second reason is very important conceptually although more subtle. A big problem with the Riemann integral is that you can have a sequence of Riemann integrable functions f_n converging pointwise to a bounded function f (that is, for all x, as a sequence of numbers,

$f_n(x) \to f(x)$) but f may *not* be Riemann integrable. This can make proving facts about limits of functions very difficult.

A third reason is that alternative, more flexible concepts of integration are not that complicated! Okay, I might be cheating a bit with that reason, but I hope I can give you some sense of this in coming sections.

8.2 Measurable sets and spaces

The fundamental difference between the Riemann integral and the Lebesgue integral is the way in which the partitioning is done: instead of partitioning the function's domain, the Lebesgue integral partitions the range and looks at sets of the form $f^{-1}([a,b))$. This sounds like a small difference, but it ends up making a big difference. The sets of the form $f^{-1}([a,b))$ can be very strange, and so the question boils down to how to define the size of these sets. This is where the concept of a measure comes in.

Ideally, we would be able to assign a measure to every set $A \subseteq \mathbb{R}$. Unfortunately it is not possible to do this in such a way that the measure satisfies certain natural properties.

Definition 8.2.1. A **measure space** is a triple $(\Omega, \mathcal{B}, \mu)$ where:

(a) Ω is a set,

(b) $\mathcal{B}$ is a σ-**algebra** of Ω, that is a family of subsets $\mathcal{B} \subseteq 2^\Omega$ such that

   (i) $\emptyset, \Omega \in \mathcal{B}$,

   (ii) closure under complementation: $B \in \mathcal{B}$ implies $\Omega \setminus B \in \mathcal{B}$, and

   (iii) closure under countable union: if for all $n \in \mathbb{N}$, $B_n \in \mathcal{B}$ then $\cup_{n\in\mathbb{N}} B_n \in \mathcal{B}$, and

   Any set in $\mathcal{B}$ is called a $\mathcal{B}$−**measurable set**.

(c) μ is a **measure** on $\mathcal{B}$, that is a nonnegative function $\mu : \mathcal{B} \to \mathbb{R} \cup \{\infty\}$ such that

   (i) Nonnegativity: $\mu(B) \geq 0$ for all $B \in \mathcal{B}$.

   (ii) Null empty set: $\mu(\emptyset) = 0$.

   (iii) Countable additivity: for all countable collections of $\mathcal{B}$-sets, $(B_n)_{n\in\mathbb{N}}$, which are pairwise disjoint, that is, $B_m \cap B_n$ for all $m \neq n$, we have $\mu(\cup_{n\in\mathbb{N}} B_n) = \sum_{n\in\mathbb{N}} \mu(B_n)$.

The pair $(X, \mathcal{B}_X)$ is called a **measurable** space.

A measure space is a **probability space** if $\mu(\Omega) = 1$.

For our notion of measure to correspond to what we think of as length, area or volume in Euclidean spaces, we require that the measure satisfy one more important property.

Definition 8.2.2. Let $\Omega = \mathbb{R}^n$ for some $n \in \mathbb{N}$ and suppose that a $\sigma-$algebra $\mathcal{B}$ is defined on Ω such that if $B \in \mathcal{B}$ then $B_c = \{c + b | b \in B\} \in \mathcal{B}$ for all $c \in \Omega$. Then measure μ on $\mathcal{B}$ is **translation-invariant** if $\mu(B) = \mu(B_c)$ for all $B \in \mathcal{B}$ and $c \in \Omega$.

We now prove the following important theorem.

Theorem 8.2.3. *There is no translation-invariant measure on* $2^{\mathbb{R}}$ *such that* $0 < \mu([0, 1]) < \infty$.

Proof. Suppose μ were such a measure. Define an equivalence relation on $\mathbb{R}$ by $x \sim y$ iff $x - y \in \mathbb{Q}$ and form the Vitali set E by choosing exactly one element from each equivalence class lying in $[0, 1)$ — note that we are using the axiom of choice to define this set. Let $\oplus$ be the addition modulo 1 operator, that is for $x, y \in [0, 1]$,

$$x \oplus y = \begin{cases} x + y & \text{if } x + y \leq 1 \\ x + y - 1 & \text{if } x + y \geq 1. \end{cases}$$

Translation-invariance implies $\mu(E \oplus q) = \mu(E)$. Moreover, $[0, 1) = \cup_{q \in \mathbb{Q} \cap [0,1)} E \oplus q$, so that

$$\mu([0, 1)) = \sum_{q \in \mathbb{Q} \cap [0,1)} \mu(E \oplus q) = \sum_{q \in \mathbb{Q}} \mu(E).$$

Then either $\mu(E) = 0$ so that $\mu([0, 1)) = 0 < \mu([0, 1])$ or $\mu(E) > 0$ so that $\mu([0, 1)) = \infty > \mu([0, 1])$. ■

So what σ-algebra will we use on $\mathbb{R}$? A starting point is the following.

Definition 8.2.4. The **Borel σ-algebra** $\mathcal{B}(\mathbb{R})$ is the smallest $\sigma-$algebra on $\mathbb{R}$ containing all open subset of $\mathbb{R}$ (that is, unions and intersections of sets of the form (a, b), (a, ∞), $(-\infty, b)$ and $(-\infty, \infty)$). Any such set is called a **Borel set**.

Since the Borel sets are a σ-algebra, the complement of any Borel set is Borel, as is the countable union of any Borel sets. We would like to define a measure on $\mathcal{B}(\mathbb{R})$, but doing so seems like a difficult task: some of the sets in $\mathcal{B}(\mathbb{R})$ are very strange, so defining an intuitive notion of length seems challenging.

Fortunately, it suffices for defining the **Lebesgue measure** to begin with the intuitive definition $\mu((a,b)) = b - a$ for $a < b$. We then carefully build up the definition of μ on other sets $\mathcal{B}(\mathbb{R})$ so as to maintain consistency with the definition of a measure. It turns out via the **Caratheodory extension theorem** that it is possible to find a measure on all of $\mathcal{B}(\mathbb{R})$ which corresponds to this definition, and the **Hahn extension theorem** guarantees us that this extension is unique. For any $B \in \mathcal{B}(\mathbb{R})$ the Lebesgue measure is merely

$$\inf \left\{ \sum_{A \in C} \mu(A) : C \text{ is a collection of open intervals whose union covers } B \right\}.$$

There is one more intuitive property that we would like our measure to satisfy. Given a measure space, one would expect that if $\mu(B) = 0$, and $A \subseteq B$ that $\mu(A) = 0$ as well. However, for B in the Borel σ-algebra, the subset A may not even be a Borel set. This gap is filled by the completion of a sigma-algebra. The idea is as follows.

Definition 8.2.5. Let $(\Omega, \mathcal{B}, \mu)$ be a measurable space and let

$$\mathcal{C} = \{C \subseteq \Omega : C \subseteq A \text{ for } A \in \mathcal{B} \text{ with } \mu(A) = 0\}.$$

The **completion** of $\mathcal{B}$ is the family $\mathcal{B}'$ constructed by adding and subtracting members of $\mathcal{C}$ to sets in $\mathcal{B}$, that is

$$\mathcal{B}' = \{B' \subseteq S : B' = (B \cup C_1) \setminus C_2 \text{ for } C_1, C_2 \in \mathcal{C}\}.$$

The completion of the Borel σ-algebra is called the **Lebesgue σ-algebra** $\mathcal{L}(\mathbb{R})$.

The Lebesgue measure can be extended (by a theorem also known as the Caratheodory extension theorem) to all sets in $\mathcal{L}(\mathbb{R})$, basically by setting $\mu(B) = \mu(B')$ for sets B, B' as in the definition of completion above. This extension is still unique by the Hahn extension theorem.

In higher-dimensional Euclidean spaces, the definitions and results are analogous: start by defining the volume of open-boxes in the intuitive way and take the same approach to build up to measure of the Borel and Lebesgue sets.

8.3 Measurable functions and the Lebesgue integral

Recall that the motivation of defining the Lebesgue measure was to allow us to extend our definition of integration to a wider family of functions. The question is now—how wide?

Definition 8.3.1. Let $(X, \mathcal{B}_X)$ and $(Y, \mathcal{B}_Y)$ be measurable spaces and $f : X \to Y$. The function f is $(\mathcal{B}_X, \mathcal{B}_Y)$-**measurable** if for all $A \in \mathcal{B}_Y$, $f^{-1}(A) \in \mathcal{B}_X$. If $(X, \mathcal{B}_X, \mu)$ is a probability space, then any measurable f is called a **random variable**.

For real-valued functions $f : \mathbb{R} \to \mathbb{R}$, typically we take $\mathcal{B}_X = \mathcal{L}(\mathbb{R})$ or $\mathcal{B}(\mathbb{R})$ and $\mathcal{B}_Y = \mathcal{B}(\mathbb{R})$. The reason for this slight inconsistency is that the composition of $(\mathcal{L}, \mathcal{B})$-measurable functions may not be measurable, whereas the composition of $(\mathcal{B}, \mathcal{B})$-measurable functions are always measurable.

Definition 8.3.2. A **simple function** $s : X \to \mathbb{R}$ is any function of the form

$$s = c_1 \chi_{E_1} + c_2 \chi_{E_2} + \cdots + c_n \chi_{E_n}$$

where $c_1, \ldots, c_n \in \mathbb{R}$ are distinct and $\{E_1, \ldots, E_n\}$ is a partition of X into nonempty measurable sets. The **Lebesgue integral of a simple function** s is defined as

$$\int_X s d\mu = \sum_{i=1}^{n} c_i \mu(E_i).$$

Here the convention $0 \cdot \infty = \infty \cdot 0 = 0$ is adopted.

Simple functions are great and all, but the following theorems capture the full power of the Lebesgue integral.

Theorem 8.3.3. *Let $(X, \mathcal{B}_X)$ be a measurable space .*

(a) Suppose $\{f_n\}_{n\in\mathbb{N}}$ with $f_n : X \to \mathbb{R}$ is a sequence of $(\mathcal{B}_X, \mathcal{L}(\mathbb{R}))$-measurable functions converging pointwise to f, that is $f(x) = \lim_{n\to\infty} f_n(x)$. Then the pointwise limitf is also $(\mathcal{B}_X, \mathcal{L}(\mathbb{R}))$-measurable.

(b) If f is a $(\mathcal{B}_X, \mathcal{L}(\mathbb{R}))$- measurable function, there exists a sequence of simple functions $s_n : X \to \mathbb{R}$ such that $s_n \to f$ pointwise. Moreover, if $f \geq 0$, this sequence s_n may be chosen to be pointwise nondecreasing, while if f is bounded, then s_n may be chosen so that $s_n \to f$ uniformly on X.

The Lebesgue integral of measurable functions is defined using the above theorem.

Definition 8.3.4. Let $\mathcal{M}_+(X, \mathcal{B}_X)$ be the space of nonnegative-valued $(\mathcal{B}_X, \mathcal{B}(\mathbb{R}))$–measurable functions and $\mathcal{M}(X, \mathcal{B}_X)$ be the space of all measurable functions.

The **Lebesgue integral** of $f \in \mathcal{M}_+(X, \mathcal{B}_X)$ is defined as

$$\int_X f(x)d\mu(x) = \sup \int_X \phi(x)d\mu(x),$$

where the supremum is taken over all simple functions $\phi \in \mathcal{M}_+(X, \mathcal{B}_X)$ with $0 \leq \phi \leq f$. The Lebesgue integral of $f \in \mathcal{M}(X, \mathcal{B}_X)$ is defined by letting $f_+(x) = \max\{f(x), 0\}$ and $f_-(x) = -\min\{f(x), 0\}$. Say that f is **integrable** if the Lebesgue integral of both f_+ and f_- exist and are finite, and define

$$\int_X f(x)d\mu(x) = \int_X f_+(x)d\mu(x) - \int_X f_-(x)d\mu(x).$$

So that is how to define an integral properly. We now state a variety of useful results about the Lebesgue integral.

Definition 8.3.5. Let $(X, \mathcal{B}_X, \mu)$ be a measure space. A set $B \in \mathcal{B}_X$ is a **null set** if $\mu(B) = 0$. A property is said to hold **almost everywhere** (or just a.e.) if the set of points where the property does not hold is a null set.

Theorem 8.3.6 (Null sets and measure). *Suppose that f is a measurable function and that $g = f$ almost everywhere. Then $\int g d\mu = \int f d\mu$.*

Theorem 8.3.7 (Riemann and Lebesgue integrals). *Suppose that $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then f is Lebesgue integrable on $[a, b]$ and these integrals coincide.*

Theorem 8.3.8 (Monotone Convergence Theorem). *Let $\{f_n\}_{n\in\mathbb{N}}$ be a monotone increasing (a.e.) sequence of functions in $\mathcal{M}_+(X, \mathcal{B}_X)$ converging pointwise (a.e.) to f then*

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu.$$

Theorem 8.3.9 (Fatou's Lemma). *Let $\{f_n\}$ be a sequence of functions in $\mathcal{M}_+(X, \mathcal{B}_X)$, then*

$$\int \liminf f_n d\mu \leq \liminf \int f_n d\mu.$$

Theorem 8.3.10 (Lebesgue's Dominated Convergence Theorem)**.** *Let $(X, \mathcal{B}_X, \mu)$ be a measure space and let $\{f_n\}_{n\in\mathbb{N}}$ converge pointwise (a.e.) to a measurable function f. If there exists an integrable function g such that $|f_n| \leq g$ for all n, then f is integrable and*

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu.$$

Definition 8.3.11. Let $(X, \mathcal{B}_X)$ be a measurable space.

(a) Measure μ on $(X, \mathcal{B}_X)$ is σ**-finite** if $\mu(X) < \infty$.

(b) If λ and μ are both σ-finite measures on $(X, \mathcal{B}_X)$, say that λ is **absolutely continuous** with respect to μ, written $\lambda << \mu$ if, for all $B \in \mathcal{B}_X$, $\mu(B) = 0$ implies $\lambda(B) = 0$.

Theorem 8.3.12 (Radon-Nikodym Theorem)**.** *Let λ and μ be σ-finite measures on $(X, \mathcal{B}_X)$ and suppose that $\lambda << \mu$. Then there is an integrable function h such that for all $B \in \mathcal{B}_X$,*

$$\lambda(B) = \int_X h(x) d\mu(x).$$

The function is unique in the sense that if g also has this property, $g = h$ almost everywhere. The function h is called the ***Radon-Nikodym derivative*** *of λ with respect to μ and is written $\frac{d\lambda}{d\mu}$.*

Definition 8.3.13. Let $S \subseteq \mathbb{R}^m$ and $T \subseteq \mathbb{R}^n$ and $F : S \rightrightarrows T$. A **measurable selection** of F is a measurable function $h : S \to T$ with $h(x) \in F(x)$ for all $x \in S$.

Theorem 8.3.14 (Measurable selection theorem)**.** *Let $S \subseteq \mathbb{R}^m$ and $T \subseteq \mathbb{R}^n$, and suppose that $F : S \rightrightarrows T$ is a nonempty, compact-valued and upper hemicontinuous correspondence. Then F has a measurable selection.*

Theorem 8.3.15 (Product Measure Theorem)**.** . *Let $(X, \mathcal{B}_X, \mu)$ and $(Y, \mathcal{B}_Y, \lambda)$ be measure spaces. Then there exists a measure π on $(X \times Y, \mathcal{B}_X \times \mathcal{B}_Y)$ such that $\pi(A \times B) = \mu(A)\lambda(B)$ for $A \in \mathcal{B}_X$ and $B \in \mathcal{B}_Y$. Moreover, if μ and λ are σ-finite, then π is unique and σ-finite.*

Definition 8.3.16. Let $E \subseteq Z = X \times Y$. An x**-section** of E is the set $E_x = \{y \in Y \mid (x, y) \in E\}$. A y-section is $E^y = \{x \in X \mid (x, y) \in E\}$. Let $f : Z \to [-\infty, \infty]$ and $x \in X$. The x-section of f is $f_x(y) = f(x, y)$. For $y \in Y$, the y-section of f is $f^y(x) = f(x, y)$.

Theorem 8.3.17 (Fubini'S Theorem)**.** *. Let $(X, \mathcal{B}_X, \mu)$ and $(Y, \mathcal{B}_Y, \lambda)$ be σ-finite and let $\pi = \mu \times \lambda$. If F is integrable with respect to π on $Z = X \times Y$, then the extended real valued functions defined almost everywhere by $f(x) = \int_Y F_x d\lambda$ and $g(y) = \int_X F^y d\mu$ have finite integrals and $\int_X f d\mu = \int_Z F d\pi = \int_Y g d\lambda$. That is to say,*

$$\int_X \left(\int_Y F(x,y) d\lambda(y) \right) d\mu(x) = \int_Z F d\pi = \int_Y \left(\int_X F(x,y) d\mu(x) \right) d\lambda(y)$$

8.4 Line integrals

An important kind of integral you will encounter a few times in the core economics sequence is the line (or path) integral. We will very briefly cover it.

Definition 8.4.1. Let $\mathbf{r} : [0,1] \longrightarrow \mathbb{R}^n$ parametrize a curve C in $\mathbb{R}^n$ which is differentiable almost everywhere and let $F : U \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a vector field. The line integral of F in the direction of $\mathbf{r}$ is defined as

$$\int_C F(\mathbf{r}) \cdot d\mathbf{r} = \int_0^1 F(\mathbf{r}(t)) \cdot \mathbf{r}'(t) dt,$$

if this integral exists. If C is closed, the integral is often written with the sign $\oint_C$.

The vector $\mathbf{r}'(t)$ is a tangent vector to $\mathbf{r}(t)$, so that the dot product on the right-hand-side of the definition is the projection of F in the direction along the curve. The line integral is independent of the parametrization of the curve (i.e., if $s(t)$ plots out the same curve as $r(t)$, the line integrals are the same), up to orientation.

Two important facts about the line integral follow.

Theorem 8.4.2 (Fundamental theorem of line integrals)**.** *Let F be a **conservative vector field**, that is, there exists a function $G : U \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}$ such that $F = \nabla G$. Then the line integral is **path-independent**, that is*

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = G(\mathbf{r}(1)) - G(\mathbf{r}(0)),$$

for all **r** *differentiable almost everywhere.*

Theorem 8.4.3 (Green's Theorem)**.** *Let C be a counterclockwise oriented, differentiable a.e. closed curve in the plane and let D be the region bounded by C. Then*

$$\oint_C F \cdot d\mathbf{r} = \iint_D \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} dxdy.$$

Green's Theorem is a special case of the very powerful (generalized) Stokes' Theorem, which is occasionally used in mechanism design (and perhaps elsewhere in economics?). Note that Green's Theorem implies that a vector field in the plane is conservative if $\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} = 0$ (in fact, this is an if and only if statement!).

8.5 Counting and combinatorics

We now cover a very important special case of a probability measure, which applies to finite and countable spaces. We have the following result about such spaces.

Theorem 8.5.1. *Let X be a finite or countable set and $\mathcal{B}_X = 2^X$. Then each measure on $\mathcal{B}_X$ is of the form*

$$\mu(A) = \sum_{x \in A} p(x),$$

for some function $p : S \longrightarrow [0, \infty]$.

Definition 8.5.2. Let X be a finite set and $\mathcal{B}_X = 2^X$. The **uniform measure** on $(X, \mathcal{B}_X)$ is defined as

$$\mu(A) = \frac{|A|}{|X|}.$$

The uniform measure can be interpreted as the probability measure associated with events that are equally likely.

We now recap some basic principles of counting (thanks to Joe Romano for notes on these).

First basic counting principle

If there are m elements $a_1, \ldots, a_m$ in one groups and n elements $b_1, \ldots, b_n$ in another, there are mn possible pairs (a_i, b_j) containing one element from each group. If there are r groups with the i th group containing n_i elements, then there are a total of $n_1 \cdots n_r$ total possible r-tuples.

Example: how many license plates numbers possible if the first 3 places must be letters and the final 4 numbers: $26^3 10^4$

Second basic counting principle

Given a population of n elements, there are n^r different samples when sampling r from the population with replacement (which follows from first principle). When sampling without replacement, there are

$$(n)_r = n(n-1)\cdots(n-r+1).$$

In the special case $r = n$, each sample represents a reordering or permutation of the population elements, and there are n ! such orderings. eg. batting orders in baseball.

Example (Chavalier de Méré's gambling problem) Suffering some gambling losses, the following problem was posed to Blaise Pascal in 1654 . Since the probability of getting a 6 in a toss of a fair die is $1/6$, de Méré thought the probability of at least one ace in four tosses is $(1/6) \cdot 4$. Then, knowing that the chance of a double six in a toss of a pair of dice is $1/36$, or a sixth as likely as in one toss, he reasoned that he needed to toss the pair 24 times so that the probability of at least one double 6 is the same as at least one 6 in four tosses. YIKES. For the first, there are 6^4 possible outcomes of tossing a die four times and 5^4 possible outcomes with no sixes. The probability of at least one 6 in four tosses is $1-(5/6)^4 \approx 0.5177$, while the probability of at least one double 6 in 24 tosses is $1-(35/36)^{24} \approx 0.4914$.

Example (The classical birthday problem) Given a room with n people, what is the probability, p_n, that at least two share the same birthday?

$$p_n = 1 - \frac{365 \cdot 364 \cdots (365-n+1)}{365^n}.$$

Note $p_{23} = 0.506$. Using $\log(1+x) \approx x$, we can approximate p_n :

$$\begin{aligned}\log(1-p_n) &= \sum_{i=1}^{n-1} \log(1-i/365) \approx \sum_{i=1}^{n-1} -i/365 \\ &= \frac{-1}{365} \cdot n(n-1)/2\end{aligned}$$

2 So,

$$p_n \approx 1 - \exp\left[\frac{-n(n-1)}{2 \cdot 365}\right]$$

Suppose you ask people, one by one, their birthday. On average, how many people must be sampled to get a duplicate? The solution, to come later, is

$$1 + \frac{364}{365} + \frac{364 \cdot 363}{365^2} + \cdots + \frac{364 \cdot 363 \cdots 2 \cdot 1}{365^{364}}$$
$$\approx \sqrt{\frac{\pi \cdot 365}{2}} - \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2 \cdot 365}}$$

Example (Flag displays) Suppose r flags of different colors are to be shown on n poles in a row. How many ways can this be done (disregarding the absolute position of the flags on the poles and limitations of number of flags on any one pole)? For the first flag, there are n choices of poles. For the next, there are $n + 1$ (either of the remaining $n - 1$ poles, or above or below the first flag). Then, there are $n + 2$ choices for the 3rd, etc. So, there are $n(n + 1) \cdots (n + r - 1)$ different possible displays.

Third basic counting principle For $r \leq n$,

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!},$$

is the number of possible combination of n objects taken r at a time (without regard to ordering).

Example (Poker hands) First, how many poker hands are there? The chance of four of a kind is

$$(13)(48)/\binom{52}{5} = 0.00024$$

Full house is

$$\frac{13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2}}{\binom{52}{5}} = 0.00144$$

Example (Mega Millions Lottery) Pick 5 out of 56 (white balls) and 1 from 46 (the mega ball number in gold). Number of picks is $\binom{56}{5} \cdot 46 = 175, 711, 536$

Example (Indistinguishable flag displays) Suppose all of the r flags are of the same color. How many ways can they be displayed on n poles? If you number the flags from 1 to r, they become distinguishable and there are $N = n(n+1)\cdots(n+r-1)$ ways, from earlier. If they are not distinguishable, there are $N/r!$ ways, or

$$\begin{pmatrix} n+r-1 \\ r \end{pmatrix}$$

(How many distinguishable distributions where no flagpole is empty? $\begin{pmatrix} r-1 \\ n-1 \end{pmatrix}$. Why? Change r to $r-n$ because need one flag on each pole to begin.) Hence, consider the positive integers r_i which satisfy

$$r_1 + r_2 + \cdots r_n = r,$$

which denotes a possibly configuration of occupancy numbers when placing r indistinguishable balls into n cells. The number of distinguishable distributions (or solutions of above equation) is

$$\begin{pmatrix} n+r-1 \\ r \end{pmatrix}.$$

Example (Investments) If you invest 25 K among 4 investments, where each investment must be in multiples of 1 K, how many strategies are possible? If not all money needs to be invested? First, $\begin{pmatrix} 28 \\ 3 \end{pmatrix}$. Adding a 5 th investment into a reserve yields for the second $\begin{pmatrix} 29 \\ 4 \end{pmatrix}$.

Example (Hypergeometric probabilities) A committee of 5 is selected from 6 men and 9 women. If selection is random, what is the probability the committee contains exactly 3 men and 2 women?

$$\begin{pmatrix} 6 \\ 3 \end{pmatrix}\begin{pmatrix} 9 \\ 2 \end{pmatrix}/\begin{pmatrix} 15 \\ 5 \end{pmatrix} = 240/1001$$

More generally, an urn contains n balls, of which r are red and $n-r$ are white. Let X by the number of red balls drawn taking m without replacement. Then,

$$P\{X=k\} = \begin{pmatrix} r \\ k \end{pmatrix}\begin{pmatrix} n-r \\ m-k \end{pmatrix}/\begin{pmatrix} n \\ m \end{pmatrix} k = \max(0, m-(n-r)), \ldots, \min(m, r).$$

(Why the $\max(0, m-(n-r))$ term? If you take all the $n-r$ whites, there are still $m-(n-r)$ remaining.) As an example, suppose you capture, tag and release 10 animals. Later, you capture

20 , and let X be the number of tagged animals out of the 10 . If there are n "alive", then

$$P(X=k)=\left(\begin{array}{c}10\\k\end{array}\right)\left(\begin{array}{c}n-10\\20-k\end{array}\right)/\left(\begin{array}{c}n\\20\end{array}\right).$$

If you actually observe $X=4$, you can estimate n as the value maximizing

$$\left(\begin{array}{c}10\\4\end{array}\right)\left(\begin{array}{c}n-10\\16\end{array}\right)/\left(\begin{array}{c}n\\20\end{array}\right)$$

the value is $\hat{n}=50$.

Here are some important facts to know about $\binom{n}{k}$.

Theorem 8.5.3. *The binomial coefficient $\binom{n}{k}$ satisfies the following:*

(a) Recursive formula: $\binom{n}{k}=\binom{n-1}{k-1}+\binom{n-1}{k}$ for all $1\leq k\leq n-1$.

(b) Symmetry: $\binom{n}{k}=\binom{n}{n-k}$.

(c) Binomial formula: $(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^ky^{n-k}$.

(d) Sums: $\sum_{k=0}^{n}\binom{n}{k}=2^n$.

(e) Stirling's Formula: $n!\sim\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ (by $\sim$ here, I mean the ratio of the two sides tends to 1 as $n\longrightarrow\infty$).

Back to measure theory...

The counting measure is an example of a **discrete measure** on $\mathbb{R}$, which is a sequence of numbers $\{s_n\}_{n\in\mathbb{N}}$ such that $\mu(\mathbb{R}\setminus\{s_n\}_{n\in\mathbb{N}})=0$. That is, the measure is **concentrated** on a (countable) sequence of numbers. These points are called the **atoms** of the measure. Discrete measures are important to understand given the following important theorem due to Lebesgue.

Theorem 8.5.4. *Let μ be a measure on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$. Then*

$$\mu=\mu_{ac}+\mu_{sc}+\mu_d,$$

*where μ_{ac} is absolutely continuous with respect to the Lebesgue measure, μ_d is a discrete measure and μ_{cs} is a **singular continuous measure**, that is, μ_{cs} is zero on all points $x\in\mathbb{R}$ and zero on the complement of some set B of Lebesgue measure zero.*

Part IV

Static Optimization

9

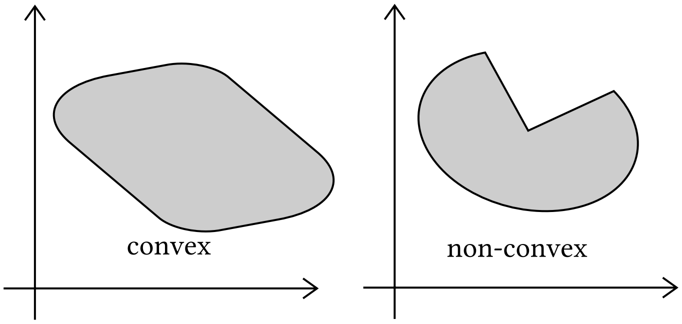
Convexity

Contents

9.1 Convex sets

We briefly introduced convexity earlier during our discussion of fixed point theorems, but now we will approach the topic in more detail because of its importance to the study of optimization problems.

Definition 9.1.1. A set $X \subseteq \mathbb{R}^n$ is **convex** if for any $\mathbf{x}, \mathbf{y} \in X$, we have $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in X$ for all $\lambda \in (0,1)$. It is **strictly convex** if for any $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \neq \mathbf{y}$ and $\lambda \in (0,1)$, we have $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \text{int}X$.



Example. The preference relation $\succeq$ on X is convex if for every $\mathbf{x} \in X$ the upper contour set $\{\mathbf{y} \in X : \mathbf{y} \succeq \mathbf{x}\}$ is convex, i.e. if $\mathbf{y} \succeq \mathbf{x}$ and $\mathbf{z} \succeq \mathbf{x}$, then $\lambda\mathbf{y} + (1-\lambda)\mathbf{z} \succeq x$ for every $\lambda \in (0,1)$. ♣

Theorem 9.1.2. *If $X \subseteq \mathbb{R}^n$ is a convex set, then $\sum_{i=1}^k \lambda_i \mathbf{x}_i \in X$ for any $\mathbf{x}_1, \ldots, \mathbf{x}_k \in X$, $\lambda_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^k \lambda_i = 1$.*

Proof. It follows by a standard induction argument from the definition of convexity (where $k = 2$), inducting to the number k. ∎

Theorem 9.1.3. *The intersection of any number of convex sets is convex.*

Non-convex sets can be made convex by filling in any holes.

Definition 9.1.4. The **convex hull** of a set $X \subseteq \mathbb{R}^n$, denoted $\mathrm{co}(X)$ is the smallest convex set containing X. The **closed convex hull** of a set $X \subseteq \mathbb{R}^n$, denoted $\overline{\mathrm{co}}(X)$ is the smallest closed-and-convex set containing X.

Theorem 9.1.5 (Caratheodory's Theorem)**.** *Let $X \subseteq \mathbb{R}^n$. Every vector in $co(X)$ can be written as the convex combination of at most $n + 1$ points in X.*

Theorem 9.1.6. *For any set $X \subseteq \mathbb{R}^n$,*

$$co(X) = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \sum_{i=1}^k \lambda_i \mathbf{x}_i, \ \text{where } \mathbf{x}_i \in X, \ \lambda_i \in [0,1], \ i = 1, \ldots, k, \ \text{and} \ \sum_{i=1}^k \lambda_1 = 1\}$$

9.2 Separating and Supporting Hyperplanes

In this section, we develop the famous separating and supporting hyperplane theorems that are very important in economic theory.

Definition 9.2.1. Fix a Euclidean space $\mathbb{R}^n$ and two sets X, Y in $\mathbb{R}^n$.

(a) A **hyperplane** is a set of the form $\{x : p \cdot x = c\}$ for some $p \neq 0$ and $c \in \mathbb{R}$.

(b) A hyperplane **separates** X and Y if X and Y lie on different sides of the hyperplane, that is $p \cdot x \geq c$ for all $x \in X$ and $p \cdot y \leq c$ for all $y \in Y$. In that case, we say that X and Y **can be separated**.

(c) **Strict separation** requires that X and Y lie in the open half-spaces, that is, that the inequalities above can be replaced by their strict versions. That is $p \cdot x > c$ for all $x \in X$ and $p \cdot y < c$ for all $y \in Y$.

(d) **Strong separation** requires that X and Y lie in different half-spaces, namely $p \cdot x > c_2$ and $p \cdot y < c_1$ where $c_1 < c_2$.

Theorem 9.2.2 (Separating Hyperplane Theorems)**.** *Let X and Y be two nonempty convex subsets of $\mathbb{R}^n$.*

(a) Sets X and Y may be separated if and only if $0 \notin \operatorname{int}(X - Y)$. This holds, in particular, if X and Y are disjoint.

(b) If, in addition, neither X nor Y contains a half-line in its boundary, i.e., a set of points $\{x+\lambda y : \lambda \geq 0\}$, then X and Y may be strictly separated.

(c) X and Y may be strongly separated if and only if $0 \notin \overline{X - Y}$. This holds, in particular, if X and Y are disjoint, with both sets closed and one of them bounded (compact).

By taking points on the boundary of a convex set and applying the separating hyperplane theorem between the point and the set, we obtain the following important result.

Theorem 9.2.3 (Supporting Hyperplane Theorem)**.** *Let X be a nonempty convex subset of $\mathbb{R}^n$ and let x_0 be on the boundary of X. Then there exists a nonzero vector $p \in \mathbb{R}^n$ and $c \in \mathbb{R}$ such that $p \cdot x_0 = c$ and $p \cdot x \leq c$ for all $x \in X$.*

*Moreover, if X is closed and we define the **support function** $\phi_X : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$ by*

$$\phi_X(p) = \inf\{c \in \mathbb{R} : p \cdot x \leq c \text{ for all } x \in X\},$$

we have that

$$X = \bigcap_{p \in \mathbb{R}^n} \{x \in \mathbb{R}^n : p \cdot x \leq \phi_X(p)\}.$$

Exercise 9.1. *Prove the following simple case of the separating hyperplane theorem: let $x \in \mathbb{R}^m$ and $Y \subset \mathbb{R}^m$ be a convex set with $x \notin Y$. Show that $\{x\}$ and Y may be separated.*

9.3 Convex functions

Definition 9.3.1. Let $X \subseteq \mathbb{R}^n$ be a convex set. A function $f : X \longrightarrow \mathbb{R}$ is **convex** if for any $\mathbf{x}, \mathbf{y} \in X$ and any $\lambda \in (0, 1)$

$$f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})$$

If the inequality is strict whenever $\mathbf{x} \neq \mathbf{y}$, f is said to be **strictly convex**.

A function $f : X \longrightarrow \mathbb{R}$ is **concave** if for any $x, y \in X$ and $\lambda \in (0, 1)$,

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y).$$

Strict concavity is defined similarly to strict convexity.

Example. Consider $f(x) = x$ versus $f(x) = x^2$; the first is only convex while the second one is strictly convex. ♣

Theorem 9.3.2 (Equivalent characterizations of convexity)**.** *A function $f : X \longrightarrow \mathbb{R}$ is convex if and only if:*

(a) the function $g : \mathbb{R} \longrightarrow \mathbb{R}$ given by $g(t) = f(x + ty)$ is convex for all $(x, y) \in X \times \mathbb{R}$ and t such that $x + ty \in X$.

(b) its ***epigraph****,* $\text{epi}(f) = \{(x, y) \in X \times \mathbb{R} | f(x) \leq y\}$ *is a convex set*

First-order characterization of convexity*: If f is differentiable, then f is convex if and only if X is convex and $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$ for all $x, y \in X$.*

Second-order characterization of convexity*: If f is twice differentiable, then f is convex if and only if X is convex and its Hessian $D^2 f(x)$ is positive semi-definite for all $x \in X$.*

Each of these theorems have an analogue for concavity, namely in (a), each such function must be concave, in (b), the hypograph (defined with the reversed inequality sign) must be convex, in (c), the inequality is reversed while in (d), the Hessian must be negative semi-definite.

Example. Here are some examples of functions that are convex or concave:

- Exponential: e^{ax} is convex on $\mathbb{R}$ for any a.
- Powers: x^a is convex on $\mathbb{R}_{++}$ when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$.
- Powers of absolute value: $|x|^a$ is convex on $\mathbb{R}$ for $a \geq 1$.

- Logarithm: $\log(x)$ is concave on $\mathbb{R}_{++}$.
- Negative entropy: $x\log(x)$ is convex on $\mathbb{R}_{++}$.
- Norms: every norm is convex on $\mathbb{R}^n$:

$$||\lambda\mathbf{x} + (1-\lambda)\mathbf{y}|| \leq ||\lambda\mathbf{x}|| + ||(1-\lambda)\mathbf{y}|| = \lambda||\mathbf{x}|| + (1-\lambda)||\mathbf{y}||$$

- Max: $f(\mathbf{x}) = \max\{x_1, ..., x_n\}$ is convex on $\mathbb{R}^n$.
- Quadratic-over-linear: $f(x, y) = \frac{x^2}{y}$ defined on $\mathbb{R} \times \mathbb{R}_{++}$ is convex.
- Log-sum-exp: $f(\mathbf{x}) = \log\left(e^{x_1} + ... + e^{x_n}\right)$ on $\mathbb{R}^n$ is convex.
- Geometric mean: $f(\mathbf{x}) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{2}}$ is concave on $\mathbb{R}^n_{++}$.
- Log-determinant: $f(\mathbf{X}) = \log\det(\mathbf{X})$ is concave on S^n_{++} (symmetric positive definite matrices).

♣

Theorem 9.3.3. *Here are some operations that preserve convexity:*

- *Nonnegative weighted sum: $f = w_1 f_1 + ... + w_m f_m$ is convex if f_1,...,f_m are convex and $w \geq 0$.*
- *Precomposition with affine mappings: $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$ is convex if f is convex.*
- *Pointwise maximum and supremum: $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$ is convex if f_1 and f_2 are convex. If for each $\mathbf{y} \in A$, $f(\mathbf{x}, \mathbf{y})$ is convex, then $g(\mathbf{x}) = \sup_{y \in A} f(\mathbf{x}, \mathbf{y})$ is convex.*
- *Scalar composition: Suppose $h : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$ and $f(x) = h(g(x))$. Asumme that both h and g are twice differentiable. Then*

$$f''(x) = h'(g(x))g''(x) + h''(g(x))\left(g'(x)\right)^2$$

So, $f'' \geq 0$ (i.e. f is convex) if either $h'' \geq 0$, $h' \geq 0$ and $g'' \geq 0$ (i.e. h is nondecreasing and convex, and g is convex), or $h'' \geq 0$, $h' \leq 0$ and $g'' \leq 0$ (i.e.h is nonincreasing and convex, and g is concave).

- *Vector composition: Now suppose $f(x) = h(g(x)) = h(g_1(x), ..., g_k(x))$, where $h : \mathbb{R}^k \to \mathbb{R}$,*

$g_i : \mathbb{R} \to \mathbb{R}$. Then

$$f''(x) = g'(x)^T D^2 h(g(x)) g'(x) + \nabla h(g(x))^T g''(x)$$

So f is convex if h is convex, nondecreasing in each argument, and g_i is convex for $i = 1, \ldots, k$, or if h is convex, nonincreasing in each argument, and g_i is concave for $i = 1, \ldots, k$.

- *Minimization: If f is convex in (x, y), C is nonempty convex set, then $g(x) = \inf_{y \in C} f(x, y)$ is convex.*

We now discuss some nice consequences of convexity/concavity. The first is a very important inequality you must know!

Theorem 9.3.4. *Any convex function f satisfies Jensen's inequality*

$$f\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{n} \lambda_i f(\mathbf{x}_i) \qquad\qquad \sum_{i=1}^{n} \lambda_i = 1, \;\; \lambda_i \geq 0 \text{ for } i = 1, \ldots, n$$

Concave functions satisfy the inequality with a $\geq$.

Convex functions are very nicely behaved in terms of continuity and differentiability.

Theorem 9.3.5. *Let $f : U \to \mathbb{R}$ be a convex function defined on an open $U \subseteq \mathbb{R}^m$. Then*

(a) f is continuous on U,

(b) f is Fréchet differentiable almost everywhere on U and Gâteaux differentiable everywhere on U,

(c) f is twice differentiable almost everywhere on U.

Even where f is not differentiable, an important analogue of the first-order characterization of convexity holds.

Definition 9.3.6. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. A vector p is a **subgradient** of f at x if $f(x) + p \cdot (y - c) \leq f(y)$ for all $x, y \in \mathbb{R}^n$. That is, the subgradients at x determine tangent lines everywhere below the function. The set of all subgradients at x is called the **subdifferentiable** and is denoted $\partial f(x)$.

We may similarly define the **supergradients** of a concave function, with the resulting

superdifferential denoted $\partial^* f(x)$, where these determine tangent lines everywhere above the function.

Theorem 9.3.7 (Properties of the subdifferential)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function.*

(a) The subdifferential $\partial f(x)$ is a closed, convex set at all $x \in \mathbb{R}^n$.

(b) If f is continuous, then $\partial f(x) \neq \emptyset$ for all x and ∂f is a lower hemicontinuous correspondence. Moreover, ∂f is monotone, that is for all $s_x \in \partial f(x)$ and $s_y \in \partial f(y)$, $(s_y - s_x) \cdot (y - x) \geq 0$.

(c) Wherever f is differentiable, $\partial f(x) = \{\nabla f(x)\}$ and wherever the subdifferential is single-valued, f is differentiable at that point.

Subdifferentials and superdifferentials will be important in producer and consumer theory. As a preview, think of f as a concave utility function. Then ∂f is the inverse demand correspondence!

This very important theorem tells us why this discussion of convexity belongs in the section on optimization.

Theorem 9.3.8 (First-order subdifferential characterization of minimum)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous convex function. Then $x \in \mathbb{R}^n$ is a global minimzer of f if and only if $0 \in \partial f(x)$.*

Here is another nice consequence of convexity.

Theorem 9.3.9. *Let X be a convex subset of $\mathbb{R}^n$. Then*

(a) If f is concave, then for all $r \in \mathbb{R}$, the upper contour set $\{\mathbf{x} \in X : f(\mathbf{x}) \geq r\}$ is convex.

(b) If f is convex, then for all $r \in \mathbb{R}$, the lower contour set $\{\mathbf{x} \in X : f(\mathbf{x}) \leq r\}$ is convex.

In many economic problems, this last property of concave functions—the convexity of upper contour sets—is exactly what we are after. However, concavity is not strictly necessary for this result. Instead it often suffices to assume the following.

Definition 9.3.10. Suppose $f : X \to \mathbb{R}$, where X is a convex subset of $\mathbb{R}^n$.

(a) The function f is **quasiconcave** if $f(a\mathbf{x} + (1-a)\mathbf{y}) \geq \min\{f(\mathbf{x}), f(\mathbf{y})\}$ for all $\mathbf{x}, \mathbf{y} \in X$ and $a \in [0, 1]$.

(b) The function f is **strictly quasiconcave** if $f(a\mathbf{x} + (1-a)\mathbf{y}) > \min\{f(\mathbf{x}), f(\mathbf{y})\}$ for all $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \neq \mathbf{y}$ and $a \in (0, 1)$.

(c) The function f is **quasiconvex** if $f(a\mathbf{x} + (1-a)\mathbf{y}) \leq \max\{f(\mathbf{x}), f(\mathbf{y})\}$ for all $\mathbf{x}, \mathbf{y} \in X$ and $a \in [0, 1]$.

(d) The function f is **strictly quasiconvex** if $f(a\mathbf{x} + (1-a)\mathbf{y}) < \max\{f(\mathbf{x}), f(\mathbf{y})\}$ for all $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \neq \mathbf{y}$ and $a \in (0,1)$

Theorem 9.3.11. *A function f with convex domain X and range $\mathbb{R}$ is quasiconcave if and only if the sets $\{\mathbf{x} \in X : f(\mathbf{x}) \geq r\}$ are convex for every $r \in \mathbb{R}$. A function f with convex domain X and range $\mathbb{R}$ is quasiconvex if and only if the sets $\{\mathbf{x} \in X : f(\mathbf{x}) \leq r\}$ are convex for every $r \in \mathbb{R}$.*

Proof. Let's prove the first claim. The second claim is proven in a similar way. Suppose f is quasiconcave. Fix $r \in \mathbb{R}$, and suppose $\mathbf{x}'$ and $\mathbf{x}''$ are both elements of $\{\mathbf{x} \in X : f(\mathbf{x}) \geq r\}$. For $a \in [0,1]$, $f(a\mathbf{x}' + (1-a)\mathbf{x}'') \geq \min\{f(\mathbf{x}'), f(\mathbf{x}'')\} \geq r$. Thus, $a\mathbf{x}' + (1-a)\mathbf{x}'' \in \{\mathbf{x} \in X : f(\mathbf{x}) \geq r\}$, and hence $\{\mathbf{x} \in X : f(\mathbf{x}) \geq r\}$ is convex. Conversely, suppose that $\{\mathbf{x} \in X : f(\mathbf{x}) \geq r\}$ is a convex set for every $r \in \mathbb{R}$. Choose any $\mathbf{x}', \mathbf{x}''$ and assume without loss of generality that $f(\mathbf{x}') \geq f(\mathbf{x}'')$. Then $\mathbf{x}' \in \{\mathbf{x} \in X : f(\mathbf{x}) \geq f(\mathbf{x}'')\}$ and $\mathbf{x}'' \in \{\mathbf{x} \in X : f(\mathbf{x}) \geq f(\mathbf{x}'')\}$. Convexity of the set $\{\mathbf{x} \in X : f(\mathbf{x}) \geq f(\mathbf{x}'')\}$ implies that, for all $a \in [0,1]$, $a\mathbf{x}' + (1-a)\mathbf{x}'' \in \{\mathbf{x} \in X : f(\mathbf{x}) \geq f(\mathbf{x}'')\}$, and thus $f(a\mathbf{x}' + (1-a)\mathbf{x}'') \geq f(\mathbf{x}'') = \min\{f(\mathbf{x}'), f(\mathbf{x}'')\}$. Therefore, f is quasiconcave. ■

Theorem 9.3.12. *A (strictly) concave function f is (strictly) quasiconcave. A (strictly) convex function f is (strictly) quasiconvex.*

Proof. The claims for "non-strict versions" follow by combining Theorems 9.3.9 and 9.3.11. The claims for "strict versions" can be shown in a similar way. ■

Finally, we have the following important theorem about quasiconcavity (it is arguably the reason we care about quasiconcavity in economics).

Theorem 9.3.13. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quasiconcave function and $g : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then the composition $g \circ f$ is a quasiconcave function.*

Note that the same need not be true of concave functions!

10

Constrained optimization

Contents

10.1 Optimization problems

Most economic models are based on the solution of optimization problems. These notes outline some of the basic tools needed to tackle these problems.

Definition 10.1.1. We consider the **parametric constrained optimization problem** of the following form:

$$\max_{x \in D(\theta)} f(x, \theta)$$

where $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ is called the **objective function**, x is a **choice variable**, $D(\theta)$ is the available **choice set**, and θ is an exogenous parameter that may affect both the objective function and the choice set. Let Θ denote the set of all possible parameter values.

Definition 10.1.2. The **solution set** is defined as

$$x^*(\theta) = \arg\max_{x \in D(\theta)} f(x, \theta)$$

The **value function** is defined as

$$V(\theta) \equiv \max_{x \in D(\theta)} f(x, \theta)$$

Note that any results derived for a maximization problem can be used in a minimization problem, which follows from the fact that

$$\begin{aligned} x^*(\theta) = \arg\min_{x \in D(\theta)} f(x, \theta) \quad &\Leftrightarrow \quad x^*(\theta) = \arg\max_{x \in D(\theta)} -f(x, \theta) \\ \text{and } V(\theta) = \min_{x \in D(\theta)} f(x, \theta) \quad &\Leftrightarrow \quad V(\theta) = -\max_{x \in D(\theta)} -f(x, \theta) \end{aligned}$$

There are several questions of interest:

1. When does a solution to the maximization problem exist for each θ?
2. What are the main properties of the solution set and the value function?
3. How can we compute the solution to the problem?
4. How do the solution set and the value function change with the parameters?

10.2 Properties of solutions

This section addressed the first two questions posed in the previous section. To answer these questions, we will use the notions of continuity and convexity of functions and continuity of correspondences, discussed in previous lectures.

The following theorem, called the **Theorem of the Maximum**, or **Berge's Theorem** provides answers on the existence of a solution and continuity properties of the solution set and the value function.

Theorem 10.2.1 (Berge's Theorem of the Maximum)**.** *Consider the class of parametric constrained optimization problems*

$$\max_{x \in D(\theta)} f(x, \theta)$$

defined over the set of parameters Θ. *Suppose that*

(i) $D : \Theta \rightrightarrows X$ *is continuous (i.e. upper and lower hemicontinuous) and compact valued;*

(ii) $f : X \times \Theta \to \mathbb{R}$ *is a continuous function.*

Then

1. $x^*(\theta)$ *is non-empty for every* $\theta \in \Theta$*;*
2. $x^*(\theta)$ *is upper hemicontinuous;*
3. V *is continuous.*

The following examples illustrate the role of some of the assumptions in the theorem.

Example. What can happen when D is not compact? Consider $\Theta = [0, 10]$, $D(\theta) = (0, 1)$, and $f(x, \theta) = x$. Then $x^*(\theta) = \emptyset$ for all $\theta \in \Theta$. ♣

Example. What can happen if D is lower hemicontinuous, but not upper hemicontinuous? Suppose that $\Theta = [0, 10]$, $f(x, \theta) = x$, and

$$D(\theta) = \begin{cases} \{0\} & \text{if } \theta \leq 5 \\ [-1, 1] & \text{otherwise} \end{cases}$$

The solution is given by

$$x^*(\theta) = \begin{cases} \{0\} & \text{if } \theta \leq 5 \\ \{1\} & \text{otherwise} \end{cases}$$

which is not upper hemicontinuous. The value function is also discontinuous. ♣

Example. What can happen if D is upper hemicontinuous, but not lower hemicontinuous? Suppose $\Theta = [0, 10]$, $f(x, \theta) = x$, and

$$D(\theta) = \begin{cases} \{0\} & \text{if } \theta < 5 \\ [-1, 1] & \text{otherwise} \end{cases}$$

Then the solution set is given by

$$x^*(\theta) = \begin{cases} \{0\} & \text{if } \theta < 5 \\ \{1\} & \text{otherwise} \end{cases}$$

which is again not upper hemicontinuous. ♣

Example. What can happen if f is not continuous? Suppose that $\Theta = [0, 10]$, $D(\theta) = [\theta, \theta + 1]$, and

$$f(x, \theta) = \begin{cases} 0 & \text{if } x < 5 \\ 1 & \text{otherwise} \end{cases}$$

Then the solution set is given by

$$x^*(\theta) = \begin{cases} [\theta, \theta + 1] & \text{if } \theta < 4 \\ [5, \theta + 1] & \text{if } 4 \leq \theta < 5 \\ [\theta, \theta + 1] & \text{otherwise} \end{cases}$$

and the value function is given by

$$V(\theta) = \begin{cases} 0 & \text{if } \theta < 4 \\ 1 & \text{otherwise} \end{cases}$$

Therefore, x^* is not upper hemicontinuous and V is not continuous. ♣

The following theorem identifies conditions for convexity of the solution set, uniqueness of the solution, and concavity of the value function.

Theorem 10.2.2. *Consider the class of parametric constrained optimization problems*

$$\max_{x \in D(\theta)} f(x, \theta)$$

defined over the convex set of parameters Θ. Supose that

(i) $D : \Theta \rightrightarrows X$ is continuous and compact valued;

(ii) $f : X \times \Theta \to \mathbb{R}$ is a continuous function.

Then

1. *If $f(\cdot, \theta)$ is a quasiconcave function in x for each θ, and D is convex valued, then x^* is convex-valued.*
2. *If $f(\cdot, \theta)$ is a strictly quasiconcave function in x for each θ, and D is convex valued, then x^* is single-valued.*
3. *If f is a concave function in (x, θ) and D is convex-valued, then V is a concave function and x^* is convex-valued.*
4. *If f is a strictly concave function in (x, θ) and D is convex-valued, then V is a strictly concave function and x^* is single-valued.*

Proof. 1. Suppose that $f(\cdot, \theta)$ is a quasiconcave function in x for each θ, and D is convex valued. Pick any $x, x' \in x^*(\theta)$. Since D is convex-valued, $x_a = ax + (1 - a)x' \in D(\theta)$ for all $a \in [0, 1]$. Also, by the quasi-concavity of f we have that $f(x_a, \theta) \geq f(x, \theta) = f(x', \theta)$. But since $f(x, \theta) = f(x', \theta) \geq f(y, \theta)$ for all $y \in D(\theta)$, we get that $f(x_a, \theta) \geq f(y, \theta)$ for all $y \in D(\theta)$. Therefore, $x_a \in x^*(\theta)$.

2. The proof is by contradiction. Suppose, aiming for a contradiction, that $x^*(\theta)$ is not single-valued at θ, which is to say that $x^*(\theta)$ contains two distinct points x and x'. As before, since D is convex-valued, $x_a = ax + (1 - a)x' \in D(\theta)$ for all $a \in (0, 1)$. By the strict quasi-concavity of $f(\cdot, \theta)$ in x, $f(x_a, \theta) > f(x, \theta) = f(x', \theta)$, which contradicts the fact that x and x' are maximizers in $D(\theta)$.

3. Suppose that f is a concave function in (x, θ) and D is convex-valued. Since concavity of f

in (x, θ) implies quasi-concavity of $f(\cdot, \theta)$ in x, it follows that x^* is convex-valued. For the concavity of V, pick any $\theta, \theta' \in \Theta$ and let $\theta_a = a\theta + (1-a)\theta'$ for some $t \in [0, 1]$. Let $x \in x^*(\theta)$ and $x' \in x^*(\theta')$, and let $x_a = ax + (1-a)x'$. Then

$$\begin{aligned} V(\theta_a) &\geq f(x_a, \theta_a) \\ &= f(ax + (1-a)x', a\theta + (1-a)\theta') \\ &\geq af(x, \theta) + (1-a)f(x', \theta') \\ &= aV(\theta) + (1-a)V(\theta') \end{aligned}$$

4. Very similar to 3.

■

10.3 Characterization of Solution

The next step is to learn how to solve optimization problems and characterize the solution to a particular problem. To do that, we focus on more restricted classes of problems than in the previous section. To begin with, we will discuss the relatively simple case of equality constraints. We will then discuss some of the techniques for optimization with inequality constraints.

We will begin by fixing the parameter θ and characterizing the solution for a given θ. We will therefore suppress the θ in the notation. In the next section, we will return the θ to the problem, and consider how θ influences the solution to the problem, which is called *comparative statics.*

Equality constraints - Lagrange multipliers

Consider the following problem

$$\max_{x \in \mathbb{R}^n} f(x)$$
$$\text{subject to } h(x) = 0.$$

Here f is a real-valued function defined on $\mathbb{R}^n$ or an open subset of $\mathbb{R}^n$, while $h = (h_1, ..., h_m)'$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Both are assumed to be in class C^1.

We will now give a rough justification for the method of solution of such problems.

Each function $h_i = 0$ defines a surface in $\mathbb{R}^n$ which is usually $n - 1$ dimensional, so that the intersection $h = 0$ is typically an $(n - m)-$dimensional subspace of $\mathbb{R}^n$, called a manifold. Denote this manifold by M.

Consider all the curves $x : \mathbb{R} \to M$ passing through a point $x \in M$. The **tangent space** $\mathcal{T}_x$ at x is the set of all derivatives of these curves at point x. We would like to express this tangent space in terms of ∇h, which is possible under the following condition.

Definition 10.3.1. The point x is called a **regular point** if the set $(\nabla h_i(x))_{i=1}^m$ is linearly independent.

We have the following fact about regular points.

Theorem 10.3.2. *Let x be a regular point on M defined by $h(x) = 0$. The tangent space of M at x is the same as $\{y : \nabla h(x) \cdot y = 0\}$.*

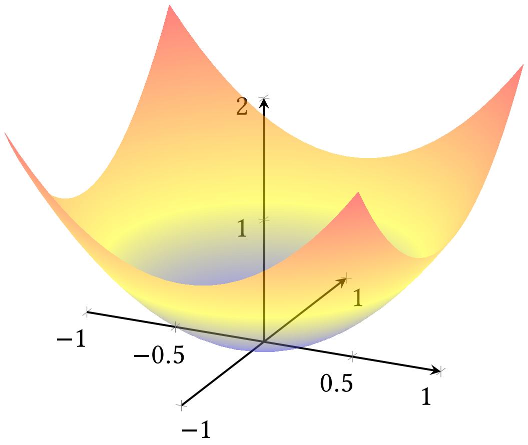
Without the assumption that x is a regular point, we could only obtain that $\{y : \nabla h(x)\dot{y} = 0\}$ is a *subspace* of the tangent space. That is, there would be tangent vectors not satisfying the condition.

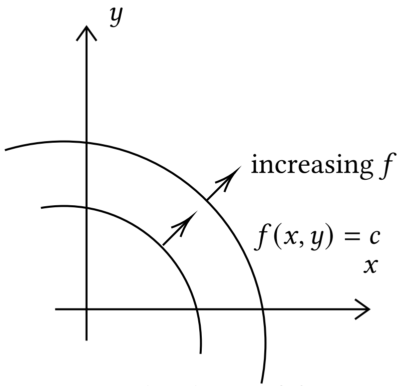
We now show that the gradient of f at any constrained optimum must be orthogonal to the tangent space, which is called the *tangency condition*. Suppose that f obtains a local extremum at x^*, subject to the constraints that $h(x) = 0$. Let y be any vector in the tangent space at x^*, that is there exists a path $x(t)$ through x^* with derivative y at 0, so $x(0) = x^*$, $x'(0) = y$. Since x^* is regular, we have that $\nabla h(x^*) \cdot y = 0$. Since x^* is a constrained optimum of f, we have that $\frac{d}{dt} f(x(t))|_{t=0} = 0$. But by the chain rule, this implies that $\nabla f(x^*) \cdot y = 0$. Thus $\nabla f(x^*)$ is orthogonal to any $y \in \mathcal{T}$.

But since x^* is regular, the orthogonal complement of the tangent space is spanned by vectors of the form ∇h_i, which implies that ∇f must be a linear combination of such vectors. That gives us the following important theorem.

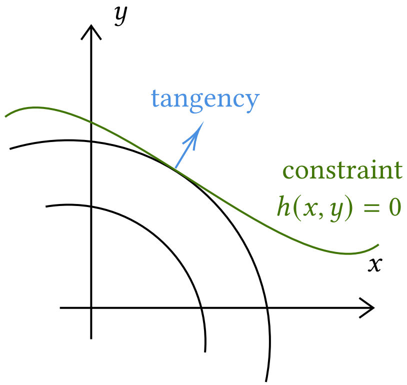
Theorem 10.3.3. *Let x^* be a constrained extremum of f subject to $h(x) = 0$, and assume that x^* is a regular point of the constraint set. Then there exist **Lagrange multipliers** $\lambda \in \mathbb{R}^m$ such that $\nabla f(x^*) + \lambda \cdot \nabla h(x^*) = 0$.*

An illustration of this method is below.





level sets of f



tangency condition

Note that the Lagrange condition is a necessary but not sufficient condition for an extremum. It also does not tell us whether the extremum of interest is a maximum or a minimum. Note also that the equation $\nabla f(x^*) + \lambda \cdot \nabla h(x^*) = 0$ together with the constraints $h(x^*) = 0$ give a total of $n+m$ equations in $n+m$ unknowns x^*, λ, and thus they comprise sufficient information to identify candidate solutions. It is convenient to introduce a **Lagrangian**

$$\mathcal{L}(x, \lambda) = f(x) + \lambda \cdot h(x),$$

with which these conditions may be written $D_x\mathcal{L} = 0$ and $D_\lambda\mathcal{L} = 0$.

Generally, checking whether a candidate optimum is a maximum or a minimum can be quite annoying. There are necessary and sufficient conditions that are somewhat difficult to check, as below.

Theorem 10.3.4. *Suppose that x^* satisfies the necessary conditions for an extremum (with x^* a regular point). Then if x^* is a local maximum, for all $y \in \mathcal{T}_{x^*}$, we have $y' D^2\mathcal{L}(x^*, \lambda^*)y \leq 0$.*

If x^* is an extremum satisfying the strict inequality $y'D^2\mathcal{L}(x^*,\lambda^*)y < 0$ for all $y \in \mathcal{T}_{x^*}$, then x^* is a local maximum.

Note that there may be local maxima that for which $y'D^2\mathcal{L}(x^*,\lambda^*)y = 0$ for some $y \in \mathcal{T}_{x^*}$. For local minima, the same conditions may be used with the reversed inequality signs.

One (slightly) simpler approach is to analyze the **bordered Hessian**.

$$\mathbf{H}(\mathcal{L}) = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \lambda^2} & \frac{\partial^2 \mathcal{L}}{\partial \lambda \partial \mathbf{x}} \\ \left(\frac{\partial^2 \mathcal{L}}{\partial \lambda \partial \mathbf{x}}\right)^\top & \frac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} \end{bmatrix} = \begin{bmatrix} 0 & \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} & \cdots & \frac{\partial g}{\partial x_n} \\ \frac{\partial g}{\partial x_1} & \frac{\partial^2 \mathcal{L}}{\partial x_1^2} & \frac{\partial^2 \mathcal{L}}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial x_1 \partial x_n} \\ \frac{\partial g}{\partial x_2} & \frac{\partial^2 \mathcal{L}}{\partial x_2 \partial x_1} & \frac{\partial^2 \mathcal{L}}{\partial x_2^2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g}{\partial x_n} & \frac{\partial^2 \mathcal{L}}{\partial x_n \partial x_1} & \frac{\partial^2 \mathcal{L}}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial x_n^2} \end{bmatrix} = \begin{bmatrix} 0 & \frac{\partial g}{\partial \mathbf{x}} \\ \left(\frac{\partial g}{\partial \mathbf{x}}\right)^\top & \frac{\partial^2 \mathcal{L}}{\partial \mathbf{x}^2} \end{bmatrix}$$

The k-th **principal minor** is the determinant of the top-left $k \times k$ submatrix in $H(\mathcal{L})$. If the **last** $n - m$ principal minors of the bordered Hessian is such that the m^{th} minor has sign $(-1)^{m+1}$ and then the subsequent minors alternate in sign, then the candidate solution is a local maximum. If the last $n - m$ principal minors are all negative, the candidate solution is a local minimum.

In general, you will not need to check these sufficient conditions in problem sets, but they are worth keeping in mind as you apply the Lagrangian approach "irl".

Exercise 10.1. *The output of a firm is given by the Cobb-Douglas production function $f(L,K) = 2.5L^{0.45}K^{0.55}$ where L is labor and K is capital. Each unit of labor costs 40, and each unit of capital costs 50. The company faces a budget constraint of 500,000 per year. What is the optimal level of consumption for the firm?*

Exercise 10.2. *Solve*

$$\min_{x,y} x \text{ subject to } y^2 + x^4 - x^3 = 0.$$

Inequality constraints

Let us now consider the following problem:

$$\max_{x \in \mathbb{R}^n} f(x,)$$

$$\text{subject to } g(x) \leq 0 \qquad\qquad \text{for } k = 1, ..., K$$

where $f(\cdot)$ is a real-valued function defined on $\mathbb{R}^n$ while $g = (g_1(\cdot), ..., g_k(\cdot))'$ maps $\mathbb{R}^n$ to $\mathbb{R}^k$. Note that this strictly generalizes the case of equality constraints discussed above, since $g_k(x) = 0$ is equivalent to $g_k(x) \leq 0$ and $-g_k(x) \leq 0$.

Definition 10.3.5. We call the optimization problem

$$\max_{x \in X} f(x) \text{ subject to } g(x) \leq 0$$

the **primal problem** and any $x \in X$ such that $g(x) \leq 0$ a **feasible solution**. Denote a solution to the primal problem, a **maximizer**, by x^* and let $f^* = f(x^*)$ be the **value** of the primal problem. If $g_k(x^*) = 0$, then we say that the g_k constraint **binds** at x^*, otherwise we say it is **slack**.

In principle, we could approach the above problem by first identifying any unconstrained optima in the feasible set, and then by identifying the extrema given *every possible subset* of the constraint set binding (using the optimization methods for equality constraints discussed before), and then by comparing all the extrema we obtained to identify the best one. Fortunately, for many problems, there are more systematic approaches, and we discuss these now.

As previously, we may define the Lagrangian.

Definition 10.3.6. Let $f, g_1, ..., g_k : X \to \mathbb{R}$. The **Lagrangian** $\mathcal{L} : X \times \mathbb{R}^k_+ \to \mathbb{R}$ is defined by

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{j=1}^{m} \lambda_j g_j(x) = f(x) - \lambda \cdot g(x),$$

where $\lambda = (\lambda_1, ..., \lambda_m)$ are the **Lagrange multipliers**.

The sign of the multiplier subtracted in the Lagrangian is now important. For a maximization problem, we have *nonnegative* multipliers and *subtract* the term $\lambda \cdot g$. Think of the term $\lambda \cdot g$ as a *penalty* in the Lagrangian. In particular, the objective is to maximize the Lagrangian, and if we violated the constraint given $\lambda \geq 0$ the term $-\lambda_k g_k(x)$ decreases the value of our objective. This is bad since we are solving a maximization problem. If the value of the multiplier is high enough, the penalty for violating the constraint will be so large that the constraint will be obeyed.

We formalize some of this intuition. First observe that for *any* feasible $x \in X$ and every $\lambda \geq 0$, we have that

$$\mathcal{L}(x, \lambda) \geq f(x).$$

Noting that the minimum of $-\lambda \cdot g(x)$ is either 0 or $-\infty$ (depending on whether x is feasible or not), we thus have that

$$f^* = \max_{x \in X} \min_{\lambda \geq 0} \mathcal{L}(x, \lambda).$$

This formulation is helpful because we have transformed a constrained maximization problem (the primal problem) to an unconstrained optimization problem.

Let's now think about what happens if we change the order of operations in this optimization problem: that is, we minimize before maximizing.

Definition 10.3.7. For a given primal problem (f, g), define the **dual function** by

$$\phi(\lambda) = \max_{x \in X} \mathcal{L}(x, \lambda) = \max_{x \in X} \{f(x) - \lambda \cdot g(x)\}.$$

The **dual problem** is

$$\min_{\lambda \geq 0} \phi(\lambda),$$

with associated **dual minimizers** λ^* and **dual value** $\phi^* = \min_{\lambda \geq 0} \phi(\lambda)$. If $\lambda \geq 0$ and $\phi(\lambda) > -\infty$, we call such λ **dual feasible**.

Since ϕ is the pointwise maximum of a family of affine functions, it is convex. If we take the bound $\mathcal{L}(x, \lambda) \geq f(x)$, fix λ and maximize over the x on the left-hand side, we obtain

$$f(x) \leq \max_{x' \in X} \mathcal{L}(x', \lambda) = \phi(\lambda),$$

which after maximizing over the x on the left-hand-side obtains

$$f^* \leq \phi(\lambda).$$

Finally, minimizing over the λ on the right-hand-side, we obtain

$$f^* \leq \phi^*.$$

This is an important theorem.

Theorem 10.3.8 (Weak Lagrangian duality)**.** *Given any primal problem (f, g), the primal and dual values satisfy*

$$f^* \leq \phi^*.$$

When $f^* = \phi^*$, then there must exist (x^*, λ^*) such that $\mathcal{L}(x^*, \lambda^*) = f^* = \phi^*$. This is convenient, due to the following theorem.

Theorem 10.3.9 (Lagrangian saddlepoints are constrained maxima)**.** *Suppose that (x^*, λ^*) is* *a saddlepoint of the Lagrangian $\mathcal{L}(x,\lambda) = f + \lambda \cdot g$, that is,*

$$\mathcal{L}(x,\lambda^*) \underset{(a)}{\leq} \mathcal{L}(x^*,\lambda^*) \underset{(b)}{\leq} \mathcal{L}(x^*,\lambda) \quad x \in X, \lambda \geqq 0$$

Then x^* maximizes f over X subject to the constraints $g_j(x) \geqslant 0$, $j = 1,\ldots,k$, and furthermore

$$\lambda_j^* g_j(x^*) = 0 \quad j = 1,\ldots,k.$$

These last equalities are called the ***complementary slackness conditions****.*

Proof. Inequality (b) implies $\lambda^* \cdot g(x^*) \leqslant \lambda \cdot g(x^*)$ for all $\lambda \geq 0$. Therefore $g(x^*) \geq 0$, so x^* satisfies the constraints. Setting $\lambda = 0$, we see that $\lambda^* \cdot g(x^*) \leq 0$. This combined with $\lambda \geqq 0$ and $g(x^*) \geq 0$ implies $\lambda^* \cdot g(x^*) = 0$ and moreover that $\lambda_j^* g_j(x^*) = 0$ for $j = 1,\ldots,m$.

Inequality (a) implies $f(x)+\lambda^*\cdot g(x) \leqslant f(x^*)$ for all x. Therefore, if x satisfies the constraints, $g(x) \geqq 0$, we have $f(x) \leqslant f(x^*)$, so x^* is a constrained maximizer. ■

Much of convex optimization theory thus looks for conditions under which we may be assured that $f^* = \phi^*$, which is called strong duality.

Definition 10.3.10. Let f^* be the primal value and ϕ^* the dual value. Then $\phi^* - f^*$ is called the **duality gap**. If the duality gap is zero, we say that the primal-dual pair satisfies **strong duality**.

The simplest strong duality condition applies when the optimization problem is convex, as in the following definition.

Definition 10.3.11 (Convex program)**.** The primal problem

$$\max_{x\in X} f(x) \text{ subject to } g(x) \leq 0$$

is **convex** if X is a convex set, f is a concave function, and each g_1, ..., g_k is a convex function of x.

It might seem strange to require f to be *concave* in the definition of a *convex* program, but this is because of a tradition in optimization theory in which minimization problems are studied rather than maximization problems (this is mostly opposite to the convention in economics).

When the optimization problem is convex, strong duality almost always applies. There are many conditions under which strong duality holds for convex problems, which are called *constraint qualifications*. We list two important ones below.

Definition 10.3.12. (a) **Linear independence constraint qualification (LICQ)**: Let x be feasible in the primal problem and let $B(x)$ denote the set of binding constraints at x. Suppose that each g_k for $k \in B(x)$ is differentiable at x. Then the LICQ holds at x if the vectors in the set $\{\nabla g_k(x) : k \in B(x)\}$ are linearly independent.

(b) **Slater's condition**: If there exists some feasible $x' \in X$ such that all inequality constraints are *strictly* satisfied (and equality constraints are also exactly satisfied), then Slater's condition holds. That is x' is in the (relative) interior of the set of feasible points.

We have the following theorem for convex programs.

Theorem 10.3.13 (Convex strong duality)**.** *Strong duality holds for a convex program if either Slater's condition holds or if the LICQ holds at the primal maximizer x^*.*

This implies that for x^* to be a maximizer of the convex program, it is *necessary and sufficient* that

$$0 \in \partial f(x^*) - \sum_{k=1}^{K} \lambda_k^* \partial g_k(x^*).$$

When the primal problem is not convex, slightly more assumptions are required to obtain first-order conditions for optimality.

Theorem 10.3.14 (Karush-Kuhn-Tucker (KKT))**.** *Suppose that the following conditions holds:*

(a) $f(\cdot), g_1(\cdot), ..., g_K(\cdot)$ are continuously differentiable in $\mathbf{x}$;

(b) $D(\theta)$ is nonempty;

(c) x^* is a solution to the optimization problem;

(d) constraint qualification holds at x^*.

Then

1. *The first-order condition: There exist non-negative numbers $\lambda_1, ..., \lambda_K$ such that*

$$\nabla f(x^*) = \sum_{k=1}^{K} \lambda_k \nabla g_k(x^*)$$

2. *Complementary slackness condition: For $k = 1, ..., K$,*

$$\lambda_k g_k(x^*) = 0$$

Example. Consider the following problem:

$$\max_{0 \leq x \leq 5} x^3 - 5x^2 + x$$

Observe that we have two constraints, $x \leq 5$ and $x \geq 0$. Let λ be the multiplier for the $x \leq 5$ constraint and μ be the multiplier for the $x \geq 0$ constraint. The Lagrangian is

$$\mathcal{L} = x^3 - 5x^2 + x + \lambda(5 - x) + \mu x$$

The first order condition gives

$$3x^2 - 10x + 1 - \lambda + \mu = 0$$

and the complementary slackness conditions are

$$\lambda(5 - x) = 0, \quad \mu x = 0$$

One way to solve this problem is by brute force. There are three possible cases to consider:

1. $x \leq 5$ binds, and hence $x \geq 0$ cannot bind. In that case $x = 5$, $\mu = 0$, and $\lambda = 26$.
2. $x \geq 0$ binds, and hence $x \leq 5$ does not bind. In that case $x = 0$, $\mu = -1$, $\lambda = 0$.
3. neither $x \leq 5$ nor $x \geq 0$ bind. In that case $\mu = \lambda = 0$, $x = 0.103$ or $x = 3.23$.

We can immediately rule case (2) since $\mu = -1$, and Kuhn-Tucker requires that $\mu \geq 0$. Now we proceed by checking whether case (1) or case (3) (or both) maximize the utility. Doing so, we find that $x^* = 5$. (Draw the function.) So the Kuhn-Tucker algorithm found local maxima and minima, and also forced us to check the endpoints. ♣

Observe that in the steps outlined above in the Kuhn-Tucker algorithm for finding candidates to the solution of the problem, some candidates might fail to be solutions to the problem. This is clear from the example above (where the candidates were $x = 0.103,\ 3.23,\ 5$, but $x^* = 5$ only). In particular, there might be solutions $(\mathbf{x}, \lambda)$ to the Kuhn-Tucker conditions that are not solutions to the optimization problem. The Kuhn-Tucker theorem establishes conditions under which the Kuhn-Tucker conditions are *necessary* conditions for an optimum. But they might not be *sufficient.* To make sure that a particular candidate found by applying the Kuhn-Tucker theorem is a maximizer of the problem, we need to check that the second-order conditions hold at that point. In particular, one needs to calculate the Hessian matrix of second derivatives and test for negative semi-definiteness. This is usually a pain to do (unless $n = 1$), so we will avoid going into details (see Mas-Colell, Whinston and Green (1995) if this excites you). The following theorem provides conditions under which the Kuhn-Tucker conditions are both necessary and sufficient (and these conditions will be satisfied in most problems in which you are expected to apply the Kuhn-Tucker theorem).

Theorem 10.3.15. *Suppose that the conditions of the Kuhn-Tucker Theorem are satisfied and that*

(a) $f(\cdot)$ *is quasiconcave in x; and*

(b) $g_1(\cdot), ..., g_K(\cdot)$ *are quasiconvex.*

Then any point x^* *that satifies the Kuhn-Tucker conditions is a solution to the constraint optimization problem.*

Observe that $g_k(\cdot)$ is quasiconvex if and only if the set $\{\mathbf{x} : g_k(\mathbf{x}) \le 0\}$ is convex. Since a finite intersection of convex sets is convex, the assumption in (b) ensures that the constraint set is convex.

Proof. The proof considers only the case for which f is concave.

A continuously differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ is quasiconvex if and only if

$$Dg(\mathbf{x}) \cdot (\mathbf{x}' - \mathbf{x}) \le 0 \quad \text{whenever} \quad g(\mathbf{x}') \le g(\mathbf{x})$$

Suppose that x^* satisfies the Kuhn-Tucker conditions. Then there are multipliers $\lambda_1, ..., \lambda_K$ such that

$$\nabla f(x^*) = \sum_{k=1}^{K} \lambda_k \nabla g_k(x^*)$$

But then, since for any feasible $\mathbf{x}$ we have that $g_k(\mathbf{x}) \leq g_k(x^*)$ whenever $\lambda_k > 0$, we have that

$$\nabla f(x^*) \cdot (\mathbf{x} - x^*) = \sum_{k=1}^{K} \lambda_k \nabla g_k(x^*) \cdot (\mathbf{x} - x^*) \leq 0$$

Note that for concave f we know that

$$f(\mathbf{x}) \leq f(x^*) + \nabla f(x^*) \cdot (x - x^*)$$

Since $\nabla f(x^*) \cdot (x - x^*) \leq 0$, we conclude that $f(\mathbf{x}) \leq f(x^*)$, and thus x^* is a solution to the problem. ■

Example. (Non-negativity constraints) Consider a consumer who is maximizing her utility $u(x)$ by choosing a bundle $x \in \mathbb{R}^n$ such that her total expenditure does not exceed her wealth and she cannot consume negative amounts of any good. Her problem is

$$\begin{aligned} &\max_{x \in \mathbb{R}^n} u(x) && \\ \text{subject to } & p_1 x_1 + \ldots + p_n x_n \leq w && \text{(budget constraint)} \\ & x_i \geq 0, \quad i = 1, \ldots, N && \text{(non-negativity constraints)} \end{aligned}$$

where p_i is the price of good i, and w is consumer's wealth. Let λ be the Lagrange multiplier on the budget constraint (which we will assume it binding, which will usually be the case given some natural assumptions on the properties of the utility function) and μ_i be the multiplier on the non-negativity constraint for good i (which may or may not bind). The Lagrangian for the problem is

$$\mathcal{L}(x, \lambda, \mu) = u(x) + \lambda \left(w - p_1 x_1 - \ldots - p_n x_n\right) + \sum_{i=1}^{n} \mu_i x_i$$

The first-order conditions become

$$\frac{\partial u}{\partial x_i} = \lambda p_i - \mu_i \qquad\qquad i = 1, \ldots, n$$

which is equivalent to

$$\frac{\partial u}{\partial x_i} \leq \lambda p_i \qquad\qquad \text{with equality if } x_i > 0$$

for $i = 1 \ldots, n$. The complementary slackness conditions are $\mu_i x_i = 0$, $i = 1, \ldots, n$.

When we have an interior solution (i.e. $x^* >> 0$), then $\nabla u(x^*) = \lambda p$ (in the vector notation), meaning that the price vector and the vector of marginal utilities are parallel. Since the price vector is perpendicular to the budget line, and the vector of marginal utilities is perpendicular to the utility level curves, this gives us the indifference curve equals the slope of the budget line. ♣

Mixture of equality and inequality constraints

When the primal program includes equality constraints, the Lagrangian duality approach discussed above continues to work, but additional care needs to be taken. In particular, since there are nonnegative multipliers associated with both $g_i(x) \leq 0$ and $-g_i(x) \leq 0$ for some g_i, the dual variables associated with these constraints, say λ_i and λ_i' can be combined into a single dual variable $\nu_i = \lambda_i - \lambda_i'$. That is why you will often see separate theorems stated for optimization programs with equality constraints. These separate theorems are strictly unnecessary, but we will note the general principle for them below.

Theorem 10.3.16. *Consider the primal problem*

$$\max_{x \in X} f(x) \text{ subject to } g(x) \leq 0 \text{ and } h(x) = 0.$$

The Lagrangian for this problem may now be written

$$\mathcal{L}(x, \lambda, \nu) = f(x) - \lambda \cdot g(x) - \nu \cdot h(x),$$

where each $\lambda_i \geq 0$ and $\nu_i \in \mathbb{R}$.

The convex strong duality and KKT theorems also hold with this modified definition of the Lagrangian.

One note of caution: for a convex program, any equality constraints must be affine functions, since this requires $h(x)$ and $-h(x)$ to be concave, which holds exactly when h is affine.

10.3.1 Deleting constraints

Sometimes we know which contraints in an optimization problem bind, and which do not. Or sometimes due to a lot of constraints, we might have a good guess. One might try to conclude that the constraints that do not bind can be ignored without changing the optimal solution. If this were true, it would simplify the problem in many occasions, as we could ignore some constraints that we believe do not bind, and solve the simplified problem and confirm the conjecture that the ignored constraints do not bind. Unfortunately, this cannot always be done. Ignoring constraints that do not bind does change the solution sometimes. For instance, if in Example 25 we change the constraint from $0 \leq x \leq 5$ to $0 \leq x \leq 3$, the new maximum will be attained at $x = 0.103$. Even though the constraint $x \leq 3$ does not bind, ignoring it implies that the maximum will change. The following theorem identifies conditions under which ignoring the constrains that do not bind is valid:

Theorem 10.3.17. *Consider the maximization problem*

$$\max_{x \in \mathbb{R}^n} f(x) \quad \textit{subject to} \quad g_k(x) \leq 0 \quad \textit{for all} \;\; k = 1, \ldots, K$$

Suppose that the conditions of the Kuhn-Tucker Theorem hold and that

(a) $f(\cdot, \theta)$ *is strictly quasiconcave;*

(b) $g_1(\cdot), \ldots, g_K(\cdot)$ *are quasiconvex;*

(c) $g_1(\cdot), \ldots, g_B(\cdot)$ *are binding constraints at the solution;*

(d) $g_{B+1}(\cdot), \ldots, g_K(\cdot)$ *are slack constraints at the solution.*

Then x^* *is a solution if and only if it is a solution to the modified problem*

$$\max_{x \in \mathbb{R}^n} f(x) \quad \textit{subject to} \quad g_k(x) = 0 \quad \textit{for all} \;\; k = 1, \ldots, B$$

Proof. (More like a sketch of the proof.) Conditions (a) and (b) insure the uniqueness of the solution to the main maximization problem. Call it x^*. Suppose, aiming for a contradiction, that there is a point $\hat{x}$ that satisfies the constraints of the second problem, for which $f(\hat{x}) > f(x^*)$. Then because $f(\cdot)$ is strictly quasiconcave, $f(x') > f(x^*)$ for all $x' = a\hat{x} + (1 - a)x^*$, $a \in (0, 1)$. Furthermore, by strict quasiconvexity of the constraints and continuity of g_is, x' satisfies all of the constraints of the first problem for a close enough to 0. But then x^* cannot be a solution to the main problem, a contradiction. ■

10.4 Comparative statics

We now return the parameter θ as a subject of study. In economics, we are often interested in questions of how changes in exogenous variables affect endogenous outcomes. *Comparative statics* is a term for these kinds of questions:

1. How does the optimal choices $x^*(\theta)$ depend on θ?

2. How does the maximized value of the objective, $V(\theta) = f(x^*(\theta), \theta)$ depend on θ?

These questions might seem straightforward, as an obvious thing to do would be, given a particular problem, to find $x^*(\theta)$ and $V(\theta)$ and then simply check how each of the two functions varies with θ (by for instance, taking derivatives with respect to θ, if the two functions are differentiable in θ)? However, it is often difficult, or not even possible to obtain explicit solution for $x^*(\cdot)$, and/or $V(\cdot)$. We will discuss two tools for the first question: the *Implicit Function Theorem* and the *Topkis' Theorem.* We will use the *Envelope Theorem* for the second question.

10.4.1 Implicit function theorem

Recall that the *Implicit Function Theorem* (Theorem 7.3.1) gives us sufficient conditions under which an implicitly defined equation as a local solution and a formula for calculating local derivatives. This is used in comparative statics by applying the implicit function theorem to the first-order conditions of the optimization problem.

The idea is as follows. For simplicity, consider an unconstrained optimization problem $\max_{x\in\mathbb{R}} f(x,\theta)$, and suppose that we know that the first order conditions are necessary and sufficient (e.g., because $f(x,\theta)$ is bounded and concave in x for any fixed θ). The first order conditions are

$$f_x(x,\theta) = 0.$$

The implicit function theorem gives us the expression

$$\frac{\partial x^*(\theta)}{\partial\theta} = -\frac{f_{x\theta}(x^*(\theta),\theta)}{f_{xx}(x^*(\theta),\theta)},$$

under the assumption that $f_{xx}(x^*(\theta),\theta) \neq 0$.

Alternatively, we can arrive at this expression directly by differentiate both sides of the first-order conditions with respect to θ, to obtain

$$f_{xx}(x^*(\theta),\theta)\frac{\partial x^*(\theta)}{\partial\theta}+f_{x\theta}(x^*(\theta),\theta)=0,$$

which can be re-organized to give the above expression.

We will now consider a few example applications of this approach.

Example. Consider $h(x(\theta),\theta) = \ln(x(\theta)) + \theta^2$. Applying the theorem, we have that $x'(\theta) = -2x(\theta)\theta$. In this case it turns out we could've firstly solved for $x(\theta)$ explicitly and then taken the derivative. Obviously, we would've ended up with the same thing. ♣

Now consider the multi-dimensional case: $\mathbf{x} \in \mathbb{R}^n$, $\theta \in \mathbb{R}^m$:

$$D_x f(\mathbf{x}^*(\theta),\theta) = 0$$

The implicit function theorem then implies that

$$D_\theta \mathbf{x}^*(\theta) = -\left[D_{xx}f(\mathbf{x}^*(\theta),\theta)\right]^{-1} D_{x\theta}h(\mathbf{x}^*(\theta),\theta),$$

under the assumption that the solution $x(\theta)$ is locally unique, all the relevant derivatives exist at the optimum, and $D_x h(\mathbf{x}^*(\theta),\theta)$ has full rank.

Example. Consider the following problem:

$$\max_{0\le x\le 2}(x-\theta)^3 - x^2$$

Answer:

The Lagrangian is

$$\mathcal{L} = (x-\theta)^3 - x^2 + \lambda(2-x) + \mu x$$

The first order conditions gives

$$3(x-\theta)^2 - 2x - \lambda + \mu = 0$$

If one of the constraints binds, then the solution does not change locally. So, we only consider solutions in the interior, where $\lambda = \mu = 0$. Thus, we have that

$$h(x(\theta), \theta) = 3(x(\theta)-\theta)^2 - 2x(\theta) = 0$$

Applying the theorem, we have that

$$\frac{\partial x(\theta)}{\partial \theta} = \frac{3x(\theta)-3\theta}{3x(\theta)-3\theta-1}$$

So, what can we say about $\frac{\partial x(\theta)}{\partial \theta}$? Often, we are only interested in the sign of the derivative, since for the magnitude we need numbers for specific parameters. In addition, we can try to say some general things about the sign of the derivative. In this example, we can see that $\frac{\partial x(\theta)}{\partial \theta} > 0$ whenever $x(\theta) > \frac{1}{3} + \theta$ or $x(\theta) < \theta$. Since we also know that x is bounded above by 2, we can say for certain that if $\theta > 2$, then $\frac{\partial x(\theta)}{\partial \theta} > 0$ ♣

Exercise 10.3. *Consider a Cournot model, in which two firms compete for profits by choosing a quantity of a product to produce. The firms face an inverse demand curve where the price $P(Q)$ is determined by the total quantity produced by both firms, $Q = q_1 + q_2$. Assume that $P(Q)$ is decreasing and concave. In addition, each firm has a marginal cost function $c(q_i)$ that is convex.*

Consider firm 1's "best response function": the optimal q_1^* as a function of q_2. Show that $q_1^*(q_2)$ is decreasing in q_2.

Exercise 10.4. *Suppose $f(\mathbf{z})$ is a concave production function with $L-1$ inputs $(z_1, ..., z_{L-1})$. Suppose also that $\frac{\partial f(\mathbf{z})}{\partial z_l} \geq 0$ for all l and $z \geq 0$ and that the matrix $D^2 f(\mathbf{z})$ is negative definite at all $\mathbf{z}$. Use the firm's first-order conditions and the implicit function theorem to prove the following statements:*

(a) An increase in the output price always increases the profit-maximizing level of output.

(b) An increase in the output price increases the demand for some input.

(c) An increase in the price of an input leads to a reduction in the demand for the input.

This implicit function approach may also be used for constrained optimization problems for which the KKT conditions are known to be necessary and sufficient (for example, for convex programs with differentiable objective and constraint functions). In this case, since the Lagrange multipliers are also a function of θ, one must be careful to write down the full set of KKT conditions and to carefully calculate the relevant derivatives.

10.4.2 Topkis' theorem

Topkis' Theorem, like the Implicit Function Theorem is used to answer how the optimal solution(s) vary with the parameter. However, it allows us to answer the question under much more general assumption. You will study Topkis's Theorem carefully in the first-year micro sequence, so I will give only a very brief introduction to the topic.

The main barrier to the implicit function theorem approach is that the first-order conditions may be hard to write down or there may be many solutions to the first-order conditions which may or may not be the solutions of interest. The Topkis theorem approach does not even require you to write the first-order conditions down!

Instead, the crucial assumption is that x and θ are complementary in the objective function $f(x, \theta)$. When f is smooth, this is equivalent to $f_{x\theta} \geq 0$. Since we know $f_{xx} \leq 0$ at a maximum, the implicit function theorem would then imply that $x'(\theta) \geq 0$. We now introduce a condition similar to $f_{x\theta} \geq 0$ that does not require differentiability.

Definition 10.4.1. The function $f : X \times \Theta \to \mathbb{R}$ with $X, \Theta \subseteq \mathbb{R}$ has **increasing differences** if for all $x, x' \in X$ with $x' \geq x$ and $\theta, \theta' \in \Theta$ with $\theta' \geq \theta$,

$$f(x', \theta') - f(x, \theta') \geq f(x', \theta) - f(x, \theta).$$

It has **strict increasing differences** if for all $x, x' \in X$ with $x' > x$ and $\theta, \theta' \in \Theta$ with $\theta' \geq \theta$, we have

$$f(x', \theta') - f(x, \theta') > f(x', \theta) - f(x, \theta).$$

The interpretation here is that the incremental benefit of increasing x, that is $f(x', \theta) - f(x, \theta)$, increases when you increase θ. Note that the definition may be symmetrically rewritten to switch the roles of x and θ.

Theorem 10.4.2. *Under the assumptions that the derivatives below are well-defined, f has increasing differences if and only if:*

(a) $f_x(x, \theta)$ is nondecreasing in θ for all x,

(b) $f_\theta(x, \theta)$ is nondecreasing in x for all θ, or

(c) $f_{x\theta} \geq 0$.

If f is twice-differentiable and $f_{x\theta} > 0$, then f has strictly increasing differences (although this condition is not necessary).

We have the following famous results.

Theorem 10.4.3 (Univariate Topkis Theorem)**.** *Suppose that the objective function $f : X \times \Theta \to \mathbb{R}$ with $X, \Theta \subseteq \mathbb{R}$ has increasing differences. Let $\theta' > \theta$ with $x \in x^*(\theta)$ and $x' \in x^*(\theta')$. Then either:*

- *$x' \geq x$, or*
- *$x' \in x^*(\theta)$ and $x \in x^*(\theta')$.*

Theorem 10.4.4 (Monotone Selection Theorem)**.** *Suppose that the objective function $f : X \times \Theta \to \mathbb{R}$ with $X, \Theta \subseteq \mathbb{R}$ has strict increasing differences. Then for all $\theta' > \theta$, $x \in x^*(\theta)$, $x' \in x^*(\theta')$, we have $x' \geq x$. That is, any selection from $x(\theta)$ is nondecreasing.*

In ECON 202, you will learn various versions of the Topkis theorem that apply for multivariate optimization problems, as well as some necessary conditions for $x(\theta)$ to be nondecreasing. I don't want to spoil Ilya's fun, so let's move on.

Exercise 10.5. *An agent is participating in a first-price auction with n other bidders. Her value for the object is v. Show that the agent's optimal bid is a nondecreasing function of her value.*

Exercise 10.6. *A firm is developing a new product and is evaluating how long to wait before launching it. A longer development time allows the firm to improve its production technology, which results in cost savings and better product quality. On the other hand, the firm knows that a direct competitor is working on a similar product, and it realizes that whoever introduces the product first will capture a significant share of the market. Specifically, if the competitor enters first, then our firm will be left with profits ω, while if it introduces its product at time t and the competitor hasn't yet entered, it will enjoy a profit $\pi(t) > \omega$, with $\pi'(t) > 0$. The firm believes that the competitor's time of entry is distributed exponentially with parameter λ, that is the probability that the firm enters at or before time t is $1 - e^{-\lambda t}$. Formulate the firm's optimization problem, and evaluate the influence of λ on the firm's optimal waiting time.*

10.4.3 Envelope theorem

The *Envelope Theorem* is a tool used to compute the derivative of the value function with respect to the parameters. We will introduce three versions of the envelope theorem: one the 'classical' differentiable envelope theorem for unconstrained optimization problems, the second is a significant generalization of the theorem with fewer assumptions, and the third is an application of the first theorem to constrained maximization problems.

Here we may allow X to be an abstract choice set (it need not be a subset of Euclidean space).

Theorem 10.4.5 (Envelope Theorem - Differentiable Form). *Suppose that for some θ, $V'(\theta)$ exists and $x^*(\theta)$ is non-empty with $x^* \in x^*(\theta)$. Then $V'(\theta) = f_\theta(x^*, \theta)$.*

Proof. By definition of the optimization problem,

$$V(\theta + \varepsilon) \geq f(x^*, \theta + \varepsilon),$$

and $V(\theta) = f(x^*, \theta)$ so that

$$V(\theta + \varepsilon) - V(\theta) \geq f(x^*, \theta + \varepsilon) - f(x^*, \theta).$$

Hence,

$$\begin{aligned} V'(\theta) &= \lim_{\varepsilon \downarrow 0} \frac{V(\theta + \varepsilon) - V(\theta)}{\varepsilon} \\ &\geq \lim_{\varepsilon \downarrow 0} \frac{f(x^*, \theta + \varepsilon) - f(x^*, \theta)}{\varepsilon} \\ &= f_\theta(x^*, \theta). \end{aligned}$$

Similarly letting $\varepsilon \uparrow 0$ establishes $V'(\theta) \leq f_\theta(x^*, \theta)$. ■

The problem with this approach is that it assumes that $V'(\theta)$ exists, which typically requires calculating $V(\theta)$ or applying some careful analysis. A very nice generalization due to Milgrom and Segal (2002) applies with many fewer assumptions.

Theorem 10.4.6 (Envelope Theorem - Integral Form). *Let $\Theta = \mathbb{R}$ and suppose that for almost all θ, $x^*(t)$ is nonempty and let $\chi^*(t)$ be a selection from this correspondence. Suppose moreover that $f_\theta(x, \theta)$ exists for all x, θ and that there exists some function $b : \Theta \longrightarrow \mathbb{R}_+$ such that $|f_\theta(x, \theta)| \leq b(\theta)$ for all θ and $\int_0^\theta b(t)dt < \infty$.*

Then $V(\theta)$ is differentiable almost everywhere, and

$$V(\theta) = V(0) + \int_0^\theta f_\theta(\chi^*(t), t)dt.$$

Finally, we consider an application of the first envelope theorem to constrained optimization programs. In particular, consider any programs for which strong duality holds and the KKT conditions are known to be necessary and sufficient for $x^*(\theta)$ to be an optimal solution to the original problem (e.g., convex differentiable programs). Moreover, suppose that $x^*(\theta)$ is differentiable in an open neighborhood of θ_0, a parameter value of interest. This implies that V is differentiable in a neighborhood of θ_0.

Recall that we may re-express the constrained optimization problem as

$$V(\theta) = \min_{\lambda \geq 0} \phi(\lambda, \theta),$$

where ϕ is the dual function (which now depends on θ),

$$\phi(\lambda, \theta) = \max_{x \in X} \mathcal{L}(x, \lambda, \theta).$$

We can apply the differentiable envelope theorem to the dual formulation of $V(\theta)$ to obtain

$$V'(\theta) = \phi_\theta(\lambda^*(\theta), \theta).$$

We can also apply the differentiable envelope theorem to the definition of ϕ to obtain

$$\phi_\theta(\lambda^*(\theta), \theta) = \mathcal{L}_\theta(x^*(\theta), \lambda^*(\theta), \theta).$$

Theorem 10.4.7 (Envelope Theorem for Lagrangians). *Consider the primal problem with continuously differentiable objective and constraint functions, and suppose that $x^*(\theta)$ is differentiable in an open neighborhood of θ_0.*

Then $V(\cdot)$ is differentiable in an open neighborhood of θ_0 and

$$V_{\theta_i}(\theta_0) = \mathcal{L}_{\theta_i}(x^*(\theta_0), \lambda^*(\theta_0), \theta_0) = f_{\theta_i}(x^*(\theta_0), \theta_0) - \lambda^*(\theta_0) \cdot g_{\theta_i}(x^*(\theta_0), \theta_0),$$

where $\lambda_1^*(\theta_0), ..., \lambda_K^*(\theta_0)$ are the Lagrange multipliers associated with $x^*(\theta_0)$.

That is, the derivative of the value function is equal to the derivative of the Lagrangian.

For constraint k that does not bind, $\lambda_k = 0$ so we can change anything that λ_k is multiplying without altering any equations. That justifies the above analysis. However, it needs to be the case that the set of binding constraints does not change with the parameters in an open neighborhood around the parameter of interest. This is ensured in the theorem by the assumption that $x^*(\hat{\theta})$ is differentiable in an open neighborhood around $\hat{\theta}$.

One might wonder why is the theorem called the envelope theorem. To answer, consider a graphical example in a one dimensional case, $x \in B \subseteq \mathbb{R}$, $\theta \in A \subseteq \mathbb{R}$, where A and B are open set. For every possible $x \in B$ draw the objective function in the θ-space. The "envelope" of these curves is the value function, V. So, the envelope theorem tells us about the derivative of the envelope function.

Example. Consider the same example as example 1 in the implicit function theorem section. Recall the problem:

$$V(\theta) = \max_{0 \leq x \leq 2} (x - \theta)^3 - x^2$$

Answer:

The Lagrangian is

$$\mathcal{L}(x, \lambda, \mu, \theta) = (x - \theta)^3 - x^2 + \lambda(2 - x) + \mu x$$

The derivative, evaluated at the optimum is:

$$\frac{\partial V(\theta)}{\partial \theta} = \frac{\partial \mathcal{L}(x^*(\theta), \lambda, \mu, \theta)}{\partial \theta} = -3(x^*(\theta) - \theta)^2$$

So, the value function is non-increasing (and in general decreasing) in θ. ♣

Example. The envelope theorem is used a lot in economics. It is often useful not only when we are interested in the value function, but even to characterize or find explicit solutions for the optimal choices $x^*(\theta)$. You will encounter such use of the envelope theorem in the first year (mostly in the micro sequence), under names such as Hotelling's Lemma, Shepherd's Lemma, Roy's Identity, etc. If you can spot a simple application of the envelope theorem when you come across these lemmas, you will not need to memorize them separately, but be able to quickly derive them if needed. Let's derive Roy's identity here. Roy's identity is used in consumer theory to get a (Marshallian) demand function from a consumer's indirect utility (which is the value function of the consumer problem). The consumer faces a problem of choosing a bundle $x \in \mathbb{R}^n$ given the price vector p and wealth w:

$$\begin{aligned} V(p, w) &= \max_{x \in \mathbb{R}^n} u(x) \\ \text{subject to } x_i &\geq 0 \qquad\qquad \text{for } i = 1, \ldots, n \\ p \cdot x &\leq w \end{aligned}$$

Let's assume that the solution is in the interior (so that we can ignore nonnegativity constraints), so the Lagrangian is

$$\mathcal{L}(x, \lambda, p, w) = u(x) + \lambda(w - p \cdot x)$$

Now by the envelope theorem

$$\begin{aligned} \frac{\partial V(p, w)}{\partial p_i} &= \frac{\partial \mathcal{L}(x, \lambda, p, w)}{\partial p_i} = -\lambda x^*(p, w) \\ \frac{\partial V(p, w)}{\partial w} &= \frac{\partial \mathcal{L}(x, \lambda, p, w)}{\partial w} = \lambda \end{aligned}$$

Therefore

$$x^*(p, w) = -\frac{\frac{\partial V(p,w)}{\partial p_i}}{\frac{\partial V(p,w)}{\partial w}}$$

Hence, if we know the indirect utility function, we can derive a consumer's demand function. This will be a very useful link in Econ 202. ♣

Exercise 10.7. *A price-taking firm produces output q from inputs z_1 and z_2 according to a differentiable concave production function $f(z_1, z_2)$. The price of its output is $p > 0$ and the prices of its inputs are $(w_1, w_2) >> 0$. However, there are two unusual things about this firm. First, rather than maximizing profit, the firm maximizes revenue. Second, the firm is cash constrained. In particular, it has only C dollars on hand before production and, as a result, its total expenditures on inputs cannot exceed C.*

Suppose one of your econometrician friends tells you that she has used repeated observations of the firm's revenues under various output prices, input prices, and levels of the financial constraint and has determined that the firm's revenue level R can be expressed as the following function of the variables (p, w_1, w_2, C):

$$R(p, w_1, w_2, C) = p\,[\gamma + \log C - \alpha \log w_1 - (1-\alpha)\log w_2]$$

What is the firm's use of output 1, $z_1(p, w_1, w_2, C)$?

Exercise 10.8. *A monopolist sells products in two markets, A and B. The demand for the firm's products in these markets are $Q_A = 100 - 0.4P_A + 0.1P_B$ and $Q_B = 120 - 0.5P_B + 0.2P_A$. The firm has a constant marginal cost of $c = 40$ and a fixed cost $F = 2500$. Use the envelope theorem to approximate the change in profits caused by a change in marginal cost to $c = 41$.*

Exercise 10.9. *In this problem, we will prove the famous "revenue equivalence theorem" in mechanism design. An agent has a value v for a good. The mechanism designer does not know the agent's value, but supposes it is drawn from an absolutely continuous distribution with cdf F, that is $\Pr(v \leq \bar{v}) = F(\bar{v})$ with $0 \leq v \leq 1$. The designer commits in advance to an allocation rule $x(v)$, which is the probability the agent is allocated the good if they report value v, and a payment rule $t(v)$, which is the amount the agent needs to pay the mechanism designer. The agent will choose their report r so as to maximize their utility*

$$U(v, r) = vx(r) - t(r).$$

Let $U^*(v) = \max_r vx(r) - t(r)$. Use the envelope theorem to express $t(v)$ as a function of $x(v)$.

Part V

Dynamic optimization

11

Dynamical systems

Contents

11.1 Continuous dynamical systems: differential equations

Many economic problems naturally lend to relationships between the values of variables and their rates of change. Such situations are well modelled by differential equations.

Definition 11.1.1. Let $x : \mathbb{R} \to \mathbb{R}^m$ be a vector-valued function which is n-times differentiable. An **ordinary differential equation (ODE)** is an equatiDon of the form

$$g(x^{(n)}(t), x^{(n-1)}(t), ..., \dot{x}(t), x(t), t) = 0.$$

If the equation may be written

$$x^{(n)}(t) = g(x^{(n-1)}(t), x^{(n-2)}(t), ..., \dot{x}(t), x(t), t),$$

the ODE is said to be **explicit**. The **order** of an ODE is the order of the highest derivative in the equation.

An ODE is **linear** if it takes the form

$$a_n(t) \cdot x^{(n)}(t) + a_{n-1}(t) \cdot x^{(n-1)}(t) + ... + a_1(t) \cdot \dot{x}(t) + a(t) \cdot x(t) + b(t) = 0,$$

else it is **nonlinear**. A linear ODE with $b(t) = 0$ is called **homogeneous**.

An ODE is **autonomous** if it depends on t only through the dependence of x and its derivatives on t.

In these notes we will exclusively study explicit differential equations. In the absence of this assumption, analysis is typically much more complicated, using analytic tools similar to the implicit function theorem.

There are typically many functions $x(t)$ that satisfy a differential equation. Thus, a variety of other conditions are imposed in order to obtain a unique solution.

Definition 11.1.2. A **boundary value** (or initial value or terminal value, depending on the context) is a condition of the form $x(t_0) = x_0$ for some $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{R}^m$.

A **transversality condition** is an ODE with an assumption on $\lim_{t\to\infty} x(t)$. Typically transversality conditions are insufficient to fully characterize a solution, so these may be coupled with other conditions.

We have the following important results about the existence and uniqueness of solutions to boundary value problems.

Theorem 11.1.3. *Consider the boundary value problem $\dot{x}(t) = g(x(t), t)$ with $x(t_0) = x_0$ and let U be an open set in $\mathbb{R}^m \times \mathbb{R}$ containing (t_0, x_0).*

(a) ***Peano's Theorem****: Suppose $g : U \longrightarrow \mathbb{R}^m$ is continuous. Then there exists a solution to the initial value problem.*

(b) ***Picard-Lindelöf Theorem****: If in addition g is Lipschitz in x on U, then there exists an interval (a, b) containing t_0 such that the solution is unique on (a, b).*

Although the above theorem appears to apply on to first-order explicit ODEs, in fact, it may be applied much more generally using a reduction of higher-order explicit ODEs to a system of first order ODEs, as follows. Suppose $x^{(n)} = g(x^{(n-1)}(t), x^{(n-2)}(t), \ldots, \dot{x}(t), x(t), t)$. Then define $y_n(t) = x^{(n-1)}(t)$, with $y_1(t) = x(t)$. The new system of differential equations is then $\dot{y}_1(t) = y_2(t)$, $\dot{y}_2(t) = y_3(t)$, ..., $\dot{y}_{n-1}(t) = y_n(t)$ and $\dot{y}_n(t) = g(y_n(t), y_{n-1}(t), \ldots, y_1(t), t)$. Existence and uniqueness for the higher-order equations can thus sometimes be obtained by careful application of the Peano or Picard-Lindelöf Theorems to this system.

We will mostly study *linear* ODEs. The analysis of nonlinear ODEs is typically quite challenging, relying mostly on numerical methods and localized linearization (or log-linearization) of the ODE (that is, exploiting methods like Taylor's Theorem). You will learn some of these methods in the last quarter of the macroeconomics sequence.

The study of *homogeneous* linear ODEs will be particularly important. This is because linear

ODEs may be written as linear transformations on $C^n(\mathbb{R}, \mathbb{R}^m)$, that is if $T : C^n(\mathbb{R}, \mathbb{R}^m) \longrightarrow C^n(\mathbb{R}, \mathbb{R}^m)$ is defined by

$$x(t) \mapsto a_n(t) \cdot x^{(n)}(t) + a_{n-1}(t) \cdot x^{(n-1)}(t) + ... + a_1(t) \cdot \dot{x}(t) + a(t) \cdot x(t),$$

then $T(\alpha x_1 + x_2) = \alpha T(x_1) + T(x_2)$ for any constant $\alpha \in \mathbb{R}$. Thus the general solution of these ODEs are the nullspace in $C^n(\mathbb{R}, \mathbb{R}^m)$ of the ODE operator. It turns out this nullspace is $m-$dimensional, so that there are m basis vectors–linearly independent homogenous solutions to the equation. Then if $x_p(t)$ is a "particular" solution of the ODE

$$a_n(t) \cdot x^{(n)}(t) + a_{n-1}(t) \cdot x^{(n-1)}(t) + ... + a_1(t) \cdot \dot{x}(t) + a(t) \cdot x(t) + b(t) = 0, \ x(t_0) = x_0$$

then so is $x_p(t) + x_H(t)$ where $x_H(t)$ is any element in the nullspace of the operator (that is, any solution to the homogeneous equation.

Exercise 11.1. *Let $a : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous function. Show that the ODE*

$$\dot{x}(t) = \sqrt{a(t)^2 + x(t)^2}, \ x(0) = x_0$$

has a unique solution.

11.1.1 First-order ODEs

We begin by considering the case when $m = 1$ and so $x : \mathbb{R} \longrightarrow \mathbb{R}$. We will often suppress the dependence on t, writing $x(t)$ as x and $\dot{x}(t)$ as $\dot{x}$. Even this one-dimensional case may not be simple to solve.

One family of one-dimensional ODEs we already know how to solve takes the form

$$\dot{x} = f(t), \ x(0) = x_0$$

for measurable f, which by the fundamental theorem of calculus has the solution $x(t) = x_0 + \int_0^t f(t)dt$. This is not so exciting, but the logic of this result can be greatly extended.

Separable First-Order ODEs

A related family of *nonlinear* ODEs that are fairly simple to solve are the so-called **separable first-order ODEs**

$$\dot{x} = a(t)P(x),$$

which can be re-organized as

$$\frac{\dot{x}}{P(x)} = a(t).$$

Using the fundamental theorem of calculus, we can integrate each side to obtain an implicit expression for the solution

$$\int \frac{1}{P(x)} dx = \int a(t) dt + C.$$

Given a boundary condition, the value of the constant C can be identified.

Linear first-order ODEs

We now focus on solving the **linear first-order ODE**

$$\dot{x}(t) = a(t)x(t) + b(t),$$

for sufficiently nice $a(t)$ and $b(t)$ (that is, both will need to be integrable, as will some functions of them). The trick is to reorganize and multiply the equation by an **integrating factor** $e^{-\int a(t)dt}$, the derivative of which is $a(t)e^{-\int a(t)dt}$, to obtain

$$e^{-\int a(t)dt}\dot{x}(t) - a(t)e^{-\int a(t)dt}x(t) = e^{-\int a(t)dt}b(t),$$

which may be rewritten as

$$\frac{d}{dt}\left\{e^{-\int a(t)dt}x(t)\right\} = e^{-\int a(t)dt}b(t).$$

Theorem 11.1.4. *The solution to the linear first-order ODE $\dot{x}(t) = a(t)x(t) + b(t)$ takes the form*

$$x(t) = \left[C + \int e^{-\int a(t)dt}b(t)dt\right] e^{\int a(t)dt}$$

for some constant C pinned down by the initial condition.

You can see that the function $e^{-\int a(t)dt}$ spans the nullspace of the linear first-order ODE operator.

A particularly important one is the constant coefficient homogeneous first-order ODE

$$\dot{x} = ax, \; x(t_0) = x_0$$

which says that the relative rate of growth of x is constant. We can rewrite this equation as

$$\frac{d}{dt}\log(x(t)) = a,$$

so that $x(t) = Ce^{at}$ where $C = x_0 e^{-at_0}$.

Exercise 11.2. *Solve the differential equation*

$$\frac{dy}{dt} + t^2 y = 5t^2,\ y(0) = 6.$$

11.1.2 First-order ODE Systems

Constant coefficients

We now consider $x(t) = [x_1(t), x_2(t) \cdots x_m(t)]'$, and equations of the form

$$\dot{x}(t) = Ax(t),\ A \in \mathbb{R}^{m\times m},\ \text{ with } x(t^0) = x^0.$$

Note that if A is a diagonal matrix, say, $A = \text{diag}(\lambda_1, ..., \lambda_m)$, then $x_i(t) = x^0 e^{\lambda_i(t-t_0)}$, as above.

Now suppose that A is diagonalizable, so that $A = P\Lambda P^{-1}$. Recall that Λ is then a diagonal matrix with the m eigenvalues $\lambda_1, \lambda_2, ..., \lambda_m$ on the diagonal (repeated as often as their algebraic multiplicity) and P is a matrix with each associated eigenvector $v_1, v_2, ..., v_m$. The ODE system can then be written

$$\begin{aligned} \dot{x}(t) &= P\Lambda P^{-1}x(t) \\ P^{-1}\dot{x}(t) &= \Lambda P^{-1}x(t) \\ \dot{y}(t) &= \Lambda y(t). \end{aligned}$$

Then $y_i(t) = C_i e^{\lambda_i t}$ for some constant C. Finally $x(t) = Py(t)$. This establishes the following.

Theorem 11.1.5. *Consider the ODE system $\dot{x}(t) = Ax(t)$ for $A \in \mathbb{R}^{m\times m}$ with m eigenvalues $\lambda_1, ..., \lambda_m$ and associated distinct eigenvectors $v_1, v_2, ..., v_m$. Then*

$$x(t) = \sum_{i=1}^{m} C_i v_i e^{\lambda_i t},$$

for some C_i pinned down by the initial conditions $x(t_0) = x_0$.

Note that if $\dot{x}(t) = Ax(t) + B$ for some constant vector B, then since $x_P(t) = -A^{-1}B$ is clearly a particular solution of the problem, then the general solution takes the form $x(t) = -A^{-1}B + \sum_{i=1}^{m} C_i v_i e^{\lambda_i t}$. The vector $x^* = -A^{-1}B$ is called the **stationary state** of the system.

If A is not diagonalizable, recall that an $m\times m$ matrix has m generalized eigenvectors. Using the Jordan decomposition (and calculating the matrix exponential directly), we obtain the following.

Theorem 11.1.6. *Given a chain of generalized eigenvector of length r, let*

$$
\begin{aligned}
y_1(t) &= v_1 e^{\lambda t} \\
y_2(t) &= (tv_1 + v_2)\, e^{\lambda t} \\
y_3(t) &= \left(\frac{t^2}{2}v_1 + tv_2 + v_3\right) e^{\lambda t} \\
&\vdots \\
y_r(t) &= \left(\frac{t^{r-1}}{(r-1)!}v_1 + \ldots + \frac{t^2}{2}v_{r-2} + tv_{r-1} + v_r\right) e^{\lambda t}
\end{aligned}
$$

The functions $\{y_i(t)\}_{i=1}^r$ form r linearly independent solutions of $\dot{x}(t) = Ax(t)$.

Using this foregoing theorem, we can write the general solution to any linear homogeneous ODE system in the same way as in Theorem 11.1.5, that is, by summing over the linearly independent basis functions associated with each eigenvalue.

In fact, all these cases may be summarized using the concept of the *matrix exponential.*

Definition 11.1.7. Let $A \in \mathbb{C}^{n\times n}$. The **matrix exponential** of A is given by the power series

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k!}A^k.$$

It turns out that the above series always converges, so that the matrix exponential is well-defined. The solution to $\dot{x}(t) = Ax(t) + B$ can then be concisely summarized as $x(t) = x^* + \exp(At)[x(0) - x^*]$.

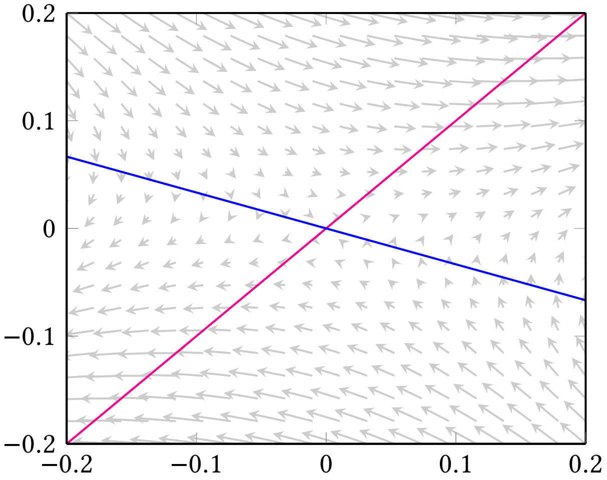
Phase plane diagrams

We will now focus on the case in which $m = 2$. The **phase plane** is a convenient way of analyzing autonomous first-order ODE systems. At point (x, y) in the plane, the value of $(\dot{x}, \dot{y})$ determines a direction in which the functions x and y change as a function of time. A **stationary point** or **equilibrium** is a point at which $\dot{x} = 0$ and $\dot{y} = 0$. A **nullcline** is the set of points for which $\dot{x} = 0$ *or* $\dot{y} = 0$.

For concreteness, let us consider the following example.

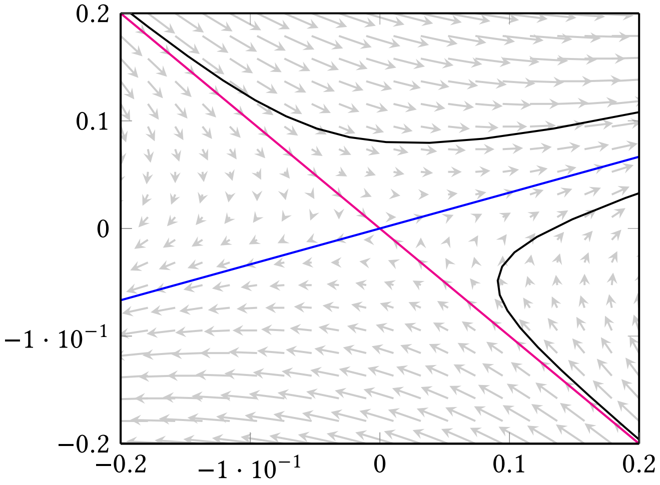
$$
\begin{aligned}
\dot{x} &= x + 3y \\
\dot{y} &= x - y
\end{aligned}
$$

It is clear that the (unique) stationary point of this system is at the origin. The nullclines are the lines $y = \frac{-1}{3}x$ and $y = x$ corresponding to $\dot{x} = 0$ and $\dot{y} = 0$ respectively. You can see a plot of the phase plane and its nullclines below.



We can calculate the solution by of the ODE system by calculating the eigenvalues and eigenvectors of the coefficient matrix. Check that the eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = -1$ with associated eigenvectors $v_1 = (3, 1)'$ and $v_2 = (1, -1)'$ respectively.

The eigenvectors are important directions in the phase plane. Recall that $x = v_i e^{\lambda_i t}$ is a solution to the system $\dot{x} = Ax$. Thus, the direction of v_i describes a straight-line trajectory in the phase plane. Outside of these straight-line trajectories, the solutions are typically curved toward the eigenvectors (with the dominant eigenvector exerting a greater "pull" than the eigenvector associated with the smaller eigenvalue).



Analyzing the phase plane, we can see some properties of the solutions to the above differential equation. Note that the straight-line trajectory associated with the negative root guides solutions back to the equilibrium point, while the other straight-line trajectory pushes solutions away from the equilibrium point. In fact, any perturbations away from the negative root's straight-line trajectory leads to a solution diverging away from the equilibrium. This form of solution is called a **saddle** and the stable line is called the **saddle-path**.

Generally, we say that this system is **unstable**, according to the definitions below.

Definition 11.1.8. Let $\bar{x}$ be an isolated steady-state of the system $\dot{x}(t) = g(x(t)), x(t) \in X$ and $t \in \mathbb{R}$. Steady-state $\bar{x}$ is **Lyapunov stable** if given any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\|x\left(t_0\right) - \bar{x}\| < \delta \text{ for any } t_0 \in \mathbb{R} \text{ implies that } \|x(t) - \bar{x}\| < \varepsilon \vee t \geq t_0.$$

A steady-state $\bar{x}$ of the system $\dot{x}(t) = g(x(t)\ )$, $x(t) \in X$ and $t \in \mathbb{R}$, is **globally asymptotically stable** if it is (Lyapunov) stable and, moreover, if for every $t_0 \in \mathbb{R}$ and $x\left(t_0\right) \in X$,

$$\|x(t) - \bar{x}\| \longrightarrow 0 \quad \text{as } t \longrightarrow \infty$$

A steady-state is **locally asymptotically stable** if it is (Lyapunov) stable and, moreover, there exists some $\delta > 0$ such that

$$\|x\left(t_0\right) - \bar{x}\| < \delta \text{ for any } t_0 \in \mathbb{R} \text{ implies that } \|x(t) - \bar{x}\| \longrightarrow 0 \quad \text{as } t \longrightarrow \infty$$

There is a subtle difference between these definitions: a system with solutions which are closed orbits around the origin (e.g., eccentric circles) is Lyapunov stable but not asymptotically stable. That is, it doesn't fly away from the equilibrium point but it doesn't lead into the equilibrium either.

What other kinds of systems can we obtain from a first-order ODE system? Recall that

$$\begin{aligned} \operatorname{tr}(A) &= \lambda_1 + \lambda_2 \\ \det(A) &= \lambda_1 \lambda_2 \end{aligned}$$

where λ_1 and λ_2 are the eigenvalues, which solve the characteristic quadratic equation

$$\begin{aligned} p_A(\lambda) &= \lambda^2 - \operatorname{tr}(A)\lambda + \det(A) = 0 \\ \Rightarrow \lambda_{1,2} &= \frac{\operatorname{tr}(A) \pm \sqrt{\operatorname{tr}(A)^2 - 4\det(A)}}{2} \end{aligned}$$

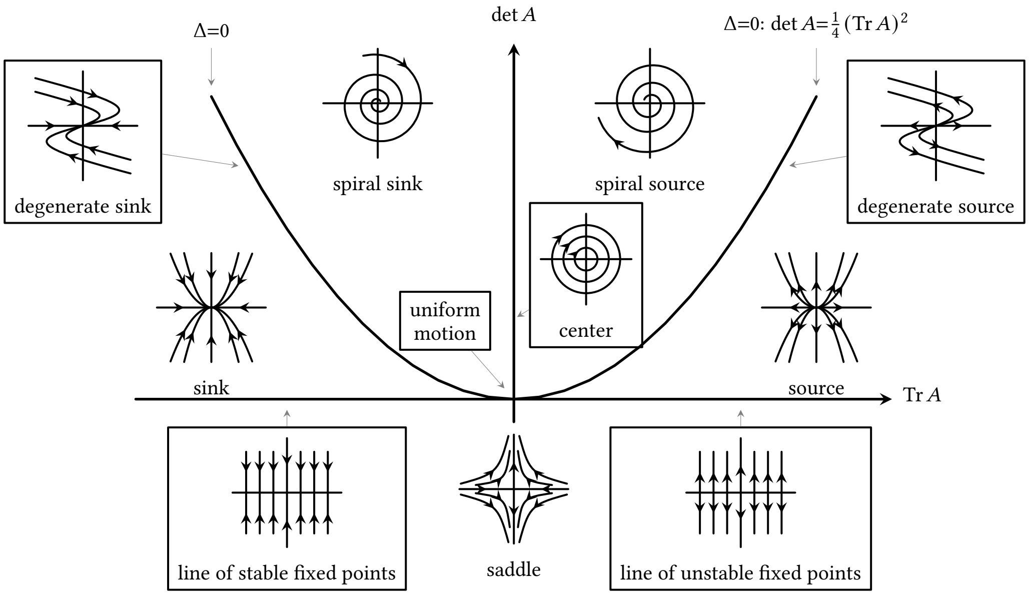


Figure 11.1. Poincaré Diagram: Classification of Phase Portraits in the (det A, Tr A)-plane

The value of the discriminant, $\Delta(A) = \text{tr}(A)^2 - 4\det(A)$ determines whether the eigenvalues are real and distinct $\Delta(A) > 0$, real and repeated $\Delta(A) = 0$ or complex $\Delta(A) < 0$.

It turns out that analysis of the trace and determinant is sufficient to determine the nature and stability of the ODE system. The full set of possibilities is illustrated in Figure 11.1 below. The main cases are as follows:

- If $\det(A) = \lambda_1\lambda_2 < 0$, the eigenvalues of the system are real numbers of opposite signs; hence, we have a **saddle point**. The saddle path (stable subspace), which is a straight line for linear systems, is determined by the eigenvector associated with the negative (stable) root.

- If $\det(A) = \lambda_1\lambda_2 > 0$, the roots are either complex numbers or real numbers of the same sign. In this case, there are two possibilities

  - If $\text{tr}(A) = \lambda_1 + \lambda_2 < 0$, the two eigenvalues are negative (if real) or have negative real parts; in either case, the system is stable. The steady-state is a **sink** .
  - If $\text{tr}(A) = \lambda_1 + \lambda_2 > 0$, both roots are positive (if real) or have positive real parts; in both cases the system is unstable. The steady-state is a **source**.

Exercise 11.3. *Consider the system*

$$\dot{x} = -5x - 2y,$$
$$\dot{y} = -x - 4y.$$

Find the general solution and sketch the phase portrait. Is the steady-state solution stable or unstable?

Higher-order linear ODEs

Recall that n-th order ODEs may be rewritten as n-dimensional systems of ODEs. This suggests that ODEs of the form

$$y^{(n)} + a_{n-1}y^{(n-1)} + a_{n-2}y^{(n-2)} + ... + a_1 y' + a_0 y = 0$$

may be solved using the constant-coefficient results above. This approach does work out, but it is in fact easier to use a short-cut motivated by the results obtained above. To do so, substitute the equation $y = e^{\lambda t}$ into the ODE, to obtain the equation

$$\lambda^n e^{\lambda t} + +a_{n-1}\lambda^{(n-1)}e^{\lambda t} + a_{n-2}\lambda^{(n-2)}e^{\lambda t} + ... + a_1\lambda e^{\lambda t} + a_0 e^{\lambda t} = 0,$$

which may be simplified, since $e^{\lambda t} \neq 0$ to

$$\lambda^n + +a_{n-1}\lambda^{(n-1)} + a_{n-2}\lambda^{(n-2)} + ... + a_1\lambda + a_0 = 0,$$

which is a polynomial in λ. It turns out this polynomial is just the characteristic equation of the equivalent ODE system, so that the "characteristic roots" λ_i are the eigenvalues of said system. If there are n distinct characteristic roots, the solution takes the form

$$y(t) = \sum_{i=1}^{n} c_i e^{\lambda_i t},$$

for some coefficients determined by the initial conditions. If there is a repeated root, then it turns out that multiplying the proposed solution by t leads to another linearly independent solution to the homogeneous equation (if the root is repeated more times, multiplying again by t leads to another linearly independent solution of the equation and so on).

In some cases it is also possible to find closed-form solutions to non-homogeneous higher-order ODEs,

$$y^{(n)} + a_{n-1}y^{(n-1)} + a_{n-2}y^{(n-2)} + ... + a_1 y' + a_0 y = b(t).$$

In order to do so, start by identifying the homogeneous solutions of the ODE, as described above. Then, it suffices to find a particular solution of the non-homogeneous equation. Start by guessing particular solutions of the "same form" as the right-hand side function $b(t)$. That is, if $b(t)$ is an k-th order polynomial in t, guess a generic kth-order polynomial in t as the particular solution and try to identify coefficients of such a polynomial which satisfy the equation. Similarly, if the right-hand side is an exponential, guess a similar exponential and so on. This method, called "undetermined coefficients" works in many cases for such equations.

Exercise 11.4. *Find the general solution to the differential equation*

$$y'' - 6y' + 2y = t.$$

11.1.3 Autonomous nonlinear systems

Many ODE systems of interest in economics are nonlinear. However, as foreshadowed earlier, such systems may typically be studied by taking local linearizations. We now briefly discuss the theory underlying this approach.

Definition 11.1.9. A steady-state x^* of $\dot{x} = F(x(t))$ (where F is continuously differentiable) is **hyperbolic** if $DF(x^*)$ has full rank.

It turns out if the steady-state is hyperbolic, then the system "behaves locally like" the linear system with coefficient matrix $DF(x^*)$. By "behaves like" we mean the following.

Definition 11.1.10. A homeomorphism $f : U \longrightarrow V$ between metric spaces U and V is a continuous function with a continuous inverse.

Theorem 11.1.11 (Hartman-Grobman Theorem)**.** *Let x^* be a hyperbolic steady-state of $\dot{x} = F(x(t))$ where F is continuously differentiable. Then there exists a neighborhood of x^* in which the solutions of $\dot{x} = F(x(t))$ are equivalent up to "direction-preserving" homeomorphism to $\dot{x} = Df(x^*)(x(t) - x^*)$.*

Here, 'direction-preserving' means it preserves the (time-)direction of trajectories in the phase plane (it is a hassle to define this technically, even though it is simple enough to understand intuitively).

The previous theorem implies that if $Df(x^*)$ has no zero eigenvalues, then the local behavior of the system around x^* can be fully determined by analysis of the eigenvalues (as in Figure 11.1). For example, if all the eigenvalues have negative real parts, x^* is locally asymptotically stable, while if any eigenvalue has a positive real part, x^* is Lyapunov unstable.

Exercise 11.5. *Consider the growth model*

$$\begin{aligned}\dot{k}(t) &= f(k(t)) - c(t) - \delta k(t)\\ \dot{c}(t) &= [f'(k(t)) - (\delta + \rho)]c(t),\end{aligned}$$

where $\rho > 0$, $\delta \in (0, 1)$ are parameters, $k(t)$ is the capital stock and the production function f satisfies

$$f(0) = 0,\ f' > 0,\ f'' < 0,\ \lim_{k\to\infty} f'(k) = 0,\ \lim_{k\downarrow 0} f'(k) = \infty.$$

Identify the steady-state of the system and describe the system's behavior locally around the steady-state. Sketch the phase portrait.

11.2 Discrete dynamical systems: difference equations

We now consider equations of the form

$$x(t + 1) = f(x(t), t),$$

where $t \in \mathbb{Z}$. It turns out the behavior of such systems tends to be very similar to that of differential equations, with a few adjustments.

The simplest difference equation is the autonomous linear first-order differencee equation

$$x(t + 1) = ax(t) + b,\ x(0) = x_0.$$

To solve this system, we can repeatedly substitute into the equation, as

$$\begin{aligned}x(1) &= ax_0 + b\\ x(2) &= a^2x_0 + ab + b\\ &\vdots\\ x(t) &= \begin{cases} x_0 + bt & \text{if } a = 1\\ \frac{b}{1-a} + a^t\left(x_0 - \frac{b}{1-a}\right) & \text{if } a \neq 1.\end{cases}\end{aligned}$$

Exercise 11.6. *Prove the above formula for $x(t)$ using induction.*

Note the similarity to the solution of the ODE $\dot{x} = ax + b$, which is $x = \frac{-b}{a} + \left(x_0 + \frac{b}{a}\right) e^{at}$. The solution to the difference equation is asymptotically stable (around $\frac{b}{1-a}$) if and only if $|a| < 1$.

Analogies of the solution methods for first-order ODE systems apply to difference equations. We state these results without proof.

Theorem 11.2.1. *The vector difference equation*

$$x(t+1) = F(x(t), t),\ x(t_0) = x_0$$

where $x : \mathbb{Z} \longrightarrow \mathbb{R}^m$ and $F : \mathbb{R}^m \times \mathbb{Z} \longrightarrow \mathbb{R}^m$ has a unique solution for all $t \geq t_0$ and as long as F is a invertible function, the solution is unique for all $t \in \mathbb{Z}$.

Theorem 11.2.2. *Suppose $A \in \mathbb{R}^{m\times m}$ with distinct eigenvalues $\lambda_1, ..., \lambda_n$, each with modulus not equal to 1. Then the unique solution to*

$$x(t+1) = Ax(t) + b,\ x(0) = x_0,$$

takes the form

$$x(t) = -[A - I_n]^{-1} b + \sum_{i=1}^{n} c_i \lambda_i^t v_i,$$

where v_i is the eigenvector associated with eigenvalue λ_i and c_i is a constant determined by the initial condition.

12

Optimal control

Contents

In this section, we introduce the optimal control problem. In the typical optimal control, the decision-maker obtains flow payoffs from the value of some variable (the “state variable”) over time but can also make choices of another variable (the “control variable”) that influence the rate of growth of the state variable. We distinguish between the finite and infinite horizon cases. While the results for the two cases are very similar, proofs for infinite horizon are substantially more difficult. Since the finite horizon proofs contain all the important intuition, these will be presented in these notes. The proofs of the other statements can be found in the book "Introduction to Modern Economic Growth" by Acemoglu.

12.1 Finite horizon

Definition 12.1.1. The **finite-horizon optimal control** problem takes the form:

$$\max_{x(t),y(t)} W(x(t), y(t)) \equiv \int_0^{t_1} f(t, x(t), y(t))dt$$

subject to

$$\dot{x}(t) = g(t, x(t), y(t)), \text{ and}$$

$$x(t) \in \mathcal{X}, y(t) \in \mathcal{Y} \text{ for all } t, \text{ and } x(0) = x_0,$$

where $t_1 \in \mathbb{R}_+, \mathcal{X}$ and, $\mathcal{Y}$ are nonempty convex subsets of $\mathbb{R}$.

Variable x is referred to as the **state variable** and y as the **control variable**. A pair of functions that jointly satisfy $\dot{x}(t) = g(t, x(t), y(t))$ $x(t) \in \mathcal{X}, y(t) \in \mathcal{Y}$ for all $t, x(0) = x_0$ are referred to as an **admissible pair**.

There are two challenges in characterizing the solution to this problem that we did not have before:

- The choice variable y is a function rather than a vector or a finite dimensional object.
- The constraint takes the form of a differential equation.

Here is the result that establishes necessary conditions for any continuous interior solution:

Theorem 12.1.2 (Pontryagin's maximum principle)**.** *Consider the problem of maximizing (1) subject to (2) and (3), with f and g continuously differentiable. Suppose that this problem has an continuous interior solution $(\hat{x}(t), \hat{y}(t)) \in \text{Int}(\mathcal{X} \times \mathcal{Y})$. Then there exists a continuously differentiable function $\lambda(\cdot)$ defined on $t \in [0, t_1]$, such that $(\hat{x}(t), \hat{y}(t))$ satisfy the following necessary conditions:*

$$\begin{aligned}
0 &= H_y(t, \hat{x}(t), \hat{y}(t), \lambda(t)) \textit{ for all } t \in [0, t_1] \\
\dot{\lambda}(t) &= -H_x(t, \hat{x}(t), \hat{y}(t), \lambda(t)) \ \textit{ for all } t \in [0, t_1] \\
\dot{x}(t) &= H_\lambda(t, \hat{x}(t), \hat{y}(t), \lambda(t)) = g(t, x(t), y(t) \textit{ for all } t \in [0, t_1] \\
\lambda\,(t_1) &= 0 \\
x(0) &= x_0
\end{aligned}$$

*where the **Hamiltonian** $H(t, x, y, \lambda)$ is defined as*

$$H(t, x, y, \lambda) = f(t, x, y) + \lambda(t) g(t, x, y)$$

Moreover, the Hamiltonian $H(t, x, y, \lambda)$ satisfies the Maximum Principle

$$H(t, \hat{x}(t), \hat{y}(t), \lambda(t)) \geq H(t, \hat{x}(t), y, \lambda(t)) \quad \textit{for all } y(t) \in \mathcal{Y} \textit{ and } t \in [0, t_1]$$

We will give an intuitive (but not completely formal) proof of this result using the method of Lagrange multipliers we developed for the static optimization case. Let $\lambda(t)$ be the Lagrange multiplier on the constraint $\dot{x}(t) = g(t, x(t), y(t))$. The Lagrangian is

$$\begin{aligned}
&\int f(t, x(t), y(t))dt + \int \lambda(t)[g(t, x(t), y(t)) - \dot{x}(t)]dt \\
&= \int H(t, x(t), y(t), \lambda(t))dt - \int \lambda(t)\dot{x}(t)dt \\
&= \int H(t, x(t), y(t), \lambda(t))dt - \lambda(t)x(t)|_0^{t_1} + \int x(t)\dot{\lambda}(t)dt \\
&= \int \left[H(t, x(t), y(t), \lambda(t)) + x(t)\dot{\lambda}(t)\right] dt - \lambda(t_1)x(t_1) + \lambda(0)x_0.
\end{aligned}$$

We maximize the integrand pointwise with respect to $x(t)$ and $y(t)$. The former gives the equation

$$-H_x(t, \hat{x}(t), \hat{y}(t), \lambda(t)) = \dot{\lambda}(t)$$

while the latter gives the equation

$$H_y(t, \hat{x}(t), \hat{y}(t), \lambda(t)) = 0.$$

Interpreting the necessary conditions

Before discussing the proof of the above theorem, let us discuss the necessary conditions. The first necessary condition $H_y = 0$ is just the first-order conditions of the "unconstrained" problem in y. The third necessary condition $\dot{x} = H_\lambda$ is simply a restatement of the constraint that $\dot{x} = g(t, x(t), y(t))$.

The second necessary condition is more subtle. The function $\lambda(t)$ is called the **co-state** variable. From the envelope theorem, we can see that

$$\lambda(t) = \frac{\partial W(t, x(t), y(t))}{\partial x},$$

so that the co-state variable is the shadow value of relaxing the constraint that $\dot{x}(t) = g(t, x(t), y(t))$. Thus, the Hamiltonian and the maximal principle may be interpreted as maximizing the sum of the immediate payoff from x, y, which is $f(t, x(t), y(t))$ plus the value of future gains that accrue from investing in future changes in the sock of the state variable. The second necessary condition then represents a kind of "no-arbitrage" condition for the variable $x(t)$. Imagine that you have $x(t)$ at time t and consider purchasing δx more x. The cost of this purchase will be $\lambda(t)\delta x$, while the benefit is the $H_x \delta x \delta t$ that you gain over the coming instant δt plus the value of the additional x you hold at time $t + \delta t$ which is $\lambda(t + \delta t)\delta x$. In order for your choice of x to be optimal, it must be that these costs and benefits equalize, that is $H_x \delta t \delta x + \lambda(t + \delta t)\delta x = \lambda(t)\delta x$. Letting $\delta t \longrightarrow 0$, this implies that $H_x + \dot{\lambda}(t) = 0$.

Sufficient conditions

The necessary conditions give us candidate solutions for the problem, but checking that a given candidate is a maximizer can be hard. The following sufficient conditions often help.

Theorem 12.1.3 (Mangasarian's Sufficiency Conditions)**.** *Let f and g be continuously differentiable, and suppose that an interior continuous path $(\hat{x}(t), \hat{y}(t)) \in \mathrm{Int}(\mathcal{X} \times \mathcal{Y})$ exists and satisfies the necessary conditions. Suppose that $\mathcal{X} \times \mathcal{Y}$ is a convex set and given the resulting $\lambda(t)$, $H(t, x, y, \lambda)$ is jointly concave in $(x, y) \in \mathcal{X} \times \mathcal{Y}$ for all $t \in [0, t_1]$. Then the pair $(\hat{x}(t), \hat{y}(t))$ is a global maximizer.*

Moreover, if $H(t, x, y, \lambda)$ is strictly concave in $(x, y) \in \mathcal{X} \times \mathcal{Y}$ for all $t \in [0, t_1]$, then the pair $(\hat{x}(t), \hat{y}(t))$ is the unique solution.

Another version of more general sufficient conditions, is given in the following theorem.

Theorem 12.1.4 (Arrow's Sufficiency Conditions)**.** *Let f and g be continuously differentiable, and suppose that an interior continuous path $(\hat{x}(t), \hat{y}(t)) \in \mathrm{Int}(\mathcal{X} \times \mathcal{Y})$ exists and satisfies the necessary conditions. Given $\lambda(t)$, let*

$$M(t, x(t), \lambda(t)) \equiv \max_{y \in \mathcal{Y}} H(t, x(t), y(t), \lambda(t))$$

If $\mathcal{X}$ is convex and $M(t, x, \lambda)$ is concave in $x \in \mathcal{X}$ for every $t \in [0, t_1]$, then $(\hat{x}(t), \hat{y}(t))$ are global maximizers. Moreover, if $M(t, x, \lambda)$ is strictly concave in $(x, y) \in \mathcal{X} \times \mathcal{Y}$ for all $t \in [0, t_1]$, then the pair $(\hat{x}(t), \hat{y}(t))$ are unique solutions.

12.2 Infinite horizon

In this section, we will consider two forms of problems with infinite horizon, one with discounting and one without. In infinite horizon problems, we will need to be concerned with the behavior of candidate solutions as $t \to \infty$. Frankly, the setup with discounting is much more important in your first-year classes (and probably economics more generally), but I have included both for completeness.

No discounting

Definition 12.2.1. The **infinite-horizon no-discounting optimal control problem** takes the form

$$\max_{x(t),y(t)} W(x(t),y(t)) \equiv \int_0^{t_1} f(t,x(t),y(t))dt$$

subject to

$$\dot{x}(t) = g(t,x(t),y(t)), \text{ and}$$

$$x(t) \in \mathcal{X}, y(t) \in \mathcal{Y} \text{ for all } t, \text{ and } x(0) = x_0 \text{ and } \lim_{t\to\infty} b(t)x(t) \geq x_1,$$

where $t_1 \in \mathbb{R}_+$; $\mathcal{X}$ and $\mathcal{Y}$ are nonempty convex subsets of $\mathbb{R}$; and $b : \mathbb{R}_+ \to \mathbb{R}_+$is a function for which $\lim_{t\to\infty} b(t)$ exists and satisfies $\lim_{t\to\infty} b(t) < \infty$.

This new requirement on the limiting behavior of x is a form of *feasibility* condition (do not confuse it with a tranversality condition!). In economic problems, it often takes the form of a "no-Ponzi" condition. It represents the goal of identifying candidate solutions which do not wander off to $-\infty$ (typically, we do not consider such solutions as being economically meaningful).

Here is the result that establishes necessary conditions for any continuous interior solution

Theorem 12.2.2. *Suppose that the no-discounting optimal control problem has a piecewise continuous interior solution $(\hat{x}(t),\hat{y}(t))$. Define $H(t,x(t),y(t),\lambda(t))$ as in Theorem 12.1.2. Then $H(t,\hat{x}(t),\hat{y}(t),\lambda(t))$ must satisfy all the same conditions as in Theorem 12.1.2 as well as the feasibility condition $\lim_{t\to\infty} b(t)\hat{x}(t) \geq x_1$. Moreover, as long as $W(t,\hat{x}(t))$ is differentiable in x and t for t sufficiently large and $\lim_{t\to\infty} V_t(t,\hat{x}(t)) = 0$, then the pair $(\hat{x}(t),\hat{y}(t))$ also satisfies the transversality condition*

$$\lim_{t\to\infty} H(t,\hat{x}(t),\hat{y}(t),\lambda(t)) = 0.$$

The transversality condition is a necessary condition for an optimal policy to exist (which is necessary for any reasonable concept of equilibrium).

Sufficiency conditions are similar to before.

Theorem 12.2.3. *Suppose that an admissible pair $(\hat{x}(t),\hat{y}(t)) \in \mathrm{Int}(\mathcal{X} \times \mathcal{Y})$ satisfies the necessary conditions above. Given $\lambda(t)$, let*

$$M(t,x(t),\lambda(t)) \equiv \max_{y\in\mathcal{Y}} H(t,x(t),y(t),\lambda(t))$$

If $\mathcal{X}$ is convex and $M(t,x,\lambda)$ is concave in $x \in \mathcal{X}$ for every $t \in \mathbb{R}_+$and $\lim_{t\to\infty} \lambda(t)(\hat{x}(t)- \widetilde{x}(t)) \leq 0$ for all $\widetilde{x}(t)$ implied by an admissible control path $\bar{y}(t)$, then $(\hat{x}(t),\hat{y}(t))$ achieves the *global maximum of the optimal control problem. Moreover, if $M(t,x,\lambda)$ is strictly concave in $(x,y) \in \mathcal{X} \times \mathcal{Y}$ for all $t \in [0,t_1]$, then the pair $(\hat{x}(t),\hat{y}(t))$ is the unique solution.*

12.3 Discounted infinite horizon

This is the most common form of optimal control problem that you will encounter in the first-year coursework.

Definition 12.3.1. The **infinite-horizon optimal control problem with discounting** takes the form

$$\max_{x(t),y(t)} W(x(t),y(t)) \equiv \int_0^{t_1} e^{-\rho t} f(x(t),y(t))dt$$

subject to

$$\dot{x}(t) = g(t,x(t),y(t)), \text{ and}$$

$$x(t) \in \mathcal{X}, y(t) \in \mathcal{Y} \text{ for all } t, \text{ and } x(0) = x_0 \text{ and } \lim_{t\to\infty} b(t)x(t) \geq x_1,$$

where $t_1 \in \mathbb{R}_+$; $\mathcal{X}$ and $\mathcal{Y}$ are nonempty convex subsets of $\mathbb{R}$; and $b : \mathbb{R}_+ \to \mathbb{R}_+$is a function for which $\lim_{t\to\infty} b(t)$ exists and satisfies $\lim_{t\to\infty} b(t) < \infty$.

In comparison to the previous section, this only difference is that the more general time dependence of the objective function f in the previous section, is restricted in this section **only** through the **time discount** $e^{-\rho t}$.

In addition, we will make use of the following assumptions:

(a) f is weakly monotone in x and y, and g is weakly monotone in (t,x,y);

(b) there exists $m > 0$ such that $\left|g_y(t,x(t),y(t))\right| \geq m$ for all t and for all admissible pairs $(x(t),y(t))$; and

(c) there exists $M < \infty$ such that $\left|f_y(x,y)\right| \leq M$ for all x and y.

Instead of working with the usual Hamiltonian, it proves simpler to work with the **current value Hamiltonian** defined as follows:

$$\hat{H}(t,x(t),y(t),\mu(t)) \equiv f(x(t),y(t)) + \mu(t)g(t,x(t),y(t))$$

Theorem 12.3.2 (Maximum principle for Discounted Infinite-Horizon Problem)**.** *Consider the discounted optimal control problem, with f and g continuously differentiable. Suppose that this problem has a piecewise continuous interior solution $(\hat{x}(t), \hat{y}(t)) \in \mathrm{Int}(\mathcal{X} \times \mathcal{Y})$. Let $V(t, x(t))$ be the value function defined as:*

$$V(t, x(t)) = \sup_{(x(s), y(s)) \in \mathcal{X} \times y} \int_t^\infty e^{-\rho s} f(x(s), y(s)) ds$$

subject to $\dot{x}(s) = g(s, x(s), y(s))$ and $\lim_{s\to\infty} b(s)x(s) \geq x_1$ Suppose that $V(t, \hat{x}(t))$ is differentiable in x and t for t sufficiently large, that $V(t, \hat{x}(t))$ exists and is finite for all t and that $\lim_{t\to\infty} \frac{\theta V(t,\hat{x}(t))}{\partial t} = 0$.

Then there exists a continuously differentiable function $\mu(\cdot)$ defined on $t \in \mathbb{R}_+$, such that $(\hat{x}(t), \hat{y}(t))$ satisfy the following necessary conditions:

$$\begin{aligned}
&\hat{H}_y(t, \hat{x}(t), \hat{y}(t), \mu(t)) = 0 && \textit{for all } t \in \mathbb{R}_+ \\
&\hat{H}_x(t, \hat{x}(t), \hat{y}(t), \mu(t)) = \rho\mu(t) - \dot{\mu}(t) && \textit{for all } t \in \mathbb{R}_+ \\
&\hat{H}_\mu(t, \hat{x}(t), \hat{y}(t), \mu(t)) = \dot{x}(t) && \textit{for all } t \in \mathbb{R}_+ \\
&x(0) = x_0 \\
&\lim_{t\to\infty} b(t)x(t) \geq x_1
\end{aligned}$$

and the transversality condition

$$\lim_{t\to\infty} e^{-\mu t} \hat{H}(t, \hat{x}(t), \hat{y}(t), \mu(t)) = 0$$

Moreover, suppose that Assumption 1.8 *holds and that either $\lim_{t\to\infty} \hat{x}(t) = x^* \in \mathbb{R}$ or 9 $\lim_{t\to\infty} \frac{\dot{x}(t)}{x(t)} = \xi \in \mathbb{R}$. Then the transversality condition can be strengthened to*

$$\lim_{t\to\infty} e^{-\rho t} \mu(t) \hat{x}(t) = 0.$$

The following theorem comprises the sufficient conditions.

Theorem 12.3.3. *Suppose that an admissible pair $(\hat{x}(t), \hat{y}(t)) \in \mathrm{Int}(\mathcal{X} \times \mathcal{Y})$ satisfies the necessary conditions above. Given $\mu(t)$, let*

$$M(t, x(t), \mu(t)) \equiv \max_{y \in \mathcal{Y}} \hat{H}(t, x(t), y(t), \mu(t))$$

Suppose that $V(t, \hat{x}(t))$ exists and is finite for all t, that for any admissible pair $(x(t), y(t))$, $\lim_{t\to\infty} e^{-\rho t} \mu(t) x(t) \geq 0$ and that $\mathcal{X}$ is convex and $M(t, x, \lambda)$ is concave in $x \in \mathcal{X}$ for every $t \in$

$\mathbb{R}_+$, then $(\hat{x}(t), \hat{y}(t))$ achieves the global maximum). Moreover, if $M(t, x, \lambda)$ is strictly concave in $(x, y) \in \mathcal{X} \times \mathcal{Y}$ for all $t \in [0, t_1]$, then the pair $(\hat{x}(t), \hat{y}(t))$ is the unique solution.

The above theorems are very useful and powerful. Given these theorems, the following strategy is used to solve problems of the discounted infinite-horizon form:

1. Use the necessary conditions to locate a candidate interior solution.

2. Verify the concavity condition in the sufficient conditions and simply check that $\lim_{t\to\infty} e^{-\rho t}\mu(t)x(t) \geq 0$ for other admissible pairs.

If these conditions are satisfied, we will have characterized a global maximum.

Exercise 12.1. *Suppose that an agent has access to a single unit of a resource and is determining its optimal consumption plan for the resource. The stock of the resource at time t is $x(t)$, while the flow payoff from consumption $c(t)$ is $u(c(t))$. The agent, who discounts the future with rate $\rho > 0$ solves*

$$\max_{x(t),c(t)} \int_0^\infty e^{-\rho t} u(c(t))dt,$$

subject to $\dot{x}(t) = -c(t)$, with $x(t) \in [0, 1]$ given $x(0) = 1$. Identify the optimal consumption rule and the resulting $x(t)$.

Exercise 12.2. *Consider the following neoclassical growth model. Maximize the social planner's utility (the utility of a representative household):*

$$\max[k(t), c(t)]_{t=0}^\infty \int_0^\infty e^{-\rho t} \ln(c(t))dt$$

subject to the law of motion of capital $\dot{k}(t) = k(t)^\alpha - \delta k(t) - c(t)$ and $k(0) > 0$ where $\alpha, \delta \in (0, 1), \rho > 0$.

Characterize the necessary conditions for optimality for $c(t)$ and $k(t)$.